

# STUDIA INFORMATICA

Formerly: *Zeszyty Naukowe Politechniki Śląskiej, seria INFORMATYKA*  
**Quarterly**

Volume 33, Number 3B (108)

Marek SIKORA

WYBRANE METODY OCENY  
I PRZYCINANIA REGUŁ DECYZYJNYCH



Silesian University of Technology Press  
Gliwice 2012

**Editor in Chief**

**Dr. Marcin SKOWRONEK**  
Silesian University of Technology  
Gliwice, Poland

**Editorial Board**

**Dr. Mauro CISLAGHI**  
Project Automation  
Monza, Italy

**Prof. Bernard COURTOIS**  
Lab. TIMA  
Grenoble, France

**Prof. Tadeusz CZACHÓRSKI**  
Silesian University of Technology  
Gliwice, Poland

**Prof. Zbigniew J. CZECH**  
Silesian University of Technology  
Gliwice, Poland

**Prof. Jean-Michel FOURNEAU**  
Université de Versailles - St. Quentin  
Versailles, France

**Prof. Jurij KOROSTIL**  
IPME NAN Ukraina  
Kiev, Ukraine

**Dr. George P. KOWALCZYK**  
Networks Integrators Associates, President  
Parkland, USA

**Prof. Stanisław KOZIELSKI**  
Silesian University of Technology  
Gliwice, Poland

**Prof. Peter NEUMANN**  
Otto-von-Guericke Universität  
Barleben, Germany

**Prof. Olgierd A. PALUSINSKI**  
University of Arizona  
Tucson, USA

**Prof. Svetlana V. PROKOPCHINA**  
Scientific Research Institute BITIS  
Sankt-Petersburg, Russia

**Prof. Karl REISS**  
Universität Karlsruhe  
Karlsruhe, Germany

**Prof. Jean-Marc TOULOTTE**  
Université des Sciences et Technologies de Lille  
Villeneuve d'Ascq, France

**Prof. Sarma B. K. VRUDHULA**  
University of Arizona  
Tucson, USA

**Prof. Hamid VAKILZADIAN**  
University of Nebraska-Lincoln  
Lincoln, USA

**Prof. Adam WOLISZ**  
Technical University of Berlin  
Berlin, Germany

**STUDIA INFORMATICA is indexed in INSPEC/IEE (London, United Kingdom)**

© Copyright by Silesian University of Technology Press, Gliwice 2012

PL ISSN 0208-7286, QUARTERLY

Printed in Poland

The paper version is the original version

---

**ZESZYTY NAUKOWE POLITECHNIKI ŚLĄSKIEJ**  
OPINIODAWCY

Prof. dr hab. Jarosław STEPANIUK  
Prof. dr hab. inż. Alicja WAKULICZ-DEJA

**KOLEGIUM REDAKCYJNE**

REDAKTOR NACZELNY – Prof. dr hab. inż. Andrzej Buchacz

REDAKTOR DZIAŁU – Dr inż. Marcin Skowronek

SEKRETARZ REDAKCJI – Mgr Elżbieta Leśko

Pamięci Agnieszki



## **SPIS TREŚCI**

<b>1. Wstęp.....</b>	<b>9</b>
<b>2. Indukcja reguł decyzyjnych i klasyfikatory regułowe.....</b>	<b>17</b>
2.1. Reprezentacja zbioru przykładów .....	18
2.2. Reguły decyzyjne .....	22
2.3. Pokryciowy algorytm indukcji .....	26
2.4. Algorytmy q-ModLEM oraz RMatrix .....	40
2.5. Klasyfikator regułowy .....	49
2.6. Ocena klasyfikatorów .....	52
2.6.1. Miary oceny klasyfikatorów .....	52
2.6.2. Metody oceny eksperimentalnej .....	58
2.6.3. Metody porównywania zdolności klasyfikacyjnych.....	60
2.7. Rozwój metod indukcji reguł decyzyjnych .....	64
2.7.1. Poprawa zdolności klasyfikacyjnych i opisowych .....	64
2.7.2. Problemy związane z rozmiarem i niedoskonałością danych.....	69
<b>3. Obiektywne miary oceny reguł i ich zbiorów.....</b>	<b>73</b>
3.1. Miary definiowane na podstawie tablicy kontyngencji .....	75
3.1.1. Przestrzeń wartości .....	77
3.1.2. Własności.....	79
3.1.3. Własności wymagane podczas oceny reguł decyzyjnych.....	87
3.1.4. Równoważność i podobieństwo ze względu na porządek reguł .....	98
3.1.5. Równoważność ze względu na rozstrzyganie konfliktów klasyfikacji.....	106
3.1.6. Wybrane miary jakości .....	109
3.1.7. Zastosowanie miar do parametryzacji modelu zbiorów przybliżonych.....	113
3.1.8. Ocena statystycznej istotności reguł decyzyjnych.....	119
3.1.9. Pomiar siły interwencji reprezentowanej przez regułę .....	123
3.2. Miary niedefiniowane bezpośrednio na podstawie tablicy kontyngencji .....	127
3.2.1. Długość reguły .....	127
3.2.2. Unikalnie pokrywane przykłady pozytywne .....	128
3.2.3. Podobieństwo reguł .....	129
3.2.4. Równomierność rozkładu pokrywanych przykładów pozytywnych .....	133
3.2.5. Miary złożone .....	135
3.3. Ocena zbioru reguł .....	141
<b>4. Efektywność i własności wybranych miar oceny reguł .....</b>	<b>145</b>
4.1. Rola miary w pokryciowym algorytmie indukcji.....	148
4.2. Badanie efektywności miar w pokryciowym algorytmie indukcji reguł .....	150
4.2.1. Identyfikacja miar najbardziej efektywnych.....	155

4.2.2. Miara złożona i adaptacyjna metoda doboru miary .....	159
4.2.3. Porównanie z innymi algorytmami pokryciowymi.....	165
4.3. Badanie podobieństwa miar ze względu na porządek reguł.....	172
4.4. Analiza równoważności i własności miar.....	177
4.4.1. Równoważność .....	177
4.4.2. Własności.....	182
4.5. Omówienie zbioru miar najbardziej efektywnych.....	186
<b>5. Wybrane metody przycinania reguł decyzyjnych i ich zbiorów.....</b>	<b>193</b>
5.1. Agregacja reguł .....	195
5.1.1. Agregacja przez łączenie zakresów warunków elementarnych.....	196
5.1.2. Agregacja na podstawie otoczki wypukłej, zawierającej łączone reguły .....	199
5.1.3. Badanie efektywności algorytmów agregacji .....	207
5.2. Redefinicja reguł na podstawie oceny ważności warunków elementarnych.....	215
5.2.1. Ocena ważności warunków elementarnych.....	215
5.2.2. Algorytm redefinicji .....	218
5.2.3. Badanie efektywności algorytmu redefinicji .....	221
5.3. Filtracja reguł.....	228
<b>6. Przykłady zastosowań algorytmów indukcji reguł decyzyjnych.....</b>	<b>237</b>
6.1. Prognozowanie zagrożeń sejsmicznych w kopalniach .....	238
6.1.1. Definicja zadania prognozy zagrożenia sejsmicznego .....	239
6.1.2. Budowa i weryfikacja jakości klasyfikatora .....	242
6.1.3. Analiza otrzymanych wyników .....	244
6.1.4. Zagadnienia związane z wdrażaniem .....	247
6.2. Prognozowanie czasu przeżycia pacjentów po przeszczepie szpiku kostnego.....	250
6.2.1. Pokryciowy algorytm indukcji reguł sterowany hipotezami definiowanymi przez użytkownika .....	252
6.2.2. Reguły decyzyjne jako wzorce przeżycia .....	255
6.2.3. Analiza danych .....	258
6.3. Opis grup genów za pomocą reguł decyzyjnych .....	267
6.3.1. Definicja problemu .....	267
6.3.2. Dostosowanie algorytmu indukcji reguł do specyfiki problemu .....	271
6.3.3. Kryteria i sposób oceny reguł .....	273
6.3.4. Metoda redukcji atrybutów biorąca pod uwagę semantykę ich wartości .....	276
6.3.5. Analiza danych .....	278
<b>7. Podsumowanie.....</b>	<b>281</b>
<b>Bibliografia .....</b>	<b>289</b>
<b>Dodatek A. Wykresy miar jakości.....</b>	<b>313</b>
<b>Dodatek B. Przykład indukcji reguł regresyjnych.....</b>	<b>323</b>
<b>Spis rysункów.....</b>	<b>327</b>
<b>Spis tabel.....</b>	<b>329</b>
<b>Streszczenie.....</b>	<b>333</b>

## CONTENTS

<b>1. Introduction.....</b>	<b>9</b>
<b>2. Decision rule induction and rule-based classifiers.....</b>	<b>17</b>
2.1. Example set representation .....	18
2.2. Decision rules .....	22
2.3. Sequential covering rule induction algorithm.....	26
2.4. The q-ModLEM and RMatrix algorithms .....	40
2.5. Rule-based classifier.....	49
2.6. Classifiers evaluation.....	52
2.6.1. Classifier evaluation measures.....	52
2.6.2. Methods of experimental evaluation.....	58
2.6.3. Methods of classification abilities comparison.....	60
2.7. Development of rule induction methods .....	64
2.7.1. Improvement of classification and description abilities .....	64
2.7.2. Problems with the size and quality of data .....	69
<b>3. Objective measures for rule and rule set evaluation .....</b>	<b>73</b>
3.1. Measures defined basing on the contingency table .....	75
3.1.1. Value space.....	77
3.1.2. Properties .....	79
3.1.3. Properties required for decision rule evaluation .....	87
3.1.4. Equivalence and similarity with respect to the rule order.....	98
3.1.5. Equivalence with respect to classification conflicts resolving .....	106
3.1.6. Selected quality measures .....	109
3.1.7. Application for rough sets parameterization.....	113
3.1.8. Evaluation of decision rule statistical significance.....	119
3.1.9. Measuring effects of interventions based on a decision rule .....	123
3.2. Measures not defined directly based on the contingency table .....	127
3.2.1. Rule length.....	127
3.2.2. Positive examples covered uniquely.....	128
3.2.3. Similarity of rules .....	129
3.2.4. Regularity of the covered positive examples distribution.....	133
3.2.5. Compound measures.....	135
3.3. Rule set evaluation .....	141
<b>4. Efficiency and properties of selected rule evaluation measures.....</b>	<b>145</b>
4.1. Role of a measure in the sequential covering rule induction algorithm .....	148
4.2. Study of the efficiency of measures in the sequential covering rule induction algorithm.....	150
4.2.1. Identification of the most efficient measures.....	155

4.2.2. The compound measure and the adaptive method of measure selection .....	159
4.2.3. Comparison with other sequential covering algorithms .....	165
4.3. Study of the measures similarity with respect to the rule order .....	172
4.4. Analysis of measures equivalence and properties .....	177
4.4.1. Equivalence.....	177
4.4.2. Properties .....	182
4.5. Discussion on the most efficient measures .....	186
<b>5. Selected methods of rule and rule set pruning.....</b>	<b>193</b>
5.1. Rule aggregation.....	195
5.1.1. Aggregation through joining of elementary conditions .....	196
5.1.2. Aggregation on the basis of the convex hull containing joined rules .....	199
5.1.3. Study of the efficiency of aggregation algorithms.....	207
5.2. Rule redefinition based upon the elementary conditions importance evaluation .....	215
5.2.1. Evaluation of elementary conditions importance .....	215
5.2.2. The redefinition algorithm .....	218
5.2.3. Study of the efficiency of the redefinition algorithm.....	221
5.3. Rule filtration.....	228
<b>6. Examples of application of decision rule induction algorithms.....</b>	<b>237</b>
6.1. Forecasting seismic hazards in coal mines .....	238
6.1.1. Definition of seismic hazard forecasting problems .....	239
6.1.2. Developing and evaluation of classifiers.....	242
6.1.3. Analysis of the obtained results .....	244
6.1.4. Implementation issues .....	247
6.2. Survival time forecasting after bone marrow transplantation.....	250
6.2.1. Sequential covering rule induction algorithm controlled by hypothesis defined by the user .....	252
6.2.2. Decision rules as the survival patterns.....	255
6.2.3. Data analysis .....	258
6.3. Gene group description by means of decision rules .....	267
6.3.1. Definition of the problem .....	267
6.3.2. Adjusting rule induction algorithm to the specific charakter of the problem .....	271
6.3.3. Criteria and the method of rule evaluation .....	273
6.3.4. Attribute reduction considering the semantics of the attributes values .....	276
6.3.5. Data analysis .....	278
<b>7. Summary.....</b>	<b>281</b>
<b>Bibliography .....</b>	<b>289</b>
<b>Appendix A. Graphs of quality measures.....</b>	<b>313</b>
<b>Appendix B. An example of regression rule induction.....</b>	<b>323</b>
<b>List of Figures.....</b>	<b>327</b>
<b>List of Tables .....</b>	<b>329</b>
<b>Abstract.....</b>	<b>335</b>

## 1. WSTĘP

W okresie ostatnich dwudziestu lat metody eksploracji (zgłębiania) danych znalazły szerokie zastosowanie w wielu dziedzinach działalności człowieka. Powszechna dostępność komputerów o stosunkowo dużej mocy obliczeniowej umożliwia wykorzystywanie coraz doskonalszych metod eksploracji danych w coraz nowszych obszarach działalności człowieka.

Eksploracja danych jest procesem wieloetapowym i najczęściej iteracyjnym, wymagającym od użytkownika nie tylko umiejętności posługiwania się określonymi metodami analitycznymi, ale również wiedzy na temat konkretnego obszaru zastosowania. Najbardziej efektywny sposób realizacji zadania eksploracji danych polega na utworzeniu zespołu składającego się z analityka oraz eksperta dziedzinowego.

Aby proces eksploracji danych uczynić przejrzystym, opracowano kilka metodyk opisujących jego etapy i ich wzajemne powiązania. Metodyki te są najczęściej związane z konkretnymi dostawcami informatycznych systemów eksploracji danych. Najpopularniejsze z nich to: CRISP-DM (m.in. SPSS-IBM, Terradata); SEMMA (SAS Institute) i nieco ogólniejsza od niej Six-Sigma (promowana przez Statsoft) oraz Virtuous Cycle of Data Mining [24]. We wszystkich tych sposobach postępowania możemy znaleźć etapy polegające na: zdefiniowaniu celu, w jakim prowadzona będzie eksploracja, przygotowaniu i wstępnomu przetworzeniu danych podlegających analizie, modelowaniu (czyli zasadniczej analizie danych), ocenie jakości modelu i interpretacji uzyskanych wyników, wdrożeniu wyników do procesów biznesowych użytkownika. Ostatni z tych etapów obejmuje również nadzorowanie wiarygodności wdrożonego modelu.

Najpopularniejsze metody stosowane w etapie modelowania to [339]: grupowanie danych, indukcja drzew, indukcja reguł, analiza podobieństwa i sąsiedztwa, maszyny wektorów podpierających oraz wnioskowanie zgodnie z regułą Bayesa. Do grupy metod eksploracji danych włącza się także niektóre metody statystyczne (np. metody regresji, analiz korelacji).

W zależności od zastosowanej metody analitycznej uzyskujemy różne formy reprezentacji wiedzy, jaką udało się odkryć na podstawie danych. Drzewa i reguły stanowią reprezentację, która uważana jest za najbardziej zbliżoną do sposobu zapisu wiedzy przez

człowieka. Z tego też powodu algorytmy indukcji drzew i reguł są najczęściej stosowane w tych wszystkich zadaniach eksploracji danych, w których jedną z ważnych cech modelu danych jest jego czytelność.

Specjalnym rodzajem reguł są reguły decyzyjne (patrz podrozdział 2.2) [220]. Są to wyrażenia warunkowe, reprezentujące lokalne zależności pomiędzy wartościami cech (atrybutów) charakteryzujących analizowany zbiór danych, przy czym cecha znajdująca się w konkluzji reguły decyzyjnej jest ustalona i nazywa się ją atrybutem decyzyjnym. Pozostałe cechy nazywane są atrybutami warunkowymi.

Reguły decyzyjne definiowane są dla celów opisowych i klasyfikacyjnych [183, 284]. Dla perspektywy opisowej najbardziej interesujący będzie zbiór złożony z reguł reprezentujących zależności nietrywialne i użyteczne dla użytkownika. Dla perspektywy klasyfikacyjnej najbardziej pożądany będzie zbiór reguł, pozwalający z jak największą dokładnością wnioskować o wartości atrybutu decyzyjnego na podstawie informacji o wartościach atrybutów warunkowych. Dla perspektywy klasyfikacyjnej istotny jest także algorytm decyzyjny, przekładający decyzje proponowane przez poszczególne reguły na ostateczną decyzję, podejmowaną przez klasyfikator. Metody indukcji reguł decyzyjnych, podobnie jak inne metody eksploracji danych, charakteryzują się obciążeniem indukcyjnym [191, 194]. Wysoka skuteczność poprawnej klasyfikacji obiektów, na podstawie których reguły utworzono, nie musi przekładać się na poprawność klasyfikacji dotychczas nieznanych obiektów. Zadaniem algorytmu decyzyjnego jest m.in. rozstrzyganie sytuacji konfliktowych, w których kilka reguł rekomenduje przyporządkowanie klasyfikowanemu obiekowi różnych wartości atrybutu decyzyjnego.

Należy podkreślić, że w zadaniach indukcji reguł cel opisowy zawsze jest istotny, również wtedy, gdy reguły definiowane są dla celów klasyfikacyjnych. To właśnie czytelność modelu danych, a nie skuteczność klasyfikacji jest szczególnie akcentowaną cechą klasyfikatorów regułowych [52, 191, 223]. Gdyby kierować się jedynie zdolnościami klasyfikacyjnymi, doszłibyśmy do wniosku, że wiele innych metod (np. maszyny wektorów podpierających, systemy neuronowo-rozmyte, rodziny klasyfikatorów) przewyższają pod tym względem klasyfikatory regułowe.

Indukcja reguł dla celów klasyfikacyjnych jest zazwyczaj ukierunkowana na wyznaczenie jak najmniej licznego zbioru, złożonego z reguł jak najbardziej dokładnych i ogólnych oraz o jak najprostszej budowie [87, 191, 284]. Pozwala to w syntetyczny sposób opisać dane, unikać problemu nadmiernego dopasowania oraz uzyskać stosunkowo wysoką dokładność klasyfikacji. Algorytmy indukcji reguł dla celów opisowych ukierunkowują proces indukcji na wyznaczenie reguł spełniających pewne ustalone przez użytkownika wymagania jakości.

Dla skutecznego stosowania algorytmów indukcji reguł niebagatelne znaczenie mają metody oceny jakości reguł [3, 8, 40, 90, 143, 132, 340, 251]. Ocena wykonywana jest

zarówno podczas indukcji (etap modelowania), jak i po wyznaczeniu reguł (etap oceny i interpretacji wyników). Ocena realizowana jest za pomocą wskaźników, które w sposób numeryczny charakteryzują jakość reguł w świetle dostępnego zbioru danych lub subiektywnych preferencji użytkownika. W literaturze zdefiniowano wiele wskaźników umożliwiających ocenę klasyfikacyjnych i opisowych zdolności reguł. Jedna grupa wskaźników przeznaczona jest do nadzorowania procesu indukcji [8, 40, 90, 251], drugą grupę stanowią wskaźniki służące do oceny opisowych zdolności reguł [83, 106, 128, 132, 151, 179, 187, 240, 329]. Wiele z nich można stosować do realizacji obu tych zadań [3, 143, 132, 340].

Wskaźniki umożliwiające ocenę reguł znane są jako miary ich jakości, atrakcyjności i użyteczności i przeznaczone są nie tylko do oceny reguł decyzyjnych. Umożliwiają one również ich ocenę m.in. z punktu widzenia dokładności, ogólności, złożoności, unikalności, aplikowalności itd. Wykorzystanie odpowiedniej miary lub zbioru miar do nadzorowania procesu indukcji i/lub selekcji reguł jest jednym z kluczowych czynników decydujących o ich zdolnościach klasyfikacyjnych i opisowych [8, 42, 143, 251, 264]. Dostosowanie odpowiedniej miary jakości do celu analizy, charakterystyki zbioru danych, a także metody indukcji i/lub selekcji reguł nie jest zadaniem trywialnym. Szczególnie trudnym problemem jest złożona, wielokryterialna ocena reguł, w której pod uwagębrane są wartości wskaźników oceniających je z różnych, często antagonistycznych punktów wiedzenia [1, 3, 117, 178].

Prace badawcze nad miarami jakości reguł prowadzone są w trzech zasadniczych kierunkach:

- definiowanie i badanie efektywności miar jakości [8, 42, 90, 143, 251, 264, 267];
- definiowanie i analiza własności miar oraz badanie ich wzajemnych powiązań [27, 43, 106, 112, 113, 307];
- stosowanie miar w algorytmach indukcji i optymalizacji reguł; w szczególności opracowanie strategii dostosowujących miarę jakości do celu analizy i specyfiki zbioru danych [8, 143, 253, 264].

Z perspektywy opisowej nadzawanym celem prowadzonych badań jest opracowanie metod indukcji reguł interesujących, a także opracowanie miar i metod umożliwiających wskazanie reguł najbardziej interesujących [169, 181, 285, 328]. Dla perspektywy klasyfikacyjnej nadzawanym celem badań jest poprawa zdolności klasyfikacyjnych wyznaczonych reguł bez znaczącego pogorszenia ich zdolności opisowych.

Poprawę opisowych i klasyfikacyjnych zdolności reguł decyzyjnych uzyskuje się także przez stosowanie technik, które ogólnie można nazwać technikami optymalizacji. Reguły optymalizowane są w trakcie indukcji oraz po jej zakończeniu [41, 86]. Optymalizacja w trakcie indukcji realizowana jest w obrębie algorytmu wyznaczającego reguły. Optymalizacja po zakończeniu indukcji może być realizowana przez algorytmy niezależne od

algorytmu indukcji. Najczęściej optymalizacja kojarzona jest z tzw. przycinaniem (ang. *pre-pruning, post-pruning*). Algorytmy przycinania uruchamiane po indukcji reguł ukierunkowane są na upraszczanie i eliminację reguł zbędnych, a także na optymalizację zapisu reguł, przez wyrażenie ich w języku o większych zdolnościach opisowych (o większej ekspresji) [5, 64, 174, 197, 224, 244, 248, 265, 266]. Optymalizacja ukierunkowana jedynie na maksymalizację zdolności klasyfikacyjnych dostępnego zbioru reguł może prowadzić do niepożądanej sytuacji, w której zbiór reguł charakteryzuje się wysoką poprawnością klasyfikacji, ale tworzących go reguły nie można oddziennie traktować jako użytecznych wzorców. Najczęściej jednak optymalizacja przynosi pożądany efekt, jakim jest podniesienie jakości reguł i uproszczenie regułowego modelu danych.

Niezależne dziedzinowo metody indukcji, optymalizacji i oceny reguł stosowane są w różnych obszarach zastosowań (np. medycynie, przemyśle, marketingu, telekomunikacji itd.). Rozwiązania najbardziej efektywne uzyskuje się jednak, dostosowując proces indukcji i wskaźniki umożliwiające ocenę reguł do specyfiki zadania oraz wymagań użytkownika. Tworzenie rozwiązań zorientowanych dziedzinowo jest obecnie najbardziej dynamicznym kierunkiem prac badawczych, dotyczących nie tylko indukcji reguł, ale również wszystkich metod eksploracji danych. Świadczy o tym duża popularność i wysoka „cytowalność” periodyków ukierunkowanych na prezentowanie prac badawczych, związanych z zastosowaniami praktycznymi (np. Applied Soft Computing, Engineering Applications of Artificial Intelligence, Expert Systems with Applications).

## Cele pracy

Niniejszą monografię można podzielić na trzy części, w których kolejno omawiane są zagadnienia związane z oceną reguł decyzyjnych, ich przycinaniem oraz zastosowaniami praktycznymi. Najwięcej miejsca poświęcono własnościom i efektywności miar dokonujących obiektywnej, a więc nieukierunkowanej na żadną specyficzną dziedzinę zastosowań, oceny reguł decyzyjnych. Szczególną uwagę skierowano na miary definiowane na podstawie tablicy kontyngencji. Poniżej przedstawiono główne cele pracy; są to:

1. zebranie, ujednolicenie i analiza własności miar definiowanych na podstawie tablicy kontyngencji; określenie zestawu własności pożądanego dla miar przeznaczonych do oceny reguł decyzyjnych, w szczególności zdefiniowanie minimalnego zbioru tych własności;
2. przedstawienie ram teoretycznych, umożliwiających badanie równoważności i podobieństwa miar definiowanych na podstawie tablicy kontyngencji; zdefiniowanie warunków równoważności miar, określenie zależności pomiędzy miarami równoważnymi, zbadanie wzajemnych powiązań pomiędzy własnościami miar;

3. przedstawienie miar dokonujących oceny reguł z punktu widzenia ich dokładności, ogólności, statystycznej istotności, efektywności interwencji, złożoności, wzajemnego podobieństwa oraz równomierności pokrywania przykładów; przedstawienie miar złożonych, dokonujących wielokryterialnej oceny reguł;
4. zbadanie efektywności i podobieństwa szerokiego spectrum miar definiowanych na podstawie tablicy kontyngencji i przeznaczonych do nadzorowania indukcji reguł, realizowanej za pomocą algorytmu pokryciowego; identyfikacja zbioru miar najbardziej efektywnych; przedstawienie metody dostosowującej miarę nadzorującą proces indukcji do kryterium oceny klasyfikatora i konkretnego zbioru przykładów treningowych;
5. zdefiniowanie algorytmów przycinania, ukierunkowanych na łączenie i redefiniowanie reguł, a także na eliminację reguł zbędnych;
6. przedstawienie przykładów praktycznych zastosowań metod indukcji i oceny reguł decyzyjnych; w szczególności zaprezentowanie zorientowanych dziedzinowo modyfikacji tych metod.

W ten sposób rozważania prowadzone w monografii dotyczą trzech etapów regułowo zorientowanego procesu eksploracji danych: modelowania, oceny wyników oraz wdrożenia modelu danych.

### **Układ pracy**

Monografia składa się z 7 rozdziałów. Rozdziały 1 i 7 stanowią wstęp i podsumowanie pracy. Rozdział 2 w większości ma charakter przeglądowy. Omówiono w nim tematykę indukcji reguł, szczególną uwagę poświęcając algorytmom pokryciowym. W podrozdziale 2.4 zaprezentowano algorytm pokryciowy, który stanowi podstawę do badania efektywności miar jakości, prezentowanych w podrozdziale 3.1.6. Reguły uzyskane za pomocą tego algorytmu są danymi wejściowymi dla przedstawionych w rozdziale 5 algorytmów przycinania. Na zakończenie rozdziału 2 omówiono metody oceny i porównywania jakości klasyfikatorów regułowych.

Rozdział 3 poświęcony jest miarom oceny jakości reguł decyzyjnych. Specjalną uwagę poświęcono miarom definiowanym na podstawie tablicy kontyngencji. Dokonano przeglądu i uporządkowano informacje na temat własności definiowanych dla tego typu miar (podrozdział 3.1.2). Na tym tle zaprezentowano minimalne zbiory własności, jakimi powinny charakteryzować się miary oceniające jakość reguł decyzyjnych (podrozdział 3.1.3). Rozdzielono własności pożądane przez miary, stosowane jedynie do nadzorowania procesu indukcji reguł, od własności przeznaczonych dla miar oceniających zdolności opisowe reguł. Wprowadzono pojęcia równoważności miar ze względu na sposób rozstrzygania konfliktów klasyfikacji oraz rozróżniono ich względową i bezwzględną równoważność z punktu widzenia

definiowanego przez nie porządku reguł (podrozdział 3.1.4). Wprowadzono także definicję podobieństwa miar ze względu na porządek reguł. Stwierdzenia i uwagi zawarte w rozdziale 3 wskazują na powiązania pomiędzy różnymi typami równoważności, pozwalając też wnioskować o własnościach jednej miary na podstawie własności miar do niej równoważnych. Daje to podstawy do analizy i kategoryzacji miar jakości, których w monografii nie analizowano. W podrozdziale 3.1.7 pokazano także możliwość wykorzystania miar jakości definiowanych na podstawie tablicy kontyngencji do definiowania tzw. przybliżonych opisów pojęć. Rozdział 3 dotyczy również statystycznej oceny reguł (podrozdział 3.1.8), pomiaru siły interwencji reprezentowanej przez reguły (podrozdział 3.1.9), a także miar niedefiniowanych bezpośrednio na podstawie tablicy kontyngencji (podrozdział 3.2). Omówiono możliwości złożonej oceny reguł, w szczególności przedstawiono propozycję wykorzystania metody estymacji funkcji użyteczności do definiowania miary złożonej, dokonującej wielokryterialnej oceny reguł (podrozdział 3.2.5). W ostatniej części rozdziału 3 przedstawiono miary wykorzystywane do oceny zbiorów reguł. Stanowią one uzupełnienie dla wskaźników mierzących efektywność klasyfikatora regułowego. Uzupełnieniem dla rozdziału 3 jest dodatek A, w którym umieszczono wykresy wybranych miar jakości definiowanych na podstawie tablicy kontyngencji.

Rozdział 4 w znacznej części poświęcony jest eksperymentalnej ocenie efektywności i podobieństwa miar jakości definiowanych na podstawie tablicy kontyngencji. Podstawą do badań był pokryciowy algorytm indukcji reguł oraz 34 miary stosowane w różnych algorytmach indukcji, przycinania i selekcji reguł. Na podstawie analizy 34 zbiorów danych zidentyfikowano grupy miar najbardziej efektywnych z punktu widzenia trzech kryteriów oceny jakości zbioru reguł. Kryteria te biorą pod uwagę zarówno klasyfikacyjne, jak i opisowe zdolności generowanych reguł. Miary najbardziej efektywne stały się podstawą do przedstawienia propozycji miar złożonych oraz adaptacyjnej metody doboru miary nadzorującej proces indukcji reguł (podrozdział 4.2.2). Aby zweryfikować stabilność uzyskanych wyników, przebadano kolejnych 14 zbiorów danych. Przeprowadzono statystyczną analizę różnic pomiędzy klasyfikatorami regułowymi, utworzonymi na podstawie różnych miar.

Analiza 48 zbiorów danych o różnorodnej charakterystyce pozwoliła również na wskazanie zależności pomiędzy miarą jakości i sposobem jej wykorzystania w pokryciowym algorytmie indukcji reguł a liczbą, dokładnością, ogólnością, wzajemnym podobieństwem i złożonością wyznaczanych reguł. Badania eksperymentalne pozwoliły na zidentyfikowanie grup miar porządkujących reguły w podobny sposób. W rozdziale 4 przeanalizowano wzajemne korelacje pomiędzy porządkami reguł w obrębie grup i pomiędzy nimi. Ostatnia część rozdziału 4 zawiera analizę własności miar. Przebadano 36 miar ze

względu na 8 własności pożądanych dla miar oceniających reguły decyzyjne. Wskazano również na możliwości rozluźnienia warunków definiujących niektóre z rozważanych własności. Uzupełnieniem badań zawartych w rozdziale 4 jest dodatek B, w którym zidentyfikowaną grupę miar najbardziej efektywnych zastosowano do nadzorowania procesu indukcji reguł regresyjnych.

Rozdział 5 traktuje o przycinaniu reguł i koncentruje się na metodach stosowanych po zakończeniu indukcji. Przedstawiono algorytmy łączenia (agregacji) i redefinicji reguł oraz eliminacji reguł zbędnych z punktu widzenia możliwości rozpoznawania przykładów treningowych lub dokładności klasyfikacji. Algorytmy agregacji i redefinicji zmieniają język reprezentacji reguł. Dzięki temu zabiegowi reguły mogą opisywać nieco bardziej złożone zależności niż umożliwia to standardowy język reprezentacji. Podstawą agregacji jest algorytm wyznaczania otoczki wypukłej, zawierającej łączone reguły. Podstawą do redefinicji są informacje na temat oceny ważności tzw. warunków elementarnych, z których zbudowane są przesłanki reguł. Ze względu na złożoność obliczeniową standardowych metod oceny ważności warunków, zaproponowano uproszczony sposób tej oceny.

W rozdziale 5 przedstawiono także 4 algorytmy dokonujące eliminacji reguł zbędnych. Podstawą dla dwóch z tych algorytmów jest spostrzeżenie, że rezultatem indukcji są często zbiory reguł pokrywające podobne obszary w przestrzeni atrybutów. Odpowiednia selekcja reguł powinna zredukować ich liczbę bez utraty możliwości opisywania przykładów treningowych. Dla dwóch kolejnych algorytmów podstawową informacją na temat użyteczności reguł jest trafność klasyfikacji utworzonego na ich podstawie klasyfikatora.

Rozdział 6 zawiera przykłady zastosowania metod indukcji, przycinania i oceny reguł w nowych obszarach. Omówiono metodykę tworzenia i wdrażania klasyfikatora regułowego, przeznaczonego do oceny zagrożeń sejsmicznych. Przedstawiono także zastosowanie medyczne, w którym powiązano indukującą regułę z analizą przeżycia oraz zaprezentowano modyfikację pokryciowego algorytmu indukcji, umożliwiającą wykorzystanie hipotez definiowanych przez użytkownika. Pozwoliło to na prowadzenie interaktywnej konstrukcji klasyfikatora.

Ostatnie z omawianych zastosowań dotyczy funkcjonalnego opisu genów za pomocą reguł decyzyjnych. Badania prowadzone w tym obszarze pozwoliły na: dostosowanie algorytmu indukcji do hierarchicznej struktury atrybutów warunkowych, opracowanie metody redukcji atrybutów sterowanej semantyką ich wartości oraz zaproponowanie złożonej miary oceny reguł.

### **Podziękowania**

Pierwsze podziękowania kieruję do śp. Prof. Adama Mrózka, który zainteresował mnie metodami maszynowego uczenia się i teorią zbiorów przybliżonych.

Większość wyników prezentowanych w monografii jest rezultatem badań prowadzonych przeze mnie w latach 2002–2006 i 2009–2012. W okresie tym miałem przyjemność współpracować z wieloma osobami z Instytutu Informatyki Politechniki Śląskiej i Instytutu Technik Innowacyjnych EMAG. Szczególne podziękowania kieruję do dr inż. Aleksandry Grucy, dr. inż. Marcina Michalaka, mgr. inż. Łukasza Wróbla, mgr. inż. Adama Gudysia, a także do prof. Andrzeja Polańskiego. Podziękowania składam również tym wszystkim osobom z Instytutu Informatyki Politechniki Śląskiej i Instytutu Technik Innowacyjnych EMAG, które udzielili mi wsparcia i pomocy w bardzo trudnych dla mnie latach 2006–2008.

Chciałbym podziękować również Profesorowi Andrzejowi Skowronowi za uwagi dotyczące konspektu niniejszej publikacji.

Recenzentom monografii – Profesor Alicji Wakulicz-Dei oraz Profesorowi Jarosławowi Stepaniukowi – składam podziękowania za uwagi, dzięki którym końcowa wersja pracy jest lepsza od wersji pierwotnej.

Badania opisane w monografii współfinansowane były m.in. przez: Komitet Badań Naukowych (5T12A00123), Europejski Fundusz Węgla i Stali (EUREKA E!3353; RFCR-CT-2010-00005 MINFIREX) i Narodowe Centrum Nauki (DEC-2011/01/D/ST6/07007).

Na zakończenie szczególnie gorące podziękowania kieruję do mojej żony Beaty oraz do dzieci – Magdy i Wojtka.

## **2. INDUKCJA REGUŁ DECYZYJNYCH I KLASYFIKATORY REGUŁOWE**

Indukcja reguł decyzyjnych należy do dziedziny maszynowego uczenia się (ang. *machine learning*), realizowanego na podstawie paradygmatu uczenia z wykorzystaniem przykładów [191]. W literaturze przedmiotu można spotkać wiele definicji maszynowego uczenia się [52, 190, 194]. W latach 1983–1997 podano najbardziej znane definicje uczenia się maszyn; ich autorami byli kolejno: Herbert Simon (1983), Marvin Minsky (1986), Ryszard Michalski (1986), Donald Michie (1991) oraz Tom Mitchell (1997). Dla celów niniejszej publikacji najbardziej odpowiednia jest definicja Ryszarda Michalskiego, która w uproszczeniu brzmi następująco: *uczenie się to konstruowanie i zmiana reprezentacji doświadczanych faktów* oraz, będąca uszczegółowieniem propozycji Michalskiego, definicja Donalda Michiego: *system uczący się wykorzystuje zewnętrzne dane empiryczne w celu tworzenia i aktualizacji podstaw dla udoskonalonego działania na podobnych danych w przyszłości oraz wyrażania tych podstaw w zrozumiałej i symbolicznej postaci.*

System uczący się to system, który zmienia swoje wewnętrzne parametry, tak aby rozpoznać i opisać dane. Definicja zaproponowana przez Michiego zakłada, że reprezentacja parametrów systemu jest symboliczna. Takie podejście do maszynowego uczenia pozwala sklasyfikować tę dziedzinę jako gałąź sztucznej inteligencji (ang. *artificial intelligence*). Klasyfikacja ta nie jest jednak precyzyjna, gdyż wiele systemów uczących się używa numerycznej reprezentacji parametrów. Lepszym rozwiązaniem jest rozważanie maszynowego uczenia się jako działu inteligencji obliczeniowej (ang. *computational intelligence*) [69]. Inteligencja obliczeniowa jest dziedziną informatyki zajmującą się problemami trudnymi do modelowania w sposób ścisły, do rozwiązywania których nie istnieją obliczeniowo efektywne algorytmy. Definicja inteligencji obliczeniowej obejmuje systemy uczące się o symbolicznej i niesymbolicznej reprezentacji parametrów.

Najbardziej popularny podział [52, 191] pozwala wyróżnić algorytmy uczenia się: poprzez zapamiętywanie, na podstawie przykładów, na podstawie wyjaśnień, z wykorzystaniem grupowania pojęciowego, przez analogię oraz ze wzmacnieniem. Uczenie się na podstawie przykładów znane jest również pod pojęciem *indukcji*. Algorytmy uczące

się mogą również zostać podzielone ze względu na dostępną informację trenującą; tutaj najczęściej rozróżnia się uczenie nadzorowane i nienadzorowane.

Uczenie nadzorowane na podstawie przykładów polega na znalezieniu hipotezy lub zbioru hipotez opisujących pewne pojęcia, które reprezentowane są przez zbiór przykładów. Termin *pojęcie* rozumiany jest jako zbiór przykładów mających charakterystyczne własności, które odróżniają go od innych pojęć znajdujących się w zbiorze przykładów. Tworząc hipotezę opisującą dane pojęcie, wszystkie przykłady będące reprezentantami tego pojęcia nazywa się przykładami pozytywnymi, a wszystkie pozostałe – przykładami negatywnymi lub kontrprzykładami.

Aby możliwe było podanie konkretyzacji algorytmu uczenia się opisów pojęć na podstawie przykładów, wymagane jest ustalenie: języka reprezentacji przykładów, języka reprezentacji hipotez, kryterium oceny jakości hipotez oraz strategii przeszukiwania przestrzeni możliwych rozwiązań.

Od języka reprezentacji hipotez w dużej mierze zależy to, czy spełniony zostanie silnie akcentowany przez Michalskiego postulat zrozumiałości [190, 191]. Postulat ten oznacza, że hipotezy zapisane są w czytelnej, zrozumiałej dla odbiorcy postaci i przez to mogą być przez niego interpretowane. Warto zauważyć, że opisy przykładów wyrażane w postaci hipotez mających cechę zrozumiałości mogą być traktowane jako wiedza, którą udało się odkryć. Z tego też względu algorytmy indukcji, których efektem działania są zrozumiałe dla człowieka hipotezy, stosowane są powszechnie w eksploracji danych (ang. *data mining*) [123, 192, 333].

Zrozumiałość hipotez leży u podstaw jeszcze jednej klasyfikacji metod maszynowego uczenia się. Metody i algorytmy uczące się hipotez o cechach zrozumiałości nazywane są symbolicznymi metodami uczenia się. Jeśli wynikiem działania algorytmu jest zestaw parametrów (np. zbiór liczb, zbiór równań matematycznych), których interpretacja bez dodatkowej wiedzy i pewnych założeń nie jest oczywista, to mamy do czynienia z niesymbolicznymi (subsztuczycznymi) metodami uczenia się.

Najpopularniejsze symboliczne reprezentacje hipotez to reprezentacje regułowe, grafowe (w szczególności drzewa decyzyjne) i formuły rachunku predykatów. Do niesymbolicznych reprezentacji najczęściej zalicza się wartości prawdopodobieństwa [282], sieci neuronowe, zbiory rozmyte i sieci neuronowo-rozmyte [186].

## 2.1. Reprezentacja zbioru przykładów

Do zapisu przykładów najczęściej stosowane są reprezentacje tabelaryczne. Reprezentacje te charakteryzują się prostotą i przejrzystością zapisu. W algorytmach indukcji reguł i drzew decyzyjnych przykłady zapisane są w tzw. tablicy decyzyjnej [220]. Każdy przykład jest w niej charakteryzowany poprzez wektor cech, nazywanych również

atrybutami. W tablicy decyzyjnej funkcjonują dwa rozłączne zbiory atrybutów. Pierwszy z nich to zbiór atrybutów warunkowych, drugi zbiór stanowią atrybuty decyzyjne. Tablicę złożoną z wielu atrybutów decyzyjnych można sprowadzić do tablicy zawierającej jeden atrybut decyzyjny. Każda wartość atrybutu decyzyjnego reprezentuje wtedy jedną ze wszystkich możliwych kombinacji wartości zastępowanych atrybutów decyzyjnych. Formalną definicję tablicy decyzyjnej zawierającej jeden wyróżniony atrybut decyzyjny  $d$  zawarto w definicji 2.1 [220].

**Definicja 2.1.** Tablicą decyzyjną  $\mathbf{DT}$  nazywamy parę  $(U, A \cup \{d\})$ , w której  $U$  jest zbiorem obiektów,  $A$  jest zbiorem atrybutów warunkowych oraz  $d$  jest atrybutem decyzyjnym.

Każdy z atrybutów  $a$ , należących do zbioru  $A \cup \{d\}$ , traktowany jest jako funkcja  $U \rightarrow V_a$ , przyporządkowująca każdemu obiekowi ze zbioru  $U$  pewną wartość należącą do zbioru  $V_a$ . Zbiór  $V_a$  jest zbiorem wartości atrybutu  $a$ . Przy takich założeniach zapis  $a(x)$  oznacza wartość, jaką obiekt  $x \in U$  charakteryzowany jest przez atrybut  $a$ . Zbiór wszystkich obiektów o identycznych wartościach atrybutu decyzyjnego nazywany jest klasą decyzyjną.

Przedstawiona definicja tablicy decyzyjnej funkcjonuje głównie w teorii zbiorów przybliżonych [220], jednak nic nie stoi na przeszkodzie, aby wykorzystywać ją również w maszynowym uczeniu się. Pomiędzy uczeniem się przez indukcję a teorią zbiorów przybliżonych występują różnice terminologiczne. W maszynowym uczeniu się obiekty tablicy decyzyjnej to przykłady, a klasy decyzyjne to pojęcia. W niniejszej monografii wykorzystana będzie terminologia z teorii zbiorów przybliżonych, jedynym odstępstwem będzie stosowanie terminu *przykład* zamiast *obiekt*.

Wartości atrybutu decyzyjnego determinują przynależność przykładów do klas decyzyjnych. W przypadku tworzenia opisu pewnej ustalonej klasy decyzyjnej wszystkie przykłady reprezentujące tę klasę traktowane są jako przykłady pozytywne, a wszystkie pozostałe – jako przykłady negatywne.

Jeśli w tablicy decyzyjnej znajdują się przykłady o identycznych wartościach wszystkich atrybutów warunkowych i różnych wartościach atrybutu decyzyjnego, to przykłady takie nazywa się sprzecznymi, a tablicę – sprzeczną tablicą decyzyjną. Tablica, która nie zawiera sprzecznych przykładów, nazywana jest tablicą niesprzeczną.

Standardowo w tablicy decyzyjnej definiuje się trzy typy atrybutów warunkowych [52]:

- nominalne – o skończonym zbiorze nieuporządkowanych wartości dyskretnych,
- porządkowe – o przeliczalnym zbiorze uporządkowanych wartości dyskretnych,
- ciągłe – o wartościach ze zbioru liczb rzeczywistych.

Pomiędzy wartościami atrybutów porządkowych istnieje jasno określony porządek. Oznacza to, że możliwe jest porównywanie wartości tych atrybutów za pomocą relacji  $=$ ,  $<$ ,  $>$ ,  $\leq$ ,  $\geq$ . Nie ma natomiast sensu wykonywanie na wartościach tych atrybutów operacji

arytmetycznych. Dziedziny atrybutów ciągłych definiowane są na skalach pomiarowych interwałowych, ilorazowych oraz absolutnych. W skali interwałowej różnice pomiędzy wartościami atrybutów mają sensowną interpretację. W skali ilorazowej sensowną interpretację mają także ilorazy wartości.

W wymienionych typach atrybutów zakłada się, że pomiędzy ich wartościami nie istnieje żaden porządek (atrybuty nominalne) lub porządek ten jest liniowy. W praktycznych zastosowaniach można spotkać się z sytuacją, w której wartości atrybutu tworzą pewną uporządkowaną strukturę. Przykładami takich struktur mogą być taksonomie lub ontologie [114]. Atrybuty typu strukturalnego charakteryzują się tym, że przykład może być opisany przez kilka wartości atrybutu. W szczególności, jeśli wartości atrybutu tworzą taksonomię, to przykład charakteryzowany przez wartość  $v$  charakteryzowany jest również przez wszystkie wartości będące w taksonomii potomkami  $v$ . Atrybut typu strukturalnego może zostać zastąpiony przez zbiór atrybutów nominalnych. Każdy atrybut z takiego zbioru odpowiada pojedynczemu węzlowi w strukturze. Zbiór wartości nowych atrybutów jest wtedy dwuargumentowy. Wartość 0 oznacza, że przykład nie jest charakteryzowany przez dany węzeł w strukturze, wartość 1 oznacza, że przykład jest przez dany węzeł charakteryzowany. W pewnych sytuacjach wykorzystanie informacji o strukturze wartości atrybutu może być pomocne w sterowaniu procesem indukcji reguł [156, 257, 315].

Definicja 2.1 zakłada, że zbiór przykładów jest nieuporządkowany. W wielu zastosowaniach przykłady stanowiące podstawę do indukcji mogą być uporządkowane. Szczególnym przypadkiem takiego porządku jest porządek temporalny, odzwierciedlający upływający czas. Informacja o czasie pozyskania danych stanowiących przykłady jest szczególnie ważna w przemysłowych i medycznych systemach monitorowania [18, 262, 276]. Tablica decyzyjna, w której przykłady uporządkowane są ze względu na upływający czas, nazywana jest temporalną tablicą decyzyjną [19, 152, 299]. W tablicy temporalnej każdy przykład ma unikalny identyfikator informujący o tym, jaką chwilę czasu opisuje ( $U = \{x_{t_1}, x_{t_2}, \dots, x_{t_m}\}$ ). W temporalnych tablicach decyzyjnych upływający czas ma charakter przedziałowy. Kolejne przykłady reprezentują następujące po sobie okna czasowe, a każde z nich zawiera zagregowany opis pewnego fragmentu rzeczywistości [19]. Identyfikator  $t+1$  wskazuje na okno czasowe, obejmujące okres pomiędzy końcem okna  $t$  a początkiem okna  $t+1$ . Zazwyczaj okna czasowe mają taką samą długość [19, 259, 262]. W pracach dotyczących indukcji tzw. reguł temporalnych [331] ciąg okien czasowych nazywany jest sekwencją zdarzeń.

Temporalna tablica decyzyjna jest szczególnym przypadkiem tzw. dynamicznego systemu informacyjnego [217, 323], w którym wraz z upływającym czasem może się zmieniać zarówno liczba przykładów, jak również liczba opisujących je atrybutów.

W temporalnych tablicach decyzyjnych informacja o upływającym czasie może zostać użyta do budowy nowych atrybutów, które przechowują informacje o dynamice zmian wartości atrybutów podstawowych. Wartości atrybutów dynamiki zmian przechowują informacje o zmianach i dynamice zmian wartości atrybutów podstawowych w kolejnych oknach czasowych. Zmiany wartości mogą być wyrażane za pomocą różnych funkcji (różnic, ilorazów, sum itd.). Dysponując informacjami o dynamice zmian wartości atrybutów podstawowych, możemy temporalną tablicę decyzyjną zamienić w standardową tablicę decyzyjną. Przekształcenie to jest bardzo proste i polega na usunięciu identyfikatorów okien czasowych przyporządkowanych przykładom. W praktycznych zastosowaniach bardzo często przydatna może okazać się również informacja o wcześniejszych wartościach danego atrybutu. W tablicy decyzyjnej pozbawionej identyfikatorów okien czasowych informację taką przechowywać będą tzw. opóźnienia.

**Definicja 2.2.** Założmy, że dana jest temporalna tablica decyzyjna  $\mathbf{DT}_t$ , w której  $U = \{x_{t1}, x_{t2}, \dots, x_{tn}\}$  oraz  $A = \{a_1, a_2, \dots, a_m\}$ . Opóźnieniem stopnia  $i$  atrybutu podstawowego  $a_j$  nazywamy atrybut  $a_j^{-i}$ , którego wartości definiowane są za pomocą wzoru (2.1):

$$a_j^{-i}(x_{tk}) = \begin{cases} ? & n \geq i \geq k \\ a_j(x_{t(k-i)}) & i < k \leq n \end{cases} \quad (2.1)$$

W definicji 2.2 symbol „?” oznacza nieokreśloną wartość atrybutu.

**Definicja 2.3.** Założmy, że dana jest temporalna tablica decyzyjna  $\mathbf{DT}_t$ , w której  $U = \{x_{t1}, x_{t2}, \dots, x_{tn}\}$  oraz  $A = \{a_1, a_2, \dots, a_m\}$ . Atrybutem dynamiki zmian atrybutu podstawowego  $a_j$  stopnia  $i$  oraz funkcji dynamiki  $f$  nazywamy atrybut  $\Delta a_j^{-i,f}$ , którego wartości definiowane są za pomocą wzoru (2.2):

$$\Delta a_j^{-i,f}(x_{tk}) = \begin{cases} ? & n \geq i \geq k \\ f(a_j(x_{t(k-i)}), \dots, a_j(x_{ti})) & i < k \leq n \end{cases} \quad (2.2)$$

**Przykład 2.1.** Założmy, że dana jest temporalna tablica decyzyjna  $\mathbf{DT}_t$ , w której  $U = \{x_{t1}, x_{t2}, \dots, x_{tn}\}$  oraz  $A = \{a_1, a_2, \dots, a_m\}$ . W tablicy  $\mathbf{DT}_t$  każdy przykład  $x_{ti}$  reprezentowany jest przez zbiór wartości  $a_1(x_{ti}), a_2(x_{ti}), \dots, a_m(x_{ti})$ . Po rozszerzeniu zbioru  $A$  o atrybuty  $a_1^{-1}, a_1^{-2}, a_2^{-1}, \Delta a_2^{-1,-}$  każdy przykład  $x_{ti}$  reprezentowany będzie przez zbiór wartości  $a_1(x_{ti}), a_2(x_{ti}), \dots, a_m(x_{ti}), a_1(x_{t(i-1)}), a_1(x_{t(i-2)}), a_2(x_{t(i-1)}), a_2(x_{ti}) - a_2(x_{t(i-1)})$ .

Wybór optymalnego zbioru opóźnień i atrybutów dynamiki jest problemem nietrywialnym i stosunkowo mało zbadanym. W praktyce zbiór opóźnień i atrybutów dynamiki dobierany jest na drodze eksperymentalnej.

Problematyka analizy danych temporalnych pochodzących z systemów monitorowania zagrożeń naturalnych była podejmowana przez autora m.in. w pracach [258, 259, 262]. Część z tych badań zostanie przedstawiona w rozdziale 6.

## 2.2. Reguły decyzyjne

Reguły stanowią reprezentację wiedzy, która uważana jest za najbardziej zbliżoną do sposobu zapisu wiedzy przez człowieka. Z tego też względu reguły uznawane są za najprostszy do interpretacji język reprezentacji hipotez będących wynikiem działania algorytmu uczącego się [191]. W najbardziej ogólnym przypadku regułę definiuje się jako wyrażenie postaci (2.3):

$$\text{jeżeli } \varphi, \text{ to } \psi. \quad (2.3)$$

W wyrażeniu (2.3)  $\varphi$  nazywana jest przesłanką, a  $\psi$  konkluzją reguły. Dla uproszczenia używa się zapisu  $\varphi \rightarrow \psi$ . Przesłanka  $\varphi$  traktowana jest jako pewna formuła logiczna, natomiast w zależności od postaci konkluzji możemy spotkać się z różnego rodzaju regułami. Jeśli konkluzja jest formułą logiczną, to mamy do czynienia z *regułą logiczną* [52, 67, 183, 284], jeśli konkluzja wskazuje na działania, jakie mają być podjęte w przypadku spełniania przesłanki, to mówimy o *regule decyzyjnej* [67, 183, 220]. W literaturze przyjęło się reguły logiczne nazywać również regułami decyzyjnymi, gdyż najczęściej konkluzja reguł logicznych zawiera formułę logiczną, informującą o przynależności przykładu do jednej z klas decyzyjnych. Część autorów stosuje termin *reguły klasyfikacyjne*, akcentując w ten sposób cel indukcji reguł [57, 76, 87].

Inne rodzaje reguł to m.in. reguły regresyjne [144, 260, 330], asocjacyjne [6], akcji [317], wzbraniające (ang. *inhibitory rules*) [61, 313] czy reguły decyzyjne definiowane dla wielokryterialnych problemów sortowania [30, 32, 284]. W tablicach decyzyjnych zawierających przykłady przeznaczone do indukcji reguł regresyjnych atrybut decyzyjny nie jest typu symbolicznego, ale ciągłego. W problemach sortowania klasy decyzyjne także uporządkowane są zgodnie z relacją preferencji.

W pracach Blachnika i Ducha można spotkać się z regułami konstruowanymi na podstawie tzw. prototypów (ang. *prototype-based rules*) [26]. Obszary pokrywane przez te reguły nie muszą być hipersześciianami, tak jak jest to w przypadku reguł logicznych. Reguły takie stanowią uogólnienie innych rodzajów reguł, w tym reguł logicznych.

Reguły decyzyjne stosowane są do opisu przykładów należących do tablicy decyzyjnej. Tablicę decyzyjną, na podstawie której wyznaczono reguły, nazywa się tablicą treningową.

Regułą decyzyjną nazywamy wyrażenie (2.4), które jest konkretyzacją (2.3):

$$\text{jeżeli } w_1 \wedge w_2 \wedge \dots \wedge w_j, \text{ to } d = v. \quad (2.4)$$

Przesłanka reguły składa się z koniunkcji warunków elementarnych. Aby przykład spełniał warunek elementarny, musi on spełniać zapisane w nim ograniczenia. Reguła (2.4) reprezentuje zależność mówiącą o tym, że przykład spełniający jednocześnie wszystkie ograniczenia zapisane w warunkach elementarnych należy do klasy decyzyjnej reprezentowanej przez wartość  $v$ .

Regułę regresyjną, w konkluzji której znajduje się wskazanie na wartość atrybutu decyzyjnego, możemy również zapisać jako wyrażenie (2.4). W regule takiej konkluzja informuje o przyporządkowaniu przykładowi spełniającemu ograniczenia zawarte w przesłance konkretnej wartości atrybutu decyzyjnego. Wartość ta jest liczbą rzeczywistą i nie jest interpretowana jako przynależność przykładu do klasy decyzyjnej.

Warunki elementarne mogą być proste lub złożone. Prosty warunek elementarny  $w$  definiowany jest jako wyrażenie o postaci  $a \text{ op } Za$ . W wyrażeniu tym  $a$  jest atrybutem warunkowym, traktowanym tutaj jako zmienna mogąca przyjmować wartości należące do dziedziny atrybutu  $a$ . Operator relacyjny  $\text{op}$  wskazuje na jeden z symboli relacyjnych należących do zbioru  $\{\leq, \geq, <, >, \neq, =, \in\}$ , a  $Za$  jest tzw. zakresem warunku i w zależności od użytego operatora relacyjnego jest wartością lub podzbiorem zbioru  $Va$ . Poza warunkami prostymi w przesłankach reguł mogą znajdować się również warunki złożone. Złożonym warunkiem elementarnym nazywać będziemy każdy warunek, który nie jest warunkiem prostym. Poniżej omówiono kilka rodzajów złożonych warunków elementarnych.

Przykładem złożonych warunków są wyrażenia o postaci  $a \text{ op } b$ , gdzie  $\text{op} \in \{\leq, \geq, <, >, =, \neq\}$  oraz  $a, b$  są dowolnymi atrybutami warunkowymi. Warunek interpretowany jest jako stwierdzenie, że pomiędzy wartościami atrybutów  $a$  i  $b$  zachodzi relacja wyrażona za pomocą operatora relacyjnego  $\text{op}$ . Przykładowo,  $a > b$  oznacza, że wartość atrybutu  $a$  jest większa od wartości atrybutu  $b$ . Warunek ten będą spełniać wszystkie przykłady, w których wartość atrybutu  $a$  będzie większa od wartości atrybutu  $b$ . Dla atrybutów ciągłych można podać bardziej złożoną postać warunku  $a \text{ op } b$ , będącą kombinacją liniową atrybutów warunkowych. Warunki o postaci  $s_1a_1 + s_2a_2 + \dots + s_la_l \text{ op } s$ , gdzie  $s, s_1, s_2, \dots, s_l \in \mathbb{R}$ ,  $a_1, a_2, \dots, a_l \in A$ , nazywane są skośnymi warunkami elementarnymi [205, 224, 265], gdyż występujące w nich równania hiperpłaszczyzn nie muszą być prostopadłe do osi żadnego z tworzących je atrybutów. Znane są także algorytmy indukcji reguł, w których złożone warunki elementarne definiowane są za pomocą krzywych wyższych stopni [21, 64].

Kolejnym przykładem złożonego warunku elementarnego jest wyrażenie  $\text{not } w$ , gdzie  $w$  jest prostym warunkiem elementarnym lub jednym z wymienionych powyżej warunków złożonych [67, 171, 266, 313].

W regułach decyzyjnych, definiowanych dla wielokryterialnych problemów sortowania [30, 32, 284], warunki elementarne mają postać  $a \leq v$  lub  $a \geq v$ . Atrybut  $a$  jest tzw. kryterium, którego wartości uporządkowane są zgodnie z relacją preferencji  $\succeq$  ( $\preceq$  oznacza odwróconą preferencję). Warunek  $a \geq v$  oznacza, że spełniające go przykłady uzyskują, ze względu na kryterium  $a$ , ocenę nie gorszą niż  $v$ . W problemach sortowania klasy decyzyjne także uporządkowane są zgodnie z relacją preferencji. Konkluzja reguły może przyjąć jedną z dwóch postaci:  $d \geq v$  lub  $d \leq v$ . W świetle przyjętych założeń interpretacja reguły zdefiniowanej dla wielokryterialnych problemów sortowania jest następująca: jeśli przykład spełnia wszystkie ograniczenia zapisane w przesłance, to przyporządkowywany jest do co najmniej (co najwyższej) tej klasy decyzyjnej, która wskazywana jest w konkluzji. Zgodnie z przyjętą przez nas definicją prostego warunku elementarnego warunki  $a \leq v$ ,  $a \geq v$  uznamy za złożone, chociaż sposób ich tworzenia podczas indukcji reguł nie odbiega znaczaco od sposobu definiowania warunków prostych  $a \leq v$ ,  $a \geq v$ .

Do złożonych warunków elementarnych można zaliczyć także propozycję przedstawioną przez Grabczewskiego [103]. W propozycji tej warunek może przyjąć postać  $dist(\cdot, x) < th$ , gdzie  $dist$  jest pewną ustaloną miarą odległości,  $x$  jest punktem w przestrzeni cech, a  $th$  jest wartością progową. Za złożone uznajemy również tzw. warunki  $M-z-N$ . W warunku  $M-z-N$   $M$  jest liczbą, a  $N$  jest zbiorem zawierającym proste lub złożone warunki elementarne. Spełnienie warunku  $M-z-N$  oznacza, że spełnionych jest jednocześnie  $M$  dowolnych warunków spośród tych zawartych w zbiorze  $N$ . Modyfikacje warunku  $M-z-N$  polegają na dodaniu przed liczbą  $M$  dodatkowego ograniczenia, np. *co najmniej*, *co najwyższej*. W terminologii angielskiej reguły zawierające warunki  $M-z-N$  nazywane są *M-of-N rules* [310].

Stosowanie złożonych warunków elementarnych prowadzi do indukcji reguł bardziej uniwersalnych, pokrywających większą liczbę przykładów. Niewątpliwie jednak stosowanie warunków złożonych wprowadza utrudnienia w interpretacji reguł. W pracach [265, 266] autor przedstawił własne propozycje, pozwalające na otrzymanie reguł zawierających złożone warunki elementarne. Zostaną one przedstawione w rozdziale 5.

O przykładach spełniających przesłankę mówi się, że pokrywają one regułę lub są pokrywane przez regułę. Zamiast zwrotu *reguła pokrywa przykłady* wymiennie może być stosowany zwrot *reguła opisuje przykłady*. Przykłady pozytywne, pokrywane przez regułę  $r$ , to przykłady należące do klasy decyzyjnej, wskazywanej w konkluzji  $r$ . Przykłady negatywne, pokrywane przez  $r$ , to przykłady nienależące do klasy decyzyjnej, wskazywanej w konkluzji  $r$ .

Przedstawiona definicja prostego warunku elementarnego i przykłady warunków złożonych mają charakter formuł logicznych, co więcej – są to formuły zdaniowe, które są

lub nie są spełniane przez przykłady (problem brakujących wartości atrybutów na razie pomijamy). Reguły zbudowane z warunków elementarnych, będących zdaniami w sensie logicznym, nazywane są regułami zdaniowymi (ang. *propositional rules*).

**Definicja 2.4.** Niech dana jest reguła decyzyjna  $\varphi \rightarrow \psi$ , w której  $\varphi \equiv w_1 \wedge w_2 \wedge \dots \wedge w_j$ , oraz tablica decyzyjna  $\mathbf{DT} = (U, A \cup \{d\})$ . Reguła  $\varphi \rightarrow \psi$  pokrywa przykład  $x \in U$  wtedy i tylko wtedy, gdy  $x \in [\varphi] = \bigcap_{i \in \{1, 2, \dots, j\}} [w_i]$ . Przez  $[w_i]$  oznaczamy zbiór wszystkich przykładów z tablicy  $\mathbf{DT}$ , które spełniają warunek  $w_i$ .

Odwołując się do logiki matematycznej, warunek  $w$  można traktować jako funkcję zdaniową zmiennej  $x \in U$ , wtedy  $[w_i] = \{x \in U : w_i(x) \text{ jest zdaniem prawdziwym}\}$ . Przyjmujemy, że do prawidłowej interpretacji formuły  $w_i(x)$  wszystkie symbole funkcyjne i relacyjne występujące w  $w_i(x)$  interpretowane są zgodnie z ich znaczeniem potocznym i intuicją matematyczną oraz  $\forall a \in A \ a(x) \in Va$ .

Ponieważ  $\psi \equiv d = v$ , do zbioru  $[\psi]$  należą wszystkie przykłady, dla których wartość atrybutu decyzyjnego jest równa  $v$ . Jeśli reguła  $r \equiv \varphi \rightarrow \psi$ , to przez  $[r] = [\varphi] \cap [\psi] = [\varphi \wedge \psi]$  oznaczać będziemy zbiór wszystkich przykładów pozytywnych pokrywanych przez regułę  $r$ .

Określenie zbioru przykładów pozytywnych i negatywnych pokrywanych przez regułę regresyjną wymaga przyjęcia dodatkowych założeń. Ograniczmy się do reguł regresyjnych o postaci (2.4). Założymy, że dane są pewne ustalone wartości  $\varepsilon_1$  i  $\varepsilon_2$ . Zbiór przykładów pozytywnych dla reguły  $\varphi \rightarrow d = v$  stanowią wszystkie przykłady ze zbioru treningowego o wartościach atrybutu decyzyjnego należących do przedziału  $[v - \varepsilon_1, v + \varepsilon_2]$ . Zbiór przykładów negatywnych stanowią wszystkie pozostałe przykłady. Wobec przyjętych założeń określenie zbioru przykładów pozytywnych i negatywnych, pokrywanych przez  $\varphi \rightarrow d = v$ , jest oczywiste.

W uczeniu się przez indukcję spotykane są także reguły pierwszego rzędu (ang. *first-order rules*), w których warunki elementarne oraz konkluzje są wyrażeniami relacyjnymi, określającymi związki pomiędzy przykładami [183, 239, 293]. Indukcją reguł pierwszego rzędu zajmuje się dziedzina maszynowego uczenia się, zwana indukcyjnym programowaniem logicznym [204].

Pojedyncza reguła nie jest traktowana jako kompletna hipoteza, ale jako pewna lokalna zależność opisująca pokrywane przez nią przykłady. Za kompletną hipotezę uznaje się zbiór reguł, pokrywający wszystkie przykłady znajdujące się w treningowej tablicy decyzyjnej. Zazwyczaj wyróżnia się także hipotezy złożone z reguł wskazujących na konkretną klasę

decyzyjną. Hipotezę częściową, opisującą wszystkie przykłady z ustalonej klasy decyzyjnej, nazywa się regułowym opisem klasy decyzyjnej lub po prostu opisem klasy decyzyjnej.

**Definicja 2.5.** Założmy, że dana jest tablica decyzyjna  $\mathbf{DT} = (U, A \cup \{d\})$ , złożona z przykładów należących do klas decyzyjnych  $X_{v1}, X_{v2}, \dots, X_{vk}$ . Zbiór reguł  $RUL_{X_v} = \{\varphi \rightarrow \psi : \psi \equiv (d = v)\}$  nazywa się regułowym opisem klasy decyzyjnej  $X_v$  wtedy i tylko wtedy, gdy  $\forall_{x \in U : d(x)=v} \exists_{r \in RUL_{X_v}} x \in [r]$ .

W literaturze można spotkać się z różnego rodzaju opisami klas decyzyjnych. W szczególności są to opisy minimalne, w których usunięcie jakiegokolwiek reguły powoduje, że warunek z definicji 2.5 przestaje być spełniony. Spotykamy także opisy satysfakcjonujące, częściowe i pełne [191, 284]. Każdy z nich może być opisem niesprzecznym (dokładnym) lub aproksymacyjnym. Opis niesprzeczny składa się jedynie z reguł niesprzecznych (dokładnych), a więc takich, dla których  $[\varphi] = [\varphi \wedge \psi]$ . Wystarczy, że jedna reguła w opisie nie jest dokładna, a opis nazywany jest opisem aproksymacyjnym. W teorii zbiorów przybliżonych rozważa się również opisy niesprzeczne, złożone z minimalnych reguł decyzyjnych. Minimalność reguły rozumiana jest w taki sposób, że usunięcie z niej jakiegokolwiek warunku elementarnego powoduje, że staje się ona regułą aproksymacyjną.

### 2.3. Pokryciowy algorytm indukcji

Dla ustalonej konkluzji wyznaczenie reguły decyzyjnej polega na zdefiniowaniu warunków elementarnych, znajdujących się w jej przesłance. Zdecydowana większość algorytmów tworzy reguły decyzyjne, złożone z prostych warunków elementarnych. Aby zdefiniować warunek elementarny, należy określić: który z atrybutów warunkowych będzie podstawą do utworzenia warunku, jakiego rodzaju operator relacyjny zostanie użyty oraz jaki będzie zakres warunku.

Podczas tworzenia reguł dla celów klasyfikacyjnych zazwyczaj dąży się do indukcji jak najmniejszej liczby reguł. Problem wyznaczenia minimalnego zbioru reguł, złożonego z dokładnych opisów wszystkich klas decyzyjnych, jest problemem NP-zupełnym [154, 270]. W realizacji praktycznych zadań najczęściej stosowane są heurystyczne, pokryciowe algorytmy indukcji, których zasadę działania ilustruje Pokryciowy algorytm indukcji reguł. Indukcja następuje kolejno dla każdej klasy decyzyjnej. Podczas indukcji reguł dla danej klasy wszystkie reprezentujące ją przykłady traktowane są jako przykłady pozytywne, a reszta przykładów – jako negatywne.

```

Pokryciowy algorytm indukcji reguł
Wejście: DT=( $U, A \cup \{d\}$ ) – treningowa tablica decyzyjna,  $X_i$  – klasy decyzyjne
Wyjście: RUL – zbiór wyznaczonych reguł
1 Begin
2   RUL:= $\emptyset$ ;
3   Foreach  $X_i$  do // dla każdej klasy decyzyjnej
4     G:= $X_i$ ;
5     While  $X_i \neq \emptyset$  do
6       r:= Wyznacz-regułę( $X_i, U, A$ );
7        $X_i := X_i - [r]$ ;
8       If Czy-dobra(r) then RUL:= RUL $\cup\{r\}$ 
9         else break;
10      end while
11       $X_i := G$ ;
12    end foreach
13 end.

```

W przedstawionym algorytmie proces indukcji realizowany jest za pomocą procedury **Wyznacz-reguły**. Generuje ona regułę, w której warunki elementarne zbudowane są na podstawie atrybutów należących do zbioru  $A$ . Przykłady pozytywne przechowywane są w sukcesywnie malejącym zbiorze  $X_i$ . Zbiór przykładów negatywnych nie zmienia się i zawiera przykłady ze wszystkich pozostałych klas decyzyjnych. Proces indukcji reguł dla danej klasy kończy się z chwilą pokrycia przez wyznaczone reguły wszystkich przykładów pozytywnych lub kiedy najnowsza z wyznaczonych reguł nie jest wystarczająco dobra (funkcja **Czy-dobra**). W zależności od postaci funkcji **Czy-dobra** i procedury **Wyznacz-reguły** można podać różne konkretyzacje algorytmu. Zadaniem funkcji **Czy-dobra** jest sprawdzenie, czy najnowsza z wyznaczonych reguł spełnia zadane przez użytkownika kryteria minimalnej jakości. Zazwyczaj są to wymagania dotyczące liczby pokrywanych przez nią przykładów pozytywnych lub proporcji pomiędzy liczbą pokrywanych przez nią przykładów pozytywnych i negatywnych [57]. Zauważmy, że w przypadku sprzecznej tablicy decyzyjnej nie da się pokryć całej klasy decyzyjnej jedynie dokładnymi regułami. Do funkcji **Czy-dobra** przekazywany może być również zbiór wszystkich dotychczas wyznaczonych reguł. Umożliwia to sprawdzenie, czy dodanie do niego najnowszej z wyznaczonych reguł spowoduje poprawę jego jakości. Na etapie indukcji jakość zbioru reguł jest najczęściej weryfikowana przez badanie jego zdolności klasyfikacyjnych na wydzielonym zbiorze przykładów lub za pomocą zasady minimalnej długości kodu (MDL – *minimal description length principle*) [52, 76, 238, 222, 297, 298], o której wspomnimy w następnym rozdziale podczas omawiania wybranych algorytmów indukcji reguł.

Procedura **Wyznacz-reguły** jest zasadniczą częścią algorytmu, związaną z procesem indukcji. Główne różnice pomiędzy pokryciowymi algorytmami indukcji reguł polegają na różnych realizacjach tej procedury. W swoich pracach [87, 89] Fürnkranz przeprowadził analizę ponad 40 programów indukcji reguł. Zdecydowana większość z nich dokonuje

indukcji reguł zawierających proste warunki elementarne. Dla atrybutów symbolicznych w warunku elementarnym zazwyczaj stosowany jest operator relacyjny  $=$ , rzadziej  $\in$ . Dla atrybutów porządkowych i ciągłych stosowane są operatory relacyjne  $\{\leq, \geq, >, <\}$ , a jeśli podczas indukcji zakres warunku elementarnego ograniczony jest z dwóch stron, to stosowany jest operator  $\in$ .

W większości algorytmów początkowo reguła pokrywa wszystkie przykłady, znajdujące się w tablicy treningowej. W uproszczeniu można powiedzieć, że przesłanka takiej reguły zawiera uniwersalne warunki elementarne, zbudowane na podstawie wszystkich atrybutów warunkowych. Uniwersalnym warunkiem elementarnym będziemy nazywali warunek, którego zakres obejmuje cały znany zbiór wartości atrybutu. Przesłankę takiej reguły spełnia każdy przykład. W rzeczywistości uniwersalnych warunków nie specyfikuje się w przesłance, gdyż nie nakładają one żadnych ograniczeń na przykłady pokrywane przez regułę. W kolejnych krokach algorytmu uniwersalne warunki elementarne specjalizowane są tak długo, dopóki reguła nie spełni wymagań minimalnej jakości lub dalsza specjalizacja nie spowoduje już poprawy jakości. Wymagania dotyczące jakości mogą być wyrażane na wiele sposobów. Najczęściej regułę specjalizuje się w taki sposób, aby pokrywała jak najwięcej przykładów pozytywnych i jak najmniej negatywnych. Część algorytmów specjalizuje reguły tak długo, aż są one dokładne. Dla nietrywialnych zbiorów danych indukcja reguł dokładnych najczęściej prowadzi do uzyskania opisów klas decyzyjnych, zbyt dopasowanych do danych treningowych. Nadmierne dopasowanie ma negatywny wpływ na jakość klasyfikatora zbudowanego na podstawie wyznaczonych reguł. Z tego też powodu po fazie specjalizacji realizowana jest faza przycinania, której wynikiem mogą być reguły niedokładne. Jakość reguł niedokładnych oceniana jest za pomocą tzw. miar jakości (ang. *rule quality measures*) [8, 40, 90, 176, 251, 264, 340]. Najczęściej miara jakości jest kryterium pozwalającym ocenić jednocześnie ogólność i dokładność reguły. Algorytm specjalizuje i przycina reguły w taki sposób, aby charakteryzowały się one jak najwyższą jakością.

Specjalizacja utożsamiana jest z fazą wzrostu reguły (ang. *growing phase*), gdyż powoduje ona, że w przesłance reguły pojawiają się warunki elementarne. Po fazie wzrostu realizowana jest faza przycinania, polegająca na usunięciu tych warunków elementarnych, które okazały się zbędne (warunki elementarne zamieniane są znowu na uniwersalne warunki elementarne). Usunięcie pewnych warunków elementarnych, utworzonych w fazie wzrostu, może spowodować, że jakość przyciętej reguły wzrośnie lub pozostanie niezmieniona. Podczas indukcji reguł dla celów klasyfikacyjnych dąży się do tworzenia reguł złożonych z jak najmniejszej liczby warunków elementarnych, dlatego warunki nieprzyczyniające się do wzrostu jakości reguły są usuwane. Faza przycinania ma sens jedynie wtedy, gdy faza wzrostu nie jest realizowana przez procedurę przeszukiwania wyczerpującego.

Faza wzrostu powoduje *specjalizację* reguły, a faza przycinania – jej *uogólnianie*. Zależność pomiędzy dokładnością a ogólnością reguły jest taka, że wzrost dokładności powoduje spadek ogólności. Zachodzi również sytuacja odwrotna. Miary jakości, stosowane w fazach wzrostu i przycinania, odzwierciedlają pewien kompromis pomiędzy oceną dokładności i ogólności reguły, przy czym, w zależności od miary, akcenty mogą być różnie rozłożone. Pewna grupa miar większą wagę przywiązuje do dokładności reguł, inna grupa dokładność i ogólność taktuje równorzędnie. W definicji miary jakości można zawrzeć również inne kryteria, odzwierciedlające na przykład preferencje użytkownika, dotyczące postaci wyznaczanych reguł [257, 345].

W przeglądzie [89] przedstawione powyżej podejście do indukcji reguły Fürnkranz nazywa strategią *od ogółu do szczegółu* (ang. *top-down strategy*).

Specjalizacja reguł realizowana jest za pomocą różnych algorytmów przeszukiwania. Dla większych zbiorów danych strategia wyczerpująca nie jest praktycznie wykorzystywana ze względu na wysoką złożoność obliczeniową. W praktyce najbardziej popularne są: strategia wspinaczki (ang. *hill climbing*) oraz przeszukiwanie wiązką (ang. *beam search*). Przy czym wspinaczka stosowana jest prawie trzykrotnie częściej [89].

W każdym kroku fazy wzrostu, realizowanej za pomocą strategii wspinaczki, sprawdzane są wszystkie sensowne specjalizacje aktualnie rozważanej reguły, a spośród nich wybierana jest ta, która uzyskuje najwyższą ocenę. Uzyskana w ten sposób reguła jest podstawą do dalszej specjalizacji. Za sensowną uznaje się jedynie taką specjalizację, która powoduje ograniczenie liczby przykładów negatywnych, pokrywanych przez regułę.

Algorytm realizujący przeszukiwanie wiązką [332] wymaga podania wartości parametru *wb*, określającego tzw. szerokość wiązki. W każdym kroku wzrostu pamiętanych jest *wb* najlepszych specjalizacji reguły. Każdy kolejny krok algorytmu polega na wygenerowaniu dla wszystkich reguł, należących do wiązki, ich sensownych specjalizacji i wyborze spośród nich kolejnej wiązki, składającej się z *wb* najlepszych reguł. Faza przycinania realizowana jest na podobnej zasadzie. Po zakończeniu indukcji reguł tworzących wiązkę jako wynikową pamięta się regułę o najlepszej jakości. Algorytm przeszukiwania wiązką jest algorymem dokładniejszym od wspinaczki, ale wymagającym większych nakładów obliczeniowych.

Po zakończeniu indukcji zbiór reguł poddawany jest dodatkowej obróbce, mającej na celu usunięcie reguł zbędnych [5, 244, 254, 264] i/lub modyfikację warunków elementarnych [52, 53, 57]. W literaturze przedmiotu etap ten nazywany jest postprocessingiem lub postpruningiem [52, 86, 87]. W niniejszej monografii etap ten nazywany będzie etapem przycinania zbioru reguł. W fazie indukcji procedura *Wyznacz-reguły* koncentruje się na optymalizacji jakości aktualnie tworzonej reguły. W etapie przycinania optymalizacja może również dotyczyć całego zbioru reguł, zasadniczym kryterium jakości będą wtedy zdolności klasyfikacyjne wyznaczonych reguł, a także ich liczba.

Indukcja reguł regresyjnych za pomocą algorytmów pokryciowych przebiega w sposób bardzo podobny do indukcji reguł decyzyjnych [144, 330]. W fazie wzrostu najczęściej stosowane są dwa podejścia do oceny warunków elementarnych. Warunki tworzone są w taki sposób, aby: minimalizować wariancję wartości zmiennej decyzyjnej przykładów pokrywanych przez regułę lub maksymalizować wartość miary oceniającej ogólność i dokładność reguły. Konkluzja budowanej reguły zmienia się dynamicznie, w najprostszym przypadku jest to średnia lub mediana wartości zmiennej decyzyjnej przykładów pokrywanych przez regułę. W przypadku bardziej zaawansowanym w konkluzji znajduje się model regresji, który wyznaczany jest po każdej konkretyzacji warunku elementarnego lub po zakończeniu fazy wzrostu.

Algorytmy indukcji reguł decyzyjnych, definiowanych dla celów opisowych, stosują idee zawarte w algorytmach indukcji reguł asocjacyjnych [148, 157, 169, 235, 284, 285] lub metody przeszukiwania wyczerpującego [234]. W regule decyzyjnej konkluzja jest ustalona (atrybut decyzyjny jest ustalony), co przyspiesza proces analizy danych. Algorytmy indukcji reguł decyzyjnych, definiowanych dla celów opisowych, zazwyczaj generują zbiór wszystkich reguł, spełniających zadane przez użytkownika wymogi minimalnej jakości. Najczęściej wymogami tymi są minimalna ogólność i dokładność. Przed indukcją atrybuty typu ciągłego poddawane są dyskretyzacji [52, 209]. Idea indukcji bazuje na spostrzeżeniu, że wymóg minimalnej ogólności może spełniać jedynie taka koniunkcja warunków elementarnych, której warunki stanowią nadzbiór zbioru warunków również spełniających ten wymóg. Spostrzeżenie to, podobnie jak w algorytmach indukcji reguł asocjacyjnych, pozwala ukierunkować proces indukcji. Reguły spełniające wymagania minimalnej jakości dodawane są do wynikowego zbioru reguł.

Zauważmy, że przesłanka reguły spełniającej wymogi minimalnej jakości może być w dalszym ciągu rozszerzana o kolejne warunki elementarne. W niektórych algorytmach [257] proces dalszego wzrostu reguł spełniających warunki minimalnej jakości jest kontynuowany. Jeśli rozszerzona reguła w dalszym ciągu spełnia wymóg minimalnej ogólności (bo wymóg minimalnej dokładności spełnia na pewno), to także ona oraz wszystkie jej dalsze specjalizacje, spełniające wymagania jakości, dodawane są do wynikowego zbioru reguł.

Obecnie krótko omówionych zostanie 6 algorytmów, umożliwiających indukcję reguł decyzyjnych. Algorytmy te stanowią kanon indukcji reguł i można śmiało powiedzieć, że na ich podstawie opracowano dziesiątki innych, często ukierunkowanych na specyficzne zastosowania, metod indukcji reguł. W opisie starano się akcentować rolę, jaką w procesie indukcji pełnią miary oceniające jakość reguł.

Pierwszym pokryciowym algorymem indukcji reguł był algorytm **AQ**, zaproponowany przez Michalskiego [189]. Algorytm ten występuje w kilku różnych postaciach, które stały

się podstawą do opracowania całej rodziny programów komputerowych (AQ15, AQ17, AQ18, AQ21). Sposób generowania reguły jest we wszystkich tych programach podobny. W pierwotnej i najbardziej popularnej postaci AQ operuje jedynie na wektorze atrybutów symbolicznych, a analiza zbioru przykładów opisanych przez atrybuty ciągłe wymaga ich uprzedniej dyskretyzacji [159]. Najnowsze programy, dla których podstawą jest algorytm AQ, przetwarzają również przykłady opisane przez atrybuty ciągłe [336].

Algorytm AQ dopuszcza, aby w warunkach elementarnych występowała tzw. wewnętrzna alternatywa. Oznacza to, że warunek elementarny, nazywany tutaj selektorem, może przyjąć postać:  $[atrbut=wartość\_1 \vee wartość\_2 \vee \dots \vee wartość\_n]$ . Koniunkcja selektorów nazywana jest kompleksem. Działanie algorytmu AQ w podstawowej formie polega na budowaniu coraz większego pokrycia tablicy treningowej. Pokrycie to składa się z reguł dokładnych. Algorytm wykorzystuje strategię przeszukiwania wiązką, szerokość wiązki  $wb$ , ma wpływ na liczbę kompleksów będących efektem budowy tzw. gwiazdy. Algorytm można skrótnie zapisać w następujący sposób:

Szablon algorytmu AQ

```
While Częściowe pokrycie nie pokrywa wszystkich przykładów pozytywnych do
    Wybierz niepokryty pozytywny przykład (tzw. ziarno);
    Określ maksymalnie ogólne kompleksy pokrywające ziarno i żadnego
    negatywnego przykładu (stwórz tzw. gwiazdę);
    Wybierz najlepszy, wg ustalonego przez użytkownika kryterium jakości,
    kompleks należący do utworzonej gwiazdy;
    Stwórz nowe, częściowe pokrycie, dodając ten najlepszy kompleks do
    aktualnego częściowego pokrycia
end while
Ostatnie częściowe pokrycie jest pokryciem uczonej klasy decyzyjnej.
```

Procedura tworzenia gwiazdy odpowiada za konstrukcję reguł.

Szablon algorytmu wyznaczania gwiazdy

```
Częściowa gwiazda pokrywa wszystkie przykłady
While Częściowa gwiazda pokrywa jakiś przykład negatywny do
    Wybierz pokryty przykład negatywny;
    Stwórz wszystkie możliwe i maksymalnie ogólne kompleksy, z których każdy
    pokrywa ziarno i nie pokrywa wybranego przykładu negatywnego;
    Wynik zapisz jako częściową gwiazdę dla ziarna;
    Wyznacz przecięcie wszystkich par kompleksów, należących do częściowej
    gwiazdy i częściowej gwiazdy dla ziarna;
    Wynik zapisz do częściowej gwiazdy;
    Wybierz  $wb$  najlepszych rozłącznych kompleksów, wchodzących w skład
    częściowej gwiazdy
end while
Ostatnia częściowa gwiazda staje się gwiazdą dla ziarna.
```

Podczas budowy gwiazdy tworzonych jest tyle kompleksów, ile jest atrybutów warunkowych, gdyż ograniczenie zakresu dowolnego selektora wystarczy, aby kompleks nie pokrywał aktualnie rozważanego przykładu negatywnego. Po przecięciu częściowej gwiazdy dla ziarna z częściową gwiazdą algorytm AQ musi dokonać wyboru  $wb$  spośród wyznaczonych kompleksów. Następnie spośród tych  $wb$  kompleksów wybierany jest najlepszy, który dodawany jest do zbioru reguł. W algorytmie AQ wykorzystuje się

jednocześnie kilka kryteriów oceny kompleksu; kryteria tebrane są pod uwagę zgodnie z porządkiem leksykograficznym. Standardowo lepszy jest ten kompleks, który pokrywa więcej przykładów pozytywnych. Jeśli jest więcej kompleksów pokrywających tę samą liczbę przykładów pozytywnych, to wybierany jest ten, który zawiera mniej selektorów. W modyfikacjach standardowej wersji algorytmu AQ dopuszcza się tworzenie reguł niedokładnych, wtedy do oceny jakości kompleksu stosowanych jest kilka miar jakości [156]. Algorytm AQ opisany jest w wielu pracach przeglądowych, dotyczących indukcji reguł, szczególnie godna polecenia jest praca Cichosza [52].

Kolejnym popularnym algorytmem jest **CN2** [53], który jest modyfikacją AQ, dodatkowo wykorzystującą pewne metody, stosowane w indukcji drzew decyzyjnych [238]. Główne różnice pomiędzy podstawową wersją AQ a CN2 to dopuszczanie przez CN2 do indukcji reguł niedokładnych oraz tworzenie listy reguł.

W CN2 podczas tworzenia reguły również konstruowana jest gwiazda; różnica w stosunku do AQ jest taka, że kompleksy tworzące gwiazdę specjalizowane są tak długo, aż gwiazda staje się zbiorem pustym. Po każdym etapie specjalizacji kompleksy są oceniane za pomocą pewnej miary jakości. Efektem działania procedury konstrukcji gwiazdy jest kompleks, który spośród wszystkich rozważanych specjalizacji uzyskał najwyższą ocenę. Podczas oceny kompleksów wykorzystuje się jedynie przykłady niepokryte przez dotychczas wyznaczone reguły. W pierwszej wersji algorytmu CN2 do oceny kompleksu stosowano negację entropii rozkładu pokrywanych przez niego przykładów. Im wyraźniej wśród przykładów pokrywanych przez kompleks dominowała jedna z klas decyzyjnych, tym wartość entropii bardziej zbliżała się do zera. W późniejszych wersjach algorytmu albo modyfikowano sposób obliczania entropii, albo do oceny jakości stosowano tzw. *m*-oszacowanie, które jest uogólnieniem estymaty Laplace'a [54, 90]. Wartość *m*-oszacowania obliczana jest zgodnie ze wzorem  $(p + Bm)/(p + n + m)$ . We wzorze tym *p* i *n* to odpowiednio liczba przykładów pozytywnych i negatywnych, pokrywanych przez regułę, *m* jest parametrem, którego wartość ustalana jest przez użytkownika, a  $B = P/(P + N)$  jest tzw. dokładnością podstawową (dokładnością bazową), obliczaną dla zadanej tablicy treningowej na podstawie liczby wszystkich przykładów pozytywnych *P* i negatywnych *N*.

Stosowanie entropii do oceny kompleksów powodowało preferowanie przez CN2 kompleksów dokładnych, dlatego do algorytmu wprowadzono drugie kryterium, zabezpieczające kompleksy przed nadmiernym dopasowaniem do danych treningowych. Kryterium tym jest wymóg statystycznej istotności kompleksów. Do oceny statystycznej istotności wykorzystano statystykę  $G^2$ , stosowaną również podczas przycinania drzew decyzyjnych. Użycie obu kryteriów entropii (lub *m*-oszacowania) i statystyki  $G^2$  powoduje, że kompleks niedokładny może okazać się lepszy od dokładnego. Jeśli wynikiem

wyznaczenia gwiazdy jest kompleks niedokładny, to przyjmuje się, że opisuje on tę klasę decyzyjną, która jest dominująca w zbiorze pokrywanych przez niego przykładów.

Opisany sposób postępowania powoduje, że w CN2 kolejno generowane reguły mogą wskazywać na różne klasy decyzyjne. Ponieważ kolejne reguły wyznaczane są na podstawie dotychczas niepokrytych przykładów, kolejność wyznaczania reguł ma duże znaczenie dla realizowanego za ich pomocą algorytmu klasyfikacji. W algorytmie CN2 klasyfikacja odbywa się z wykorzystaniem listy reguł, a kolejność reguł na liście jest zgodna z kolejnością ich indukcji. Na końcu listy znajduje się tzw. reguła domyślna (ang. *default rule*), pokrywająca każdy przykład i wskazująca na najliczniejszą klasę decyzyjną. Reguła domyślna gwarantuje, że hipoteza uzyskana przez algorytm CN2 jest pełna i opisuje cały zbiór treningowy. Reguła ta jest potrzebna, gdyż wymóg statystycznej istotności kompleksów powoduje, że dla pewnej liczby przykładów próba wyznaczenia pokrywających je reguł może zakończyć się niepowodzeniem. Szczegółowo mechanizm klasyfikacji za pomocą uporządkowanej listy reguł omówiony zostanie w dalszej części rozdziału.

Trzeci z omawianych algorytmów to niezwykle popularny i efektywny algorytm **RIPPER** [57]. Tworzy on listę reguł decyzyjnych. Opisy klas decyzyjnych budowane są, począwszy od najmniej licznej klasy. Indukcja pojedynczej reguły odbywa się przy wykorzystaniu strategii wspinaczki. Początkowo przesłanka reguły złożona jest ze wszystkich uniwersalnych warunków elementarnych, w fazie specjalizacji warunki te są kolejno ograniczane. Najlepsza jest taka specjalizacja, której wynikiem jest warunek elementarny, maksymalizujący wartość kryterium korzyści informacji (ang. *information gain*). Kryterium to porównuje entropię rozkładów przykładów pokrywanych przez regułę niezawierającą ocenianego warunku elementarnego i regułę zawierającą ten warunek. Pierwotnie kryterium korzyści informacji stosowane było w algorytmie indukcji drzew decyzyjnych C4.5 do oceny podziału węzła [238] oraz w algorytmie indukcyjnego programowania logicznego FOIL [239].

Przed indukcją reguł każda klasa decyzyjna dzielona jest na dwa rozłączne zbiory: zbiór specjalizacji (oryginalnie nazywany *growing set*), stanowiący 2/3 wejściowego zbioru przykładów, oraz zbiór przycinania (*pruning set*), zawierający pozostałe przykłady treningowe. Specjalizacja każdej reguły odbywa się na zbiorze specjalizacji i trwa tak długo, dopóki powoduje udokładnianie reguły. Po fazie specjalizacji uruchamiana jest faza przycinania. W fazie tej stosowane jest inne kryterium jakości niż podczas specjalizacji. Kryterium to wyraża się wzorem  $(p-n)/(p+n)$ , gdzie  $p$  i  $n$  to odpowiednio liczba przykładów pozytywnych i negatywnych pokrywanych przez przyciętą regułę. Po fazie przycinania sprawdzane jest, czy wśród przykładów pokrywających regułę ponad 50% stanowią przykłady pozytywne. Jeśli tak jest, algorytm sprawdza, jaka jest minimalna długość kodu, potrzebna do zakodowania dotychczas wyznaczonego zbioru reguł (łącznie

z wyznaczoną właśnie regułą), oraz informacji potrzebnej do przyporządkowania każdemu przykładowi, należącemu do zbioru specjalizacji, poprawnej klasy decyzyjnej. Przy czym zakłada się, że kodowane są jedynie przykłady niepokryte przez żadną z dotychczas wyznaczonych reguł oraz te, które przez wyznaczone reguły przyporządkowane są do błędnych klas decyzyjnych. Uzyskana w ten sposób długość kodu nie powinna być większa niż  $d$  bitów od wyznaczonej w podobny sposób długości kodu zbioru reguł, który nie zawiera właśnie wyznaczanej reguły. Liczba  $d$  jest parametrem programu i standardowo wynosi 64. Jeśli reguła spełnia ograniczenia minimalnej dokładności i długości kodu, to dodawana jest na koniec listy reguł, a ze zbioru specjalizacji usuwa się wszystkie przykłady, pokrywające wyznaczoną regułę. Proces ten powtarza się tak długo, dopóki nie zostaną pokryte wszystkie przykłady ze zbioru specjalizacji lub nie będzie spełnione kryterium dokładności reguły lub długości kodu. W standardowej wersji algorytm RIPPER nie dokonuje indukcji reguł dla najliczniejszej klasy decyzyjnej; dla klasy tej tworzona jest reguła domyślna, która umieszczana jest na końcu listy.

Nierozerwalną częścią algorytmu RIPPER jest optymalizacja wyznaczonych reguł. Wynikiem działania algorytmu jest lista reguł  $\langle r_1, r_2, \dots, r_n \rangle$ . Są one kolejno optymalizowane, począwszy od  $r_{n-1}$ , aż do  $r_1$ . Każda z reguł  $r_i$  zastępowana jest przez dwie reguły kandydatki  $r_i^{rp}$  i  $r_i^{rv}$ , w ten sposób otrzymywane są dwie nowe listy reguł. Reguła  $r_i^{rv}$  to reguła  $r_i$ , którą algorytm redefiniuje (specjalizuje i przycina) w standardowy dla RIPPERA sposób. Podczas tej redefinicji pod uwagę brane są tylko przykłady pokrywane unikalnie przez  $r_i$ . Indukcja reguły  $r_i^{rp}$  przebiega w taki sposób, aby na zbiorze przycinania maksymalizować zdolności klasyfikacyjne całej listy. W efekcie optymalizacji otrzymujemy trzy listy reguł: listę wejściową, listę wejściową z  $r_i^{rp}$  zamiast  $r_i$  oraz listę wejściową z  $r_i^{rv}$  zamiast  $r_i$ . Spośród tych list wybierana jest ta, która charakteryzuje się najkrótszą długością kodu.

Najbardziej rozpowszechnioną implementacją algorytmu RIPPER jest ta zawarta w systemie Weka [333]. Ciekawą implementację, dokonującą indukcji reguł także dla najliczniejszej klasy decyzyjnej, zawarto w systemie Rapid-Miner.

Kolejny interesujący algorytm indukcji reguł to **ITRULE** [99]. Wykorzystuje on jedynie proste warunki elementarne *atrybut=wartość*. Przed rozpoczęciem indukcji atrybuty ciągłe muszą być poddane dyskretyzacji [52, 159, 209]. Rezultatem działania algorytmu są tzw. probabilistyczne reguły decyzyjne, które poza standardowym wyrażeniem *jeżeli φ, to ψ* zawierają na końcu frazę *z prawdopodobieństwem....* Wyznaczając reguły, algorytm ITRULE stosuje strategię przeszukiwania w głąb oraz kryterium jakości, znane jako miara  $J$  [279]. Genezą miary  $J$  jest teoria informacji. Aby zminimalizować liczbę reguł

kandydujących, które w procesie poszukiwania należałyby przetestować, algorytm wykorzystuje pewne teoretyczne własności miary  $J$ . Własności te pozwalają stwierdzić, czy dalsza specjalizacja reguły może prowadzić do wzrostu wartości miary  $J$ . Jeśli osiągnięcie wyższej wartości miary jest niemożliwe, proces indukcji jest przerwany. Do przypisania wartości prawdopodobieństwa wyznaczonym regułom ITRULE stosuje  $m$ -oszacowanie.

Indukcję reguł decyzyjnych można również przeprowadzić, stosując metody oferowane przez **teorię zbiorów przybliżonych**. Teoria ta została zaproponowana przez Zdzisława Pawlaka na początku lat 80. XX wieku [220]. Teoria zbiorów przybliżonych proponuje nowe spojrzenie na wiedzę i jej reprezentację. U jego podstaw leży przekonanie, że wiedza to zdolność do klasyfikacji. Przez klasyfikację rozumie się umiejętność odróżniania obiektów-elementów otaczającej rzeczywistości. Klasyfikacji dokonujemy na podstawie znajomości cech, jakie ma dany obiekt lub grupa obiektów. Zatem do zdefiniowania wiedzy musimy mieć pewien zbiór obiektów, które chcemy klasyfikować, oraz zbiór cech, które te obiekty opisują. Takie założenia prowadzą wprost do zdefiniowanego już pojęcia tablicy decyzyjnej  $DT = (U, A \cup \{d\})$  i klas decyzyjnych. W zbiorze przykładów  $U$  definiowana jest relacja nieroróżnialności przykładów ze względu na zbiór atrybutów  $B \subseteq A$ :

$$IND(B) = \{<x, y> \in U \times U : \forall a \in B \quad a(x) = a(y)\}. \quad (2.5)$$

Nazwa relacji pochodzi od angielskiego słowa *indiscernibility*, oznaczającego nieroróżnialność. Relacja  $IND$  jest relacją równoważności. Klasy abstrakcji relacji  $IND(B)$  nazywane są zbiorami B-elementarnymi. Dla dowolnego podzbioru zbioru przykładów  $X$ , w szczególności dla klas decyzyjnych, definiuje się dolne i górne przybliżenia wyznaczane przez zbiór atrybutów  $B$ :

$$\underline{B}X = \{x \in U : [x]_{IND(B)} \subseteq X\}, \quad \overline{B}X = \{x \in U : [x]_{IND(B)} \cap X \neq \emptyset\}. \quad (2.6)$$

Jeżeli zbiór przykładów tworzących klasę decyzyjną można wyrazić jako sumę pewnych zbiorów B-elementarnych, to znając wartości atrybutów należących do zbioru  $B$ , można w jednoznaczny sposób podać również wartość atrybutu decyzyjnego.

Zależność pomiędzy wartościami atrybutów przykładów należących do zbioru B-elementarnego a przyporządkowaniem tych przykładów do odpowiedniej klasy decyzyjnej można przedstawić w postaci reguły decyzyjnej. Założymy, że dany jest zbiór B-elementarny  $[x]_{IND(B)} \subseteq X_v$ ,  $B = \{a_1, a_2, \dots, a_k\}$  oraz  $X_v$  jest klasą decyzyjną utożsamianą z wartością  $v$  atrybutu decyzyjnego; wówczas reguła:

$$\text{jeżeli } a_1 = a_1(x) \wedge a_2 = a_2(x) \wedge \dots \wedge a_k = a_k(x), \text{ to } d = v \quad (2.7)$$

wyraża zależność pomiędzy wartościami atrybutów opisujących przykład  $x$  a przynależnością tego przykładu do klasy decyzyjnej  $X_v$ . Przykład  $x$  można nazwać przykładem-generatorem reguły. Ponieważ relacja nieroróżnialności jest relacją

równoważności, wybór przykładu-generatora jest nieistotny. Wystarczy, że należy on do danego zbioru B-elementarnego, który jest klasą abstrakcji relacji  $IND(B)$ . Zauważmy, że reguła (2.7) jest dokładna, gdyż wszystkie przykłady należące do klasy abstrakcji należą również do klasy decyzyjnej. W zależności od tego, ile zbiorów B-elementarnych jest potrzebnych do pokrycia klasy decyzyjnej, tyle reguł będzie opisywało klasę decyzyjną.

W sytuacji gdy klasy decyzyjnej nie można przedstawić w postaci sumy zbiorów B-elementarnych, można podać jej przybliżony opis za pomocą przybliżeń B-dolnego i B-górnego. Ponieważ B-dolne przybliżenie zawiera się w przybliżonym zbiorze, reguły utworzone na podstawie dolnego przybliżenia będą regułami dokładnymi. Reguły utworzone na podstawie B-górne przybliżenia będą regułami niedokładnymi. W terminologii zbiorów przybliżonych reguły wygenerowane na podstawie dolnego przybliżenia klasy decyzyjnej nazywa się regułami pewnymi (ang. *certain rules*), a reguły wygenerowane na podstawie górnego przybliżenia klasy – regułami możliwymi (ang. *possible rules*). W tablicy decyzyjnej zazwyczaj znajdują się przykłady reprezentujące kilka klas decyzyjnych; w teorii zbiorów przybliżonych wprowadza się definicję B-obszaru pozytywnego tablicy decyzyjnej, oznaczanego jako  $POS_B(DT)$ . Zbiór  $POS_B(DT)$  definiowany jest jako suma B-dolnych przybliżeń wszystkich klas decyzyjnych. Przykłady należące do obszaru pozytywnego są szczególnie interesujące, gdyż reguły wyznaczone na ich podstawie są regułami dokładnymi.

Załóżmy, że dane są niesprzeczna tablica decyzyjna oraz klasa decyzyjna  $X$ . Teoria zbiorów przybliżonych dostarcza narzędzi pozwalających na taki wybór zbioru  $B$ , aby reguła utworzona na podstawie przykładu  $x$ , spełniającego warunek  $[x]_{IND(B)} \subseteq X$ , była tzw. regułą minimalną. Regułę nazywa się minimalną, jeśli jest ona dokładna, a usunięcie z jej przesłanki jakiegokolwiek warunku elementarnego powoduje spadek jej dokładności. Reguły minimalne tworzy się na podstawie tzw. reduktów względem decyzji, nazywanych także reduktami względnymi (ang. *relative reducts*) [220, 270].

Rozróżniane są dwa typy reduktów względnych: redukt względny w tablicy decyzyjnej oraz redukt względny dla przykładu (ang. *object-related relative reduct*). Reduktem względnym w tablicy decyzyjnej  $DT = (U, A \cup \{d\})$  nazywa się zbiór atrybutów  $B \subseteq A$ , spełniających następujące dwa warunki:

- $POS_B(DT) = POS_A(DT)$ ,
- dla każdego zbioru  $C \subset B$  warunek pierwszy nie jest spełniony.

Reduktem względnym dla przykładu  $x \in U$  nazywamy zbiór atrybutów  $B \subseteq A$ , spełniający następujące dwa warunki:

- $\{y \in [x]_{IND(B)} : d(y) \neq d(x)\} = \{y \in [x]_{IND(A)} : d(y) \neq d(x)\}$ ,
- dla każdego zbioru  $C \subset B$  warunek pierwszy nie jest spełniony.

Dla tablicy decyzyjnej lub dla ustalonego przykładu może istnieć kilka reduktów względnych. Zbiór atrybutów stanowiących redukt względny w tablicy decyzyjnej jest minimalnym zbiorem, który z taką samą dokładnością jak cały zbiór atrybutów odróżnia dowolne dwa przykłady, należące do różnych klas decyzyjnych. Redukt względny dla przykładu  $x$  jest minimalnym zbiorem atrybutów pozwalających z taką samą dokładnością jak cały zbiór atrybutów odróżnić ten przykład od przykładów z innych klas decyzyjnych. Spośród wszystkich reduktów szczególnie interesujący jest redukt zawierający najmniej atrybutów, problem poszukiwania takiego reduktu jest NP-zupełny [270]. Dokładne algorytmy wyznaczania reduktów zaprezentowano m.in. w [168, 270, 292]. W praktycznych zastosowaniach zamiast algorytmów dokładnych stosowane są heurystyki pozwalające na wyznaczenie pewnej ograniczonej liczby reduktów, znalezienie reduktu quasi-najkrótszego, a także reduktów aproksymacyjnych i dynamicznych [16, 17, 198, 199, 208, 211, 301, 303, 343].

Redukty stanowią podstawę do wyznaczenia minimalnych reguł decyzyjnych. Zgodnie z wyrażeniem (2.7), warunki elementarne reguł mają postać *atribut=wartość*. Postać warunków elementarnych wymaga, aby atrybuty o wartościach ciągłych zostały poddane dyskretyzacji. Bez dyskretyzacji do zbiorów B-elementarnych, zbudowanych na podstawie atrybutów ciągłych, należałoby niewiele przykładów, a same warunki elementarne byłyby bardzo specyficzne. W tolerancyjnym modelu zbiorów przybliżonych [227, 271], w którym zamiast relacji nieroróżnialności użyto relacji podobieństwa, możliwe jest pominięcie dyskretyzacji oraz stosowanie warunków elementarnych *atribut ∈ V*, gdzie *V* jest pewnym zbiorem wartości atrybutu *atribut*.

Wyznaczanie reguł decyzyjnych może odbywać się na podstawie reduktów względnych w tablicy decyzyjnej lub na podstawie reduktów względnych dla przykładów. W praktyce reguły tworzone są jedynie na podstawie wybranych reduktów. Często stosowanym rozwiązaniami jest wyznaczenie reguł jedynie na podstawie najkrótszego reduktu [208, 246]. Program RSES i biblioteka RSESlip [17] oferują możliwość: wyznaczenia wszystkich minimalnych reguł decyzyjnych, utworzenia reguł na podstawie najkrótszego reduktu lub na podstawie zadanej przez użytkownika liczby reduktów. Reguły tworzone na podstawie reduktów aproksymacyjnych pokrywają więcej przykładów, zawierają także mniej warunków elementarnych.

W obrębie teorii zbiorów przybliżonych opracowano również algorytmy indukcji minimalnych opisów klas decyzyjnych. Najpopularniejszym algorytmem tego typu jest

zaproponowany przez Grzymałę-Busse **LEM2** [119]. Reguły generowane są oddziennie dla każdej klasy decyzyjnej, a zasada tworzenia opisu klasy podobna jest do zasady generowania kolejnych pokryć.

Algorytm LEM2 poszukuje tzw. lokalnego pokrycia klasy decyzyjnej. Lokalne pokrycie składa się z minimalnych kompleksów. Minimalnym kompleksem dla klasy decyzyjnej  $X_v$  nazywany jest najmniejszy, w sensie inkluzji, zbiór  $K$ , składający się z wyrażeń  $w \equiv (atrybut = wartość)$ , o własnościach  $[K] \neq \emptyset$  oraz  $[K] \subseteq X_v$ , przy czym  $[K]$  definiowane jest zgodnie ze wzorem (2.8):

$$[K] = \bigcap_{w \in K} [w]. \quad (2.8)$$

We wzorze (2.8) zapis  $[w] = [atrybut = wartość]$  oznacza zbiór wszystkich przykładów z treningowej tablicy decyzyjnej, dla których wartość atrybutu *atrybut* równa się *wartość*. Dla ustalonej klasy decyzyjnej  $X_v$  każdy minimalny kompleks może być utożsamiony z minimalną regułą decyzyjną o postaci (2.7).

Lokalnym pokryciem klasy decyzyjnej  $X_v$  nazywana jest rodzina kompleksów  $\mathfrak{I}$ , o następujących własnościach:

1.  $\bigvee_{K \in \mathfrak{I}} K$  jest minimalnym kompleksem dla  $X_v$ .
2.  $\bigcup \mathfrak{I} = X_v$ .
3. Dla każdego  $\mathfrak{R} \subset \mathfrak{I}$  warunek 2 nie jest spełniony.

Budowa minimalnego kompleksu podobna jest do budowy reguły w algorytmach pokryciowych. W fazie wzrostu do kompleksu  $K$  dodawane są kolejno te wyrażenia *atrybut = wartość*, które maksymalizują moc zbioru  $[atrybut = wartość] \cap G$ . Początkowo zbiór  $G$  zawiera wszystkie przykłady z klasy decyzyjnej; wraz ze wzrostem kompleksu jest on ograniczany do przykładów niepokrytych jeszcze przez tworzony kompleks. Rozrost kompleksu kończy się z chwilą, gdy spełniony zostanie warunek  $[K] \subseteq X_v$ . Następnie uruchamiana jest faza przycinania, polegająca na usuwaniu z kompleksu wyrażeń *atrybut = wartość* tak długo, jak długo spełniany jest warunek  $[K] \subseteq X_v$ . Po zakończeniu przycinania, na podstawie kompleksu tworzona jest reguła o postaci (2.7), a z klasy decyzyjnej usuwane są wszystkie przykłady pokrywające tę regułę. Następnie rozpoczyna się procedura tworzenia kolejnego kompleksu.

Tworzenie kompleksów kończy się z chwilą pokrycia wszystkich przykładów należących do dolnego lub górnego przybliżenia rozważanej klasy decyzyjnej. Z dolnych przybliżeń uzyskiwane są reguły pewne, a z górnych przybliżeń – reguły możliwe.

Aby wyznaczone pokrycie klasy decyzyjnej spełniało trzeci warunek lokalnego pokrycia klasy decyzyjnej, algorytm LEM2 realizuje procedurę optymalizacji zbioru kompleksów.

Optymalizacja polega na usunięciu z pokrycia nadmiarowych kompleksów (oraz wyznaczonych na ich podstawie reguł).

W standardowej wersji LEM2 działa jedynie w przestrzeni atrybutów dyskretnych, a wykorzystanie atrybutów ciągłych wymaga ich uprzedniej dyskretyzacji. Znane są jednak modyfikacje algorytmu, pozwalające na indukcję reguł bez wstępnej dyskretyzacji atrybutów ciągłych. Modyfikacje takie zawarte są w zaproponowanym przez Stefanowskiego algorytmie MODLEM [283] oraz w algorytmie MLEM2 [126]. Implementacje algorytmu LEM2 zawarto m.in. w systemach LERS [119] i RSES [17].

Stefanowski i Vanderpooten [285] opracowali algorytm indukcji reguł decyzyjnych, przeznaczonych do celów opisowych. Algorytm nosi nazwę **Explore**. Umożliwia on znalezienie wszystkich koniunkcji warunków elementarnych, spełniających zdefiniowane przez użytkownika wymogi minimalnej jakości (określające np. minimalną ogólność i minimalną dokładność reguł). Explore rozpoczyna proces indukcji od reguły zawierającej jeden warunek elementarny. Następnie realizowana jest faza wzrostu, w której przesłanka rozszerzana jest o kolejne warunki elementarne, przy czym algorytm działa zgodnie ze strategią przeszukiwania wszerz, ponieważ wymagania minimalnej jakości może spełnić każda z tworzonych reguł. Po dodaniu kolejnego warunku elementarnego sprawdza się, czy aktualna postać reguły spełnia wymagania minimalnej jakości; jeśli tak jest, zostaje ona umieszczona w wynikowym zbiorze reguł. Oznacza to także, że dla zbioru warunków elementarnych, zawartych w tej regule, oraz dla każdego jego nadzbioru poszukiwanie reguł zostało zakończone. Jeśli aktualna postać reguły nie spełnia wymagań minimalnej jakości, to algorytm sprawdza, czy jest szansa na to, aby reguła spełniła je po dodaniu kolejnych warunków. W standardowej wersji algorytmu jest to równoważne sprawdzeniu, czy reguła pokrywa wystarczającą liczbę przykładów pozytywnych. Jeśli szansy takiej nie ma, proces indukcji reguły kończy się niepowodzeniem. Podczas wzrostu reguły opisującej klasę decyzyjną  $X$  rozważane są jedynie warunki elementarne, pokrywające co najmniej jeden przykład z tej klasy.

Najbliższym idei sekwencyjnego pokrywania algorytmem indukcji reguł regresyjnych jest algorytm stosowany przez Janssena i Fürnkranza [144]. Nie ma on własnej nazwy, gdyż jest kopią stosowanego przez tych autorów algorytmu pokryciowego przeznaczonego do indukcji reguł decyzyjnych. Konkluzja reguł zawiera wyrażenie  $d=M$ , w którym  $M$  jest medianą wartości zmiennej decyzyjnej przykładów pokrywanych przez regułę. W fazie wzrostu miarą jakości reguły jest m.in. współczynnik korelacji. Aby możliwe było obliczenie korelacji pomiędzy przesłanką a konkluzją, konieczne jest określenie zbiorów przykładów pozytywnych i negatywnych. Autorzy przyjęli, że przykładami pozytywnymi będą wszystkie te przykłady, których wartość atrybutu decyzyjnego mieści się w przedziale  $[M - \sigma, M + \sigma]$ , gdzie  $\sigma$  jest odchyleniem standardowym, obliczanym na podstawie wartości zmiennej

decyzyjnej przykładów aktualnie pokrywanych przez regułę. Zauważmy, że wraz ze wzrostem reguły, wartości  $M$  i  $\sigma$  mogą się zmieniać. Oznacza to, że rozmiar „klasy decyzyjnej” również może ulegać zmianie. Proces wzrostu reguły kończy się z chwilą, gdy regułę pokrywają nie więcej niż 3 przykłady lub gdy wszystkie przykłady pokrywane przez regułę mieszczą się w przedziale  $[M - \sigma, M + \sigma]$ . Ta ostatnia sytuacja utożsamiana jest z osiągnięciem reguły dokładnej. Faza przycinania nie jest realizowana, a jako reguła wyjściowa pamiętana jest ta, która podczas wzrostu otrzymała najwyższą ocenę.

## 2.4. Algorytmy q-ModLEM oraz RMatrix

W niniejszym rozdziale przedstawione zostały dwa algorytmy indukcji reguł. Pierwszy z nich użyty zostanie w dalszej części monografii m.in. do badania wpływu, jaki ma zastosowana podczas indukcji miara jakości reguł na jakość wynikowego zbioru reguł. Reguły będące rezultatem działania tego algorytmu stanowią również podstawę dla omawianych w rozdziale 5 metod uogólniania i filtracji reguł. Prezentowane algorytmy autor stosował również do realizacji wielu zadań analitycznych, o niektórych z nich będzie mowa w rozdziale 6.

Pierwszy algorytm działa według zasady generowania kolejnych pokryć, nie wymaga dyskretyzacji atrybutów ciągłych, a w fazach wzrostu i przycinania stosuje strategię wspinaczki. Zasadniczy wpływ na wynikową postać reguły ma miara jakości, stosowana w fazach wzrostu i przycinania, przy czym ujęta miara może być dowolną ze znanych miar jakości reguł. Idea indukcji realizowana przez algorytm najbliższa jest idei zawartej w algorytmach LEM2 i MODLEM, z tego też powodu algorytmowi nadamy nazwę q-ModLEM. Litera q jest pierwszą literą słowa *quality* i akcentuje, że zasadniczy wpływ na postać wyznaczonych reguł ma zastosowana miara jakości.

Indukcja reguły rozpoczyna się od pustej przesłanki i konkluzji wskazującej na aktualnie rozważaną klasę decyzyjną (linie 7 i 8). Faza wzrostu realizowana jest tak długo, dopóki reguła nie stanie się dokładna lub dalsze dodawanie warunków elementarnych nie powoduje już udokładniania reguły (linie 11 i 19). Zmienna boolowska *growth* zabezpiecza przesłankę przed dodaniem do niej zbędnych warunków elementarnych. Wzrost reguły realizowany jest przez procedurę *Znajdz-najlepszy-warunek*, która dla każdego atrybutu  $a$  znajduje najlepszy warunek elementarny. W liniach 15, 16, 17 algorytm spośród najlepszych warunków znalezionych oddziennie dla każdego atrybutu wybiera ten, który jest generalnie najlepszy.

Jakość warunku elementarnego weryfikowana jest poprzez obliczenie (na całym zbiorze przykładów  $U$ ) wartości miary  $q$  dla reguły, której przesłanka zbudowana jest z wyrażenia  $\varphi$ , rozszerzonego o oceniany warunek (linie 17,18).

Pokryciowy algorytm indukcji reguł - q-ModLEM

Wejście:  $\text{DT} = (U, A \cup \{d\})$  – treningowa tablica decyzyjna,  $X_v$  – klasy decyzyjne,  $q$  – miara jakości, zastosowana do oceny tworzonych reguł  
 $\text{prec}$  – zmienna Boolowska decydująca o tym, czy wynikiem fazy wzrostu jest reguła dokładna, czy reguła o najwyższej jakości  $q$

Wyjście: RUL – zbiór wyznaczonych reguł

```

1  Begin
2      RUL:=∅;
3      Foreach  $X_v$  do // dla każdej klasy decyzyjnej
4          G:= $X_v$ ;
5          While  $X_v \neq \emptyset$  do
6              growth:=True;
7               $\varphi := \emptyset$  // stwórz przesłankę reguły bez żadnych
                    // warunków elementarnych, przesłankę spełniają wszystkie
                    // przykłady ze zbioru  $U$ 
8               $\psi := (d=v)$  // stwórz konkluzję reguły, wskazującą na rozważaną klasę
                    // decyzyjną  $X_v$ 
9               $r_{\text{best}} := \varphi \rightarrow \psi$  // reguła najlepsza w danej chwili
10              $q(r_{\text{best}}) := -\infty$ 

                // faza wzrostu reguły
11             While (not( $[\varphi] \subseteq G$ ) and growth) do
12                  $w_{\text{best}} := \emptyset$  // warunek elementarny, dodawany do przesłanki reguły
13                 Foreach  $a \in A$  do
14                      $w := \text{Znajdz-najlepszy-warunek}(a, \varphi, \psi, U, q)$ ;
15                      $a_1 := \varphi \wedge w$ ; // dołącz na próbę warunek  $w$  do przesłanki  $\varphi$ 
16                      $a_2 := \varphi \wedge w_{\text{best}}$ ; // dołącz na próbę warunek  $w_{\text{best}}$  do przesłanki  $\varphi$ 
17                     If  $q(a_1 \rightarrow \psi) > q(a_2 \rightarrow \psi)$  then  $w_{\text{best}} := w$  else
18                         if  $q(a_1 \rightarrow \psi) = q(a_2 \rightarrow \psi)$  then jako nowy  $w_{\text{best}}$  wybierz ten spośród
                                //  $w$  i  $w_{\text{best}}$ , który pokrywa więcej przykładów pozytywnych;
19                          $a_1 := \emptyset$ ;  $a_2 := \emptyset$ ;
20                 end foreach

                // czy dodanie warunku powoduje zwiększenie dokładności reguły
21                 If  $(([\varphi \wedge w_{\text{best}}] \cap (U - G)) \subset ([\varphi] \cap (U - G)))$ 
22                     then
23                          $\varphi := \varphi \wedge w_{\text{best}}$ ; // dodaj najlepszy warunek do przesłanki  $\varphi$ 
24                         // czy pamiętana jest reguła doklana czy reguła o najwyższej jakości
25                         If prec then  $r_{\text{best}} := \varphi \rightarrow \psi$ 
26                             else if  $q(\varphi \rightarrow \psi) > q(r_{\text{best}})$  then  $r_{\text{best}} := \varphi \rightarrow \psi$ 
27                         else
28                             growth:=False;
29                         end while

                // faza przycinania reguły
30                  $r_{\text{best}} := \text{Skroc-regule}(r_{\text{best}}, U, q)$ ;

31                 RUL := RUL ∪ { $r_{\text{best}}$ };
32                  $X_v := X_v - [r_{\text{best}}]$ ;
33                  $\text{Uprosc-zapis-reguly}(r_{\text{best}})$ 
34             end while
35              $X_v := G$ ;
36         end foreach

            // faza przycinania zbioru reguł
37             // Przytnij-zbior-regul(RUL);
38         end.

```

Najlepszy z rozważanych warunków elementarnych dodawany jest do przesłanki  $\varphi$  (linia 21). Jeśli wartość zmiennej  $\text{prec}$  równa jest  $true$ , to przesłanka  $\varphi$  staje się nową przesłanką

reguły (linia 22). Jeśli wartość zmiennej *prec* równa jest *false*, sprawdzane jest, czy rozszerzona reguła  $\varphi \rightarrow \psi$  charakteryzuje się wyższą jakością niż najlepsza z reguł dotychczas rozważanych w fazie wzrostu (linia 23). W tym przypadku efektem realizacji fazy wzrostu jest reguła, która uzyskała najwyższą ocenę. Po zakończeniu fazy wzrostu realizowana jest faza przycinania (procedura *Skroc-regule*), polegająca na usuwaniu zbędnych warunków elementarnych. Po skróceniu reguła dodawana jest do wynikowego zbioru reguł oraz ograniczany jest zbiór  $X_v$  przykładów pozytywnych, niepokrytych jeszcze przez żadną z dotychczas wyznaczonych reguł. Aktualna postać zbioru  $X_v$  ma znaczenie dla procedury *Znajdz-najlepszy-warunek* oraz decyduje o tym, czy proces indukcji kolejnych reguł ma być kontynuowany (linia 5).

Zadaniem procedury *Uprosc-zapis-reguły* jest uproszczenie zapisu warunków elementarnych, utworzonych na podstawie atrybutów ciągłych. Dodawanie warunków elementarnych odbywa się w taki sposób, że w przesłance reguły mogą znaleźć się wyrażenia  $a > v1 \wedge a < v2$ , w których  $v1 < v2$ . Jednym zadaniem procedury *Uprosc-zapis-reguły* jest identyfikacja takich wyrażeń i zamiana ich na warunek  $a \in (v1, v2)$ .

Po zakończeniu indukcji w obrębie danej klasy decyzyjnej przywracana jest wejściowa postać tej klasy (linia 32).

W linii 34 wymieniono procedurę *Przytnij-zbior-reguł*, której zadaniem jest ograniczenie zbioru wyznaczonych reguł. W rozdziale 5 omówiono kilka propozycji procedur przycinania zbiorów reguł. Opisując algorytm q-ModLEM, sygnalizujemy jedynie, że taki proces jest realizowany po zakończeniu indukcji wszystkich reguł. Realizacja procedury przycinania zbioru reguł nie jest jednak warunkiem koniecznym zakończenia działania algorytmu, dlatego też procedurę *Przytnij-zbior-reguł* umieszczono w komentarzu.

Do pełnego omówienia algorytmu q-ModLEM pozostało opisanie procedur *Znajdz-najlepszy-warunek* oraz *Skroc-regule*. Procedura *Znajdz-najlepszy-warunek* dla atrybutów symbolicznych i porządkowych generuje warunki elementarne o postaci *atrybut=wartość*, a dla atrybutów ciągłych – warunki *atrybut>wartość* lub *atrybut<wartość*. W szczególności jeśli atrybuty porządkowe zakodowane są za pomocą liczb całkowitych, to można potraktować je jako atrybuty ciągłe. Trzeba jednak mieć na uwadze, że uzyskane dla takich atrybutów granice zakresów warunków elementarnych nie będą miały sensownej interpretacji. Wiedząc jednak, w jaki sposób oryginalne wartości atrybutu zostały zakodowane, można sensownie interpretować tak uzyskane warunki.

Procedura - Znajdz-najlepszy-warunek;

Wejście:  $a$  - atrybut warunkowy,  $\varphi$  - aktualna postać przesłanki reguły,  $\psi$  - konkluzja reguły,  $U$  zbiór wszystkich przykładów,  $q$  - miara jakości, zastosowana do oceny reguł

Wyjście:  $w$  - aktualnie najlepszy warunek elementarny

```

1  Begin
2    If  $a$  jest atrybutem symbolicznym lub porządkowym
3      then
4         $V$ ; // stwórz zbiór wartości atrybutu  $a$ , występujących w zbiorze  $[\varphi]$ 
5         $eval\_w:=-\infty$ ; // ustal minimalną wartość oceny warunku elementarnego
6         $w:=\emptyset$ ; // utwórz pusty warunek elementarny
7        Foreach  $v \in V$  do // dla każdej wartości ze zbioru  $V$ 
8           $w\_temp:=(a=v)$  // utwórz tymczasowy warunek elementarny  $a=v$ 
9           $a:=\varphi \wedge w\_temp$ ; // utwórz przesłankę będącą koniunkcją  $\varphi$  i  $w\_temp$ 
10          $eval\_w\_temp:=q(a \rightarrow \psi)$ ;
11         If  $eval\_w\_temp > eval\_w$ 
12           then
13              $eval\_w:=eval\_w\_temp$ ;
14              $w:=w\_temp$ ;
15             else if (( $eval\_w\_temp = eval\_w$ ) and ( $[w \wedge \varphi \wedge \psi] \subset [w\_temp \wedge \varphi \wedge \psi]$ )) then
16                $w:=w\_temp$ ;
17                $w\_temp:=\emptyset$ 
18             end foreach
19           else
20             // atrybut jest typu ciągłego
21              $L$ ; // stwórz posortowaną rosnąco listę wartości atrybutu  $a$ , jakie
22             // występują w zbiorze  $[\varphi]$ ,  $L(i)$  jest  $i$ -tym elementem tej listy
23              $eval\_w:=-\infty$ ; // ustal minimalną wartość oceny warunku elementarnego
24              $w:=\emptyset$ ; // utwórz pusty warunek elementarny
25             Foreach  $i \in \{1, \dots, \text{length}(L)-1\}$  do
26                $v:=(L(i)+L(i+1))/2$  // ustal punkt graniczny dla zakresu warunku
27               // elementarnego; rozpatruj tylko te punkty, które
28               // rozdzielają przykłady należące do aktualnie
29               // rozpatrywanej klasy decyzyjnej od przykładów
30               // należących do innych klas
31                $w\_temp\_l:=(a < v)$ 
32                $w\_temp\_r:=(a > v)$ 
33                $a1:=\varphi \wedge w\_temp\_l$ ; // utwórz przesłankę jako koniunkcję  $\varphi$  i  $w\_temp\_l$ 
34                $a2:=\varphi \wedge w\_temp\_r$ ; // utwórz przesłankę jako koniunkcję  $\varphi$  i  $w\_temp\_r$ 
35                $eval\_w\_temp\_l:=q(a1 \rightarrow \psi)$ ;
36                $eval\_w\_temp\_r:=q(a2 \rightarrow \psi)$ ;
37               // wybierz najlepszy spośród warunków  $w\_temp\_l$  i  $w\_temp\_r$  i zapisz wyniki
38               // w  $w\_temp\_r$ 
39               If ( $eval\_w\_temp\_l > eval\_w\_temp\_r$ ) or
40               (( $eval\_w\_temp\_l = eval\_w\_temp\_r$ ) and ( $[w\_temp\_r \wedge \varphi \wedge \psi] \subset [w\_temp\_l \wedge \varphi \wedge \psi]$ ))
41               then
42                  $eval\_w\_temp\_r:=eval\_w\_temp\_l$ ;
43                  $w\_temp\_r:=w\_temp\_l$ ;
44               // zapamiętaj jako  $w$  lepszy z warunków  $w$  i  $w\_temp\_r$ 
45               If ( $eval\_w\_temp\_r > eval\_w$ ) or
46               (( $eval\_w\_temp\_r = eval\_w$ ) and ( $[w \wedge \varphi \wedge \psi] \subset [w\_temp\_r \wedge \varphi \wedge \psi]$ ))
47               then
48                  $eval\_w:=eval\_w\_temp\_r$ ;
49                  $w:=w\_temp\_r$ ;
50                  $w\_temp\_l:=\emptyset$ 
51                  $w\_temp\_r:=\emptyset$ 
52             end foreach
53         end foreach.

```

Dla atrybutów symbolicznych i porządkowych procedura `Znajdz-najlepszy-warunek` sprawdza wszystkie możliwe warunki elementarne  $a = v$  (linia 8) i wybiera spośród nich ten, który maksymalizuje jakość reguły będącej koniunkcją ocenianego warunku i aktualnej przesłanki  $\varphi$  (linie od 11 do 16). Rozważane są jedynie te wartości atrybutu  $a$ , które występują w zbiorze przykładów pokrytych przez aktualną postać przesłanki reguły  $[\varphi]$ . Jakość reguły obliczana jest na całym zbiorze przykładów  $U$  (linia 10).

Dla atrybutów ciągłych zasada wyboru najlepszego warunku jest taka sama jak dla atrybutów symbolicznych. Najlepszy jest ten warunek, który maksymalizuje jakość reguły utworzonej z aktualnej przesłanki i ocenianego warunku (linie od 31 do 38). W nieco inny sposób ustalana jest postać warunków elementarnych, które są wyrażeniami  $a > v$  lub  $a < v$ . Przed sprawdzeniem jakości warunków elementarnych wszystkie aktualnie rozważane wartości atrybutu  $a$  sortowane są rosnąco i umieszczane na liście  $L$  (linia 20). Wartość  $v$  to tzw. punkt graniczny, leżący pośrodku, pomiędzy dwoma kolejnymi wartościami  $L(i)$  i  $L(i+1)$  (linia 24). W praktyce rozpatruje się jednie te pary wartości atrybutu, które oddzielają przykład pozytywny od przykładu negatywnego. Mając ustalony punkt graniczny, można utworzyć dwa warunki elementarne:  $a < v$  (`w_temp_1` linia 25),  $a > v$  (`w_temp_r` linia 26). Następnie weryfikowana jest jakość dwóch reguł; pierwsza zawiera przesłankę będącą koniunkcją warunku `w_temp_1` i przesłanki  $\varphi$ , druga jest koniunkcją `w_temp_r` i  $\varphi$ . Ocena warunku elementarnego równa jest wartości miary jakości  $q$ , obliczonej dla reguły zawierającej ten warunek (linie 29, 30). Jeśli ocena lepszego spośród warunków `w_temp_1` i `w_temp_r` jest wyższa od oceny najlepszego z dotychczas rozpatrywanych warunków, to zostaje on zapamiętany jako aktualnie najlepszy warunek elementarny (linie 33, 34 lub 38, 39). Jeśli jakość warunków jest identyczna, to lepszy jest ten, który w połączeniu z  $\varphi$  pokrywa więcej przykładów pozytywnych (linie 15, 32, 36).

Procedura `Skroc-regule` realizuje prosty algorytm usuwania warunków elementarnych. Warunki usuwane są dopóty, dopóki jakość skróconej reguły nie jest gorsza od jakości reguły wejściowej. Poszukiwanie najlepszego w danej chwili warunku do usunięcia (warunek ten oznaczany jest jako `alpha_best`) realizowane jest za pomocą strategii wspinaczki. Jeśli po usunięciu warunku elementarnego jakość reguły, obliczana na całym zbiorze przykładów  $U$ , nie maleje, oznacza to, że warunek może zostać usunięty (linia 17). Jeśli po usunięciu warunku jakość reguły wzrasta, to staje się ona bazową jakością, w odniesieniu do której realizowany będzie dalszy proces skracania (linia 18). W szczególnych przypadkach procedura `Skroc-regule` może zostać zmodyfikowana

w taki sposób, aby dopuszczała do nieznacznego (ustalonego przez użytkownika) pogorszenia się jakości skracanej reguły.

Algorytm q-ModLEM stosowany był do analizy ogólnodostępnych danych benchmarkowych, gdzie porównywano jego efektywność m.in. z algorytmami MODLEM, RIPPER, PART oraz systemami tworzącymi reguły na podstawie reductów [251, 253, 254, 264]. W pracach [252, 258, 268] przedstawiono również wyniki zastosowania algorytmu do prognozowania zagrożeń sejsmicznych w kopalniach węgla kamiennego, diagnostyki urządzeń oraz analizy czynników wpływających na czas przeżycia pacjentów po przeszczepie szpiku kostnego.

```

Procedura - Skroc-regule;
Wejście:  $r=\varphi \rightarrow \psi$  - reguła poddawana skracaniu,  $U$  - zbiór wszystkich przykładów,
 $q$  - miara jakości zastosowana do oceny reguł
Wyjście:  $r_{sh}$  - skrócona reguła
1 Begin
2   eval_max:= $q(\varphi \rightarrow \psi)$ ;
3   while eval_max≤eval_short do
4      $\alpha_{best}:=\emptyset$ ;
5     eval_alpha_best:=-∞;
6     foreach  $w \in \varphi$  do // dla każdego warunku elementarnego, należącego do
      // przesłanki  $\varphi$ 
7        $\alpha:=\varphi$ ; usuń z  $\alpha$  warunek  $w$ ;
8       eval_alpha:= $q(\alpha \rightarrow \psi)$ ;
9       If eval_alpha>eval_alpha_best
10        then
11           $\alpha_{best}:=w$ ;
12          eval_alpha_best:= eval_alpha;
13           $\alpha:=\emptyset$ 
14        end foreach
15        If eval_alpha_best≥eval_max
16        then
17          Usuń z  $\varphi$  warunek  $\alpha_{best}$ ;
18          eval_max:= $q(\varphi \rightarrow \psi)$ ;
19        end while
20         $r_{sh}:=\varphi \rightarrow \psi$ ;
21 end.
```

W części z przywoływanych prac autor wykorzystywał wersję algorytmu q-ModLEM, w której w fazach wzrostu i przycinania stosowane były różne miary oceniające. Ponadto w celu przyspieszenia obliczeń faza wzrostu kończyła się nie z chwilą uzyskania reguły możliwie najdokładniejszej lub najlepszej (o najwyższej jakości), ale kiedy jakość specjalizowanej reguły zaczynała spadać. Genezą takiej modyfikacji było spostrzeżenie, że dla wielu zbiorów danych funkcja odzwierciedlająca jakość reguły w kolejnych krokach specjalizacji miała jedno maksimum. Różnice całkowitej dokładności klasyfikacji pomiędzy różnymi konkretyzacjami algorytmu q-ModLEM nie były jednak znaczące. Zasadnicze różnice dotyczyły liczby wyznaczanych reguł (patrz rozdział 4).

Zauważmy, że algorytm q-ModLEM można zastosować również do indukcji reguł regresyjnych [267]. Przyjmijmy założenie, że w konkluzjach reguł znajdują się wyrażenia

$d = v$ , gdzie  $v$  jest wartością definiowaną na podstawie zbioru przykładów, aktualnie pokrywanego przez regułę. Założymy również, że ustalone są wartości progowe  $\varepsilon_1, \varepsilon_2$ , pozwalające na podstawie  $[v - \varepsilon_1, v + \varepsilon_2]$  zdefiniować zbiory przykładów pozytywnych i negatywnych oraz pozytywnych i negatywnych, pokrywanych przez regułę. Niezależnie od tego, czy wartości  $\varepsilon_1, \varepsilon_2$  definiowane są w sposób statyczny czy dynamiczny (wraz ze zmieniającą się postacią reguły), możliwe jest zastosowanie algorytmu q-ModLEM do indukcji reguł regresyjnych. Indukcja będzie przebiegać w jednym wykonaniu głównej pętli (linia 3), gdyż w danych nie ma wydzielonych klas decyzyjnych.

Drugi z proponowanych algorytmów wzorowany jest na metodach indukcji reguł, stosowanych w zbiorach przybliżonych. Algorytm nie działa na zasadzie generowania kolejnych pokryć, lecz tworzy jedną regułę dla każdego przykładu znajdującego się w tablicy treningowej.

Pierwotna wersja algorytmu została opracowana dla tolerancyjnego modelu zbiorów przybliżonych [227, 271]. W modelu tym zamiast relacji nieroróżnialności użyto relacji podobieństwa. Aby możliwe było określenie, które przykłady są do siebie podobne, należy ustalić, w jaki sposób będzie mierzona odległość pomiędzy wartościami atrybutów. Konieczne jest także ustalenie wartości tzw. progów tolerancji, określających dopuszczalne odległości pomiędzy wartościami podobnymi. Wartości progów tolerancji powinny być dobierane w taki sposób, aby za podobne zostało uznanych jak najwięcej par przykładów, należących do identycznych klas decyzyjnych, oraz za niepodobne zostało uznanych jak najwięcej par przykładów, należących do różnych klas decyzyjnych. Do wyznaczania progów tolerancji stosowano różnego rodzaju algorytmy przeszukiwania i optymalizacji [166, 208, 290, 292, 295]. Autor przedstawił propozycję wykorzystania do oceny jakości wektora progów tolerancji odpowiednio zaadaptowanych miar, stosowanych do oceny jakości reguł decyzyjnych [254]. Pozwoliło to na tworzenie klasyfikatorów złożonych ze stosunkowo niewielkiej liczby reguł.

W prezentowanym dalej algorytmie pominięto część związaną z ustalaniem wartości progów tolerancji. Przyjęto, że wartości te są dla każdego atrybutu równe 0. Oznacza to, że za podobne uznawane są jedynie identyczne wartości atrybutów. Z tego też powodu prezentowany algorytm operuje na atrybutach typu dyskretnego lub poddanych dyskretyzacji atrybutach ciągłych. Konsekwencją przyjętych założeń jest dopuszczenie w przesłankach reguł jedynie warunków elementarnych o postaci *atrybut=wartość*.

Reguły generowane są kolejno dla każdego przykładu treningowego, konkluzja wskazuje na klasę decyzyjną, przypisaną do aktualnie rozpatrywanego przykładu  $x$  (linia 6). Zasadnicze znaczenie dla postaci przesłanki ma procedura *Utworz-ranking-atrybutów*, która dla  $x$  tworzy listę atrybutów warunkowych. Pierwszy na liście znajduje

się atrybut, który odróżnia najwięcej par przykładów  $\langle x, x_j \rangle$ , gdzie  $x_j$  jest przykładem należącym do klasy decyzyjnej innej niż  $x$ . Ostatni na liście jest atrybut odróżniający najmniejszą liczbę takich par. W pierwotnej wersji algorytmu ranking atrybutów tworzony był na podstawie analizy wiersza tzw. uogólnionej macierzy odróżnialności [292], zawierającego informację o odróżnialności przykładu  $x$  od przykładów reprezentujących klasy decyzyjne inne niż  $x$  (stąd nazwa algorytmu *Row of indiscernibility Matrix*). W obecnej wersji algorytm wykorzystuje twierdzenie prezentowane w pracy [208], pozwalające na wyznaczenie rankingu atrybutów bez tworzenia macierzy odróżnialności. Ogranicza to złożoność obliczeniową procesu tworzenia reguły.

Załóżmy, że dane są: przykład generator  $x$  oraz atrybut  $a$ . Przez  $W_a^x$  oznaczamy liczbę par przykładów  $\langle x, x_j \rangle$ , takich że  $a(x) \neq a(x_j)$  i  $d(x) \neq d(x_j)$ . Przez  $[IND(\{a\})]/U = \{W^1, W^2, \dots, W^j\}$  oznaczamy zbiór wszystkich klas abstrakcji relacji  $IND(\{a\})$ . Ponieważ relacja  $IND$  jest relacją równoważności, zbiory te tworzą podział zbioru przykładów  $U$ . Do jednego z tych zbiorów należy wartość  $a(x)$ ; oznaczmy ten zbiór jako  $W^{i0}$ . Wówczas wartość  $W_a^x$  można obliczyć zgodnie ze wzorem (2.9):

$$W_a^x = \sum_{k \in \{1, 2, \dots, j\} - \{i0\}} 1 \left| \left( W^k \cap (U - X_{d(x)}) \right) \right| \quad (2.9)$$

We wzorze (2.9) liczba 1 akcentuje to, że pierwszy przykład w parze  $\langle x, x_j \rangle$  jest ustalony. Zbiór  $X_{d(x)}$  to klasa decyzyjna, do której należy przykład generator  $x$ . Do zbioru  $W^k \cap (U - X_{d(x)})$  należą przykłady charakteryzowane przez inną niż  $x$  wartość atrybutu  $a$ , które jednocześnie nie należą do klasy decyzyjnej  $X_{d(x)}$ .

Obliczając wartość wyrażenia (2.9), a następnie sortując uzyskane wyniki zgodnie z porządkiem nierośącym, uzyskamy interesujący nas ranking atrybutów. Przyjmuje się, że w rankingu tym kolejność atrybutów o identycznych wartościach  $W_a^x$  jest dowolna.

Zgodnie z kolejnością atrybutów na liście  $R$ , do przesłanki reguły dodawane są kolejne warunki elementarne  $a = v$ . W warunkach tych  $a$  jest kolejnym atrybutem z listy oraz  $v = a(x)$  (linia 12). Każdorazowo po dodaniu nowego warunku elementarnego aktualna postać reguły jest oceniana (linia 16). W  $r\_best$  przechowana jest reguła o najwyższej jakości (linia 16). Reguła  $r\_best$  po zakończeniu fazy wzrostu przekazywana jest do procedury skracania.

Procedura skracania przebiega w sposób identyczny jak w przypadku algorytmu q-ModLEM. Przed umieszczeniem reguły w wynikowym zbiorze reguł algorytm sprawdza, czy nie znajduje się w nim identyczna reguła, wyznaczona na podstawie innego przykładu

(linia 23). Podobnie jak w algorytmie q-ModLEM, ostatnim, niewymaganym etapem jest przycinanie zbioru reguł.

```

Algorytm indukcji reguł - RMatrix
Wejście: DT=(U,A∪{d}) – treningowa tablica decyzyjna, A – zbiór atrybutów warunkowych dyskretnych lub po dyskretyzacji, q – miara jakości, zastosowana do oceny tworzonych reguł
Wyjście: RUL – zbiór wyznaczonych reguł
1 Begin
2   RUL:=∅;
3   Foreach x∈U do // dla każdego przykładu należącego do U
4     growth:=True;
5     φ:=∅ // stwórz przesłankę reguły, niezawierającą żadnych
           // warunków elementarnych, przesłankę spełniają wszystkie
           // przykłady ze zbioru U
6     ψ:=(d=d(x)) // stwórz konkluzję reguły, wskazującą na klasę decyzyjną,
           // do której należy przykład x
           // do rozważanej aktualnie klasy decyzyjnej
7     r_best:=φ→ψ // najlepsza w danej chwili reguła
8     q(r_best) :=-∞

           // faza wzrostu reguły
9     R:=Utworz-ranking-atrybutow(x, A, U); // R jest listą atrybutów
           // warunkowych
10    Foreach i∈{1,...,|A|} do
11      While (not([φ]⊆G) and growth) do
12        w:=(aR(i)=aR(i)(x)); // utwórz warunek elementarny a=v,
           // gdzie a=aR(i) oraz v=aR(i)(x)
           // czy zwiększyła się dokładność reguły
13        If (([φ∧w]∩(U-G))⊂([φ]∩(U-G)))
14          then
15            φ:=φ∧w // dodaj warunek w do przesłanki φ
16            If q(φ→ψ)>q(r_best) then r_best:=φ→ψ
17          else
18            growth:=False;
19          end while
20        end foreach

           // faza przycinania reguły
21        r_best:=Skroc-regułe(r_best, U, q);
22
23        If (not r_best∈RUL) then RUL:=RUL∪{r_best};
24      end foreach

           // faza przycinania zbioru reguł
25      // Przytnij-zbior-reguł(RUL);
26  end.

```

Algorytm RMatrix reprezentuje swego rodzaju podejście kompromisowe pomiędzy indukcją wszystkich minimalnych reguł decyzyjnych a generowaniem reguł na zasadzie kolejnych pokryć. W założeniu tego algorytmu leży przekonanie, że warto podjąć próbę wygenerowania z każdego przykładu treningowego przynajmniej jednej reguły o maksymalnie najwyższej jakości. Ograniczenie liczby reguł może być realizowane na etapie przycinania zbioru reguł. Klasyfikatory uzyskane przez algorytm RMatrix porównywano z klasyfikatorami wyznaczanymi na podstawie reduktów względnych, a także z algorytmem

q-ModLEM [254]. W pracy [262] zaprezentowano wyniki zastosowania teorii zbiorów przybliżonych i algorytmu RMatrix do rozwiązywania problemu prognozowania zagrożeń sejsmicznych i metanowych w kopalniach.

## 2.5. Klasyfikator regułowy

Na podstawie wyznaczonych reguł decyzyjnych możliwe jest zdefiniowanie klasyfikatora regułowego, którego zadaniem jest przyporządkowanie przykładów do odpowiadających im klas decyzyjnych. Klasyfikowane przykłady zazwyczaj pochodzą spoza tablicy decyzyjnej, na podstawie której wyznaczano reguły. Oczywiście możliwe jest również klasyfikowanie przykładów należących do tablicy treningowej.

W praktycznych zastosowaniach klasyfikacji podlegają nowe, nieznane dotychczas przykłady, dla których znamy jedynie wartości atrybutów warunkowych. Przyporządkowanie przykładu do klasy decyzyjnej jest jedynie prognozą dotyczącą wartości atrybutu decyzyjnego. Jeśli w tablicy treningowej nie zawarto wszystkich możliwych przykładów, jakie można utworzyć na podstawie atrybutów warunkowych, to klasyfikacja przykładów spoza tablicy treningowej jest procesem wnioskowania indukcyjnego, które jest wnioskowaniem zawodnym.

Proces klasyfikacji przebiega zgodnie z pewnym ustalonym algorytmem, który nazywany jest algorytmem decyzyjnym lub schematem klasyfikacji. W klasyfikatorach regułowych najczęściej wykorzystywany jest jeden z trzech schematów: klasyfikacja za pomocą listy reguł, klasyfikacja przez reguły o największym zaufaniu lub głosowanie (ang. *voting*).

W klasyfikacji przeprowadzanej za pomocą listy reguł przykład przyporządkowany jest do klasy decyzyjnej, na którą wskazuje pierwsza ze znajdujących się na liście reguł, która pokrywa klasyfikowany przykład. Założmy, że dany jest przykład  $x$  oraz lista reguł  $L = \langle r_1, r_2, \dots, r_l \rangle$ , przez  $\varphi_{r_i}$  oraz  $\psi_{r_i}$  oznaczamy odpowiednio przesłankę i konkluzję reguły  $r_i$ . Przyporządkowanie przykładu do zbioru  $[\psi_{r_i}]$  jest równoznaczne z przyporządkowaniem przykładu do klasy decyzyjnej, wskazywanej w konkluzji reguły  $r_i$ . Schemat klasyfikacji, stosujący listę reguł, można przedstawić jako ciąg wyrażeń warunkowych.

```

Jeśli  $x \in [\varphi_{r_1}]$ , to  $x \in [\psi_{r_1}]$ 
    wpp. jeśli  $x \in [\varphi_{r_2}]$ , to  $x \in [\psi_{r_2}]$ 
        wpp.
            ...
                wpp. jeśli  $x \in [\varphi_{r_l}]$ , to  $x \in [\psi_{r_l}]$ 
                    // wpp.  $x \in [\psi_{\text{default}}]$ .

```

Na końcu ciągu wyrażeń warunkowych umieszczono przyporządkowanie przykładu do klasy decyzyjnej, wskazywanej przez regułę domyślną. Zazwyczaj reguła ta wskazuje na najliczniejszą klasę decyzyjną. Klasyfikacja przykładu przez regułę domyślną jest opcjonalna, z tego też powodu odpowiadający jej fragment pseudokodu został umieszczony w komentarzu. Jeśli reguła domyślna nie jest wykorzystana, przykłady niepokryte przez żadną z reguł tworzących listę nie zostaną sklasyfikowane.

Podczas klasyfikacji przebiegającej zgodnie ze schematem największego zaufania lub głosowania reguły dzielone są na grupy wskazujące na identyczną klasę decyzyjną. Z każdą regułą związana jest wartość liczbową, odzwierciedlającą poziom zaufania, jakim obdarzana jest reguła. Poziom zaufania to najczęściej wartość liczbową, obliczana na podstawie treningowego zbioru przykładów. Podczas klasyfikacji możemy mieć do czynienia z jedną z następujących sytuacji:

1. klasyfikowany przykład pokrywany jest przez reguły wskazujące na jedną klasę decyzyjną,
2. klasyfikowany przykład pokrywany jest przez reguły wskazujące na więcej niż jedną klasę decyzyjną,
3. klasyfikowany przykład nie jest pokrywany przez żadną z reguł biorących udział w klasyfikacji.

W sytuacji 1. decyzja podejmowana przez klasyfikator jest jednoznaczna, przykład zostaje zaklasyfikowany do klasy wskazywanej przez pokrywające go reguły. W sytuacji 2. występuje konflikt, który rozstrzygany jest na podstawie informacji o zaufaniu do reguł pokrywających klasyfikowany przykład. Oznaczmy przez  $conf(r)$  poziom zaufania do reguły  $r$  oraz przez  $RUL_v$  zbiór reguł, których konkluzje wskazują na klasę decyzyjną  $X_v$ . Dla każdego klasyfikowanego przykładu  $x$  obliczany jest poziom zaufania  $x$  do każdej z klas decyzyjnych, oznaczany jako  $conf(x, X_v)$ . W schemacie największego zaufania  $conf(x, X_v)$  wyznacza się zgodnie ze wzorem (2.10), a w głosowaniu – zgodnie ze wzorem (2.11).

$$conf(x, X_v) = \max\{conf(r) : x \in [\varphi r] \wedge r \in RUL_v\}, \quad (2.10)$$

$$conf(x, X_v) = \sum_{r \in RUL_v : x \in [\varphi r]} conf(r). \quad (2.11)$$

Przykład klasyfikowany jest do tej klasy decyzyjnej, dla której wartość  $conf(x, X_v)$  jest największa. W takiej sytuacji, w której niemożliwe jest wskazanie klasy decyzyjnej o największym zaufaniu (bo można wskazać dwie takie klasy lub większą ich liczbę), przykład nie podlega klasyfikacji lub przyporządkowany jest w sposób losowy do jednej z klas o największym poziomie zaufania. Warto wspomnieć, że najczęściej rozróżnia się dwa

typy głosowania: proste (ang. *simple voting*) oraz ważone (ang. *weighted voting*). W głosowaniu prostym dla każdej reguły biorącej udział w klasyfikacji  $conf(r)=1$ .

W sytuacji 3., gdy przykład nie jest rozpoznawany przez żadną z reguł biorących udział w klasyfikacji, najprostszym rozwiązaniem jest uznanie go za niesklasyfikowany lub przyporządkowanie go do najliczniejszej klasy. Inne rozwiązania bazują na informacji o tzw. częściowym dopasowaniu przykładu do reguł lub o odległości przykładu od reguł biorących udział w klasyfikacji.

W dalszej części monografii do klasyfikacji przykładów wykorzystano głosowanie ważone oraz schemat największego zaufania. Poziom zaufania do reguł utożsamiany jest z ich jakością, a jakość oceniana jest za pomocą jednej ze znanych z literatury miar (podrozdział 3.1.6). Klasyfikacja odbywa się zawsze przez reguły całkowicie dopasowane, a przykład nierozpoznany przez żadną z reguł traktowany był jak źle sklasyfikowany.

W pracach Grzymały-Busse, Bruhy oraz Ślęzaka i Widza [42, 120, 305, 306] badano różnice pomiędzy różnymi schematami klasyfikacji, wykorzystującymi m.in. siłę reguł i częściowe dopasowanie. Interesującą metodykę, którą można również stosować do rozstrzygania konfliktów klasyfikacji, przedstawia Wakulicz-Deja wraz ze współpracownikami [214, 326]. Jej podstawą jest grupowanie reguł i przeprowadzanie klasyfikacji (lub ogólniej – wnioskowania) na podstawie reguł należących do grupy najbardziej podobnej do klasyfikowanego przykładu. Metoda ta umożliwia także wnioskowanie na podstawie częściowego dopasowania. Autorzy podkreślają, że w dużych zbiorach reguł (np. wyznaczonych na podstawie reduktów względnych) proponowany przez nich sposób postępowania znacznie skraca czas podejmowania decyzji.

„Klasyfikacja” przykładów przez zbiór reguł regresyjnych odbywa się na podobnych zasadach. W konkluzji reguły regresyjnej znajduje się stała lub funkcja pozwalająca obliczyć wartość atrybutu decyzyjnego. Gdy używana jest lista reguł lub schemat największego zaufania, decyzję o przyporządkowaniu wartości atrybutu decyzyjnego podejmuje jedna reguła. W schemacie największego zaufania przyjmuje się, że każda reguła wskazuje na inną „klasę decyzyjną”. W schemacie głosowania sytuacja jest nieco inna. Zamiast ustalać stopnie zaufania przykładu testowego do klas decyzyjnych, decyzja o przyporządkowaniu wartości atrybutu decyzyjnego podejmowana jest najczęściej na podstawie:

- średniej ważonej wartości atrybutu decyzyjnego, obliczonej na podstawie przykładów treningowych, pokrywanych przez reguły pokrywające przykład testowy;
- mediany wartości atrybutu decyzyjnego, obliczonej na podstawie przykładów treningowych, pokrywanych przez reguły pokrywające przykład testowy.

W pierwszym przypadku wagi przyporządkowane poszczególnym regułom mogą być równe lub wprost proporcjonalne do jakości reguł. Reguła domyślna wskazuje na medianę

wartości atrybutu decyzyjnego wszystkich lub jedynie niepokrytych przykładów treningowych [144].

## 2.6. Ocena klasyfikatorów

Podstawowym celem systemu klasyfikującego (klasyfikatora) jest jak najlepsza klasyfikacja danych. Założmy, że dana jest tablica decyzyjna  $T=(U, A \cup \{d\})$ , w której  $V_d = \{v_1, v_2, \dots, v_k\}$ . Klasyfikator można utożsamić z pewną funkcją decyzyjną (2.12), która każdemu przykładowi należącemu do zbioru  $U$  przyporządkowuje odpowiadającą mu wartość atrybutu decyzyjnego. Jest to równoważne z zaklasyfikowaniem przykładu do jednej z klas decyzyjnych, występujących w tablicy  $T$ , przy czym przyporządkowanie to odbywa się na podstawie informacji o wartościach atrybutów warunkowych klasyfikowanego przykładu.

$$f : U \rightarrow V_d \cup \{c\}. \quad (2.12)$$

Zbiór wartości funkcji decyzyjnej rozszerzony jest o stałą  $c \notin V_d$ . Klasyfikator przyporządkowuje przykładowi wartość stałej  $c$  wtedy, kiedy zastosowany w nim algorytm decyzyjny nie jest w stanie przyporządkować przykładu do żadnej z klas decyzyjnych. Wartość stałej  $c$  jest nieistotna, ważne jest jedynie, żeby nie była to żadna z wartości atrybutu decyzyjnego.

Zwróćmy uwagę na wyraźne rozgraniczenie terminów: klasyfikator i algorytm decyzyjny (schemat klasyfikacji). Na klasyfikator składają się model danych oraz schemat klasyfikacji, który wykorzystuje ten model do przyporządkowania przykładów do klas decyzyjnych. W klasyfikatorach regułowych model danych utożsamiany jest ze zbiorem reguł, a schemat klasyfikacji decyduje o sposobie, w jakim reguły używane są do klasyfikacji.

### 2.6.1. Miary oceny klasyfikatorów

Podstawową miarą oceny klasyfikatora jest łączna dokładność klasyfikacji (ang. *overall classification accuracy*), potocznie zwana dokładnością klasyfikacji (2.13).

$$Acc(f, T) = \frac{|\{x \in U : f(x) = d(x)\}|}{|U|}. \quad (2.13)$$

W liczniku wyrażenia (2.13) obliczana jest moc zbioru przykładów, które zostały poprawnie przyporządkowane do odpowiadających im klas decyzyjnych. Aby obliczyć dokładność klasyfikacji, potrzebujemy pewnej tablicy decyzyjnej  $T$ , w której każdy przykład charakteryzowany jest przez wartość atrybutu decyzyjnego. Wartość ta jest konfrontowana z wartością przyporządkowaną przykładowi przez klasyfikator. Zamiast stosować dokładność klasyfikacji, można posługiwać się błędem klasyfikacji  $Err(f, T) = 1 - Acc(f, T)$ . Dokładność i błąd klasyfikacji mogą być wyrażane w procentach.

Dokładność klasyfikacji jest dobrą miarą oceny klasyfikatora, jeśli błędne decyzje podejmowane przez klasyfikator mają równe znaczenie. Oznacza to, że nie jest istotne, do której klasy decyzyjnej naprawdę należy błędnie sklasyfikowany przykład.

W wielu zastosowaniach błędy klasyfikacji nie mają równego znaczenia. Najczęściej przytaczanym przykładem jest medycyna i klasyfikacja pacjentów jako zdrowych lub chorych. Przyporządkowanie osoby chorej do grupy osób zdrowych jest bardziej niebezpieczne (kosztowne w sensie błędu klasyfikacji) niż błąd polegający na zaklasyfikowaniu pacjenta zdrowego jako chorego. W sytuacjach, w których błędy popełniane przez klasyfikator nie mają równego znaczenia, mówi się, że koszty błędnej klasyfikacji przykładów do poszczególnych klas decyzyjnych są różne. Z różnymi kosztami błędnej klasyfikacji możemy spotykać się także w prognozowaniu zagrożeń naturalnych, diagnostyce urządzeń, analizie finansowej itd. Problem różnego znaczenia błędnej klasyfikacji dotyczy także danych o nierównomiernym rozkładzie przykładów reprezentujących klasy decyzyjne. Dla tego typu danych ocena klasyfikatora za pomocą całkowitej dokładności klasyfikacji prowadzi do preferowania klasyfikatorów dobrze klasyfikujących przykłady z klas większościowych oraz błędnie klasyfikujących przykłady z klas mniejszościowych. Całkowita dokładność klasyfikacji nie jest tutaj dobrą miarą, gdyż nie bierze ona pod uwagę faktu, że rozkład przykładów jest nierównomierny [228]. Gdy koszty błędnej klasyfikacji nie są dla każdej klasy decyzyjnej identyczne, najrozsądniej jest ocenić klasyfikator ze względu na każdą klasę decyzyjną oddzielnie. Zbiory zdefiniowane wyrażeniami (2.14–2.17) stosuje się w definicjach miar oceniających jakość klasyfikatora ze względu na konkretną klasę decyzyjną.

$$TP(f, X_v, T) = \{x \in U : f(x) = v = d(x)\}, \quad (2.14)$$

$$FP(f, X_v, T) = \{x \in U : f(x) = v \neq d(x)\}, \quad (2.15)$$

$$TN(f, X_v, T) = \{x \in U : f(x) \neq v \neq d(x)\}, \quad (2.16)$$

$$FN(f, X_v, T) = \{x \in U : f(x) \neq v = d(x)\}. \quad (2.17)$$

Do zbioru  $TP$  (ang. *True Positive*) należą te przykłady z klasy  $X_v$ , które zostały poprawnie zaklasyfikowane przez klasyfikator  $f$ . Do zbioru  $FP$  (ang. *False Positive*) należą te przykłady spoza klasy  $X_v$ , które przez  $f$  zostały błędnie zaklasyfikowane do  $X_v$ . Zbiór  $TN$  (ang. *True Negative*) składa się z przykładów nienależących do klasy  $X_v$ , które przez  $f$  zostały również zaklasyfikowane do innych niż  $X_v$  klas decyzyjnych. Do zbioru  $FN$  (ang. *False Negative*) należą te przykłady z klasy  $X_v$ , które przez klasyfikator  $f$  zostały błędnie przyporządkowane do innych klas decyzyjnych.

Wskaźnik  $TPR$  (2.18) (ang. *true positive rate*) informuje o tym, ile spośród klasyfikowanych przykładów, należących do klasy  $X_v$ , zostało do niej przyporządkowanych w sposób poprawny. Wskaźnik  $TPR$  można nazwać dokładnością klasy decyzyjnej  $X_v$ . Wskaźnik  $FPR$  (2.19) (ang. *false positive rate*) informuje o tym, ile spośród przykładów nienależących do klasy  $X_v$  zostało błędnie zaklasyfikowanych jako należące do tej klasy decyzyjnej (ten rodzaj błędnej klasyfikacji nazywany jest również *falszywym alarmem*).

Wskaźnik  $Sens$  (2.20) zwany jest wrażliwością lub czułością (ang. *sensitivity*) klasyfikatora ze względu na klasę decyzyjną  $X_v$ . Określa on zdolność klasyfikatora do wykrywania przykładów reprezentujących klasę  $X_v$ . Innymi słowy, wrażliwość to prawdopodobieństwo warunkowe klasyfikacji przykładu do wybranej klasy, pod warunkiem że przykład rzeczywiście do tej klasy należy.

$$TPR(f, X_v, T) = \frac{|TP(f, X_v, T)|}{|X_v|}, \quad (2.18)$$

$$FPR(f, X_v, T) = \frac{|FP(f, X_v, T)|}{|(U - X_v)|}, \quad (2.19)$$

$$Sens(f, X_v, T) = TPR(f, X_v, T) = \frac{|TP(f, X_v, T)|}{|TP(f, X_v, T)| + |FN(f, X_v, T)|}, \quad (2.20)$$

$$Spec(f, X_v, T) = 1 - FPR(f, X_v, T) = \frac{|TN(f, X_v, T)|}{|TN(f, X_v, T)| + |FP(f, X_v, T)|}, \quad (2.21)$$

$$PPV(f, X_v, T) = \frac{|TP(f, X_v, T)|}{|TP(f, X_v, T)| + |FP(f, X_v, T)|}, \quad (2.22)$$

$$NPV(f, X_v, T) = \frac{|TN(f, X_v, T)|}{|TN(f, X_v, T)| + |FN(f, X_v, T)|}. \quad (2.23)$$

Znaczennością lub specyficznością (ang. *specificity*) klasyfikatora ze względu na klasę decyzyjną  $X_v$  nazywamy wskaźnik określony wzorem (2.21). Specyficzność określa, na ile decyzja o przyporządkowaniu przykładu do wybranej klasy jest charakterystyczna wyłącznie dla tej klasy (specyficzność jest uzupełnieniem prawdopodobieństwa klasyfikacji przykładu do klasy  $X_v$ , pod warunkiem że przykład do tej klasy nie należy).

Wskaźnik  $PPV$  (2.22) (ang. *positive predictive value*) informuje o tym, jakie jest prawdopodobieństwo, że przykład zaklasyfikowany do klasy  $X_v$  rzeczywiście do niej należy.

Wskaźnik  $NPV$  (2.23) (ang. *Negative predictive value*) informuje o tym, jakie jest

prawdopodobieństwo, że przykład niezaklasyfikowany do klasy  $X_v$  rzeczywiście do niej nie należy.

Każdy z prezentowanych powyżej wskaźników obliczany jest dla konkretnej klasy decyzyjnej. Szczególnym rodzajem klasyfikatorów są klasyfikatory binarne, w których występują jedynie dwie klasy decyzyjne. Dla klasyfikatorów binarnych wartości wskaźników (2.18–2.23) obliczane są zazwyczaj dla tej klasy decyzyjnej, która bardziej interesuje użytkownika. W literaturze klasę tę nazywa się *primary class* (drugą klasę decyzyjną nazywa się *secondary class*). W większości przypadków bardziej interesująca jest klasa reprezentowana przez mniejszą liczbę przykładów, stąd dla oznaczenia tej klasy stosowany jest również termin *minority class* (drugą klasę decyzyjną nazywa się wtedy *majority class*).

W dziedzinie informatyki zwanej wyszukiwaniem informacji (ang. *information retrieval*) do określenia wrażliwości metody wyszukiwania istotnych dokumentów stosowany jest termin *recall*, natomiast wskaźnik *PPV* funkcjonuje pod nazwą *precision*.

Zwiększenie wrażliwości klasyfikatora najczęściej odbywa się kosztem jego specyficzności. Powstaje zatem pytanie: w jaki sposób porównywać jakość dwóch klasyfikatorów o różnej wrażliwości i specyficzności? Problem ten najczęściej rozpatrywany jest dla dwuklasowych zadań klasyfikacji, w których wrażliwość i znamienność obliczane są dla *primary class* (oznaczmy ją  $X_p$ ). Jeśli koszt błędnej klasyfikacji jest nieznany, poszukuje się klasyfikatorów maksymalizujących liczbę poprawnych przyporządkowań do klasy  $X_p$ , przy jednoczesnej minimalizacji liczby fałszywych alarmów, generowanych przez klasyfikator. W pracy [10] przedstawiono argumentację, że z punktu widzenia maksymalizacji prawdopodobieństwa przyporządkowania przykładu do klasy  $X_p$  i jednoczesnej minimalizacji liczby fałszywych alarmów najlepszy jest klasyfikator o maksymalnej wartości miary *SSS* (ang. *the sum of specificity and sensitivity*) (2.24):

$$SSS(f, T) = Sens(f, X_p, T) + Spec(f, X_p, T) - 1. \quad (2.24)$$

Miara *SSS* nazywana jest również statystyką Youdena, miarą *gain* [124] lub *Class-gain* [258], dla podkreślenia jej roli w porównywaniu klasyfikatorów binarnych. Miara *SSS* ocenia zdolność klasyfikatora do unikania niepowodzeń, a więc błędnych klasyfikacji. Inne miary stosowane do oceny klasyfikatorów binarnych to  $G\text{-}mean} = \sqrt{Sens \cdot Spec}$  oraz  $F\text{-}measure} = (2 \cdot PPV \cdot Sens) / (PPV + Sens)$  [333]. Miary *SSS*, *G-mean* i *F-measure* nie są równoważne z punktu widzenia ustanawianego przez nie porządku klasyfikatorów.

Zauważmy, że na podstawie (2.18–2.21) wzór (2.24) można zapisać jako (2.25):

$$SSS(f, T) = TPR(f, X_p, T) + TPR(f, X_s, T) - 1. \quad (2.25)$$

We wzorze tym przez  $X_s$  oznaczono drugą z klas decyzyjnych (*secondary class*). Maksymalizacja miary SSS jest zatem równoważna maksymalizacji dokładności klas decyzyjnych, a tym samym maksymalizacji średniej arytmetycznej, obliczonej na podstawie dokładności klas decyzyjnych (2.26):

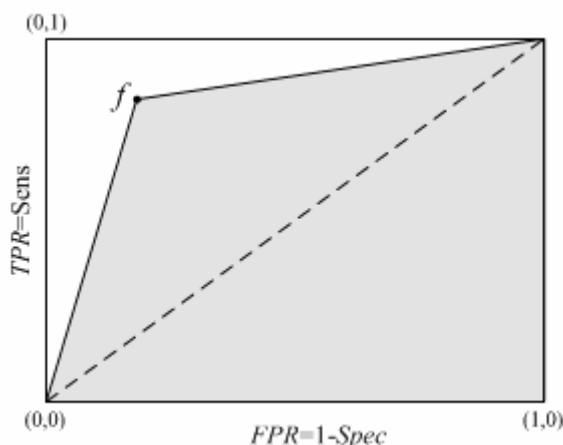
$$\text{AvgAcc}(f, T) = \frac{1}{|Vd|} \sum_{v \in Vd} TPR(f, X_v, T). \quad (2.26)$$

Dla dwuklasowego problemu klasyfikacji zachodzi następująca zależność pomiędzy SSS a  $\text{AvgAcc}$  (2.27):

$$SSS(f, T) = 2 \cdot \text{AvgAcc}(f, T) - 1. \quad (2.27)$$

W literaturze anglojęzycznej miara  $\text{AvgAcc}$  znana jest jako *balanced accuracy*. W niniejszej monografii miarę  $\text{AvgAcc}$  będziemy nazywać *średnią dokładnością klas decyzyjnych*.

Innym bardzo często stosowanym wskaźnikiem do oceny klasyfikatorów binarnych jest obszar pod tzw. krzywą ROC (ang. *receiver operating characteristic*). Krzywą ROC stosuje się do wizualnej oceny klasyfikatora. Jest ona rysowana w układzie kartezjańskim wartości  $FPR$  i  $TPR$  [75] (rys. 2.1). Z każdym klasyfikatorem dyskretnym (tzn. takim, który w sposób jednoznaczny przyporządkowuje klasyfikowany przykład do jednej z klas decyzyjnych) utożsamiony jest jeden punkt  $f(s_{FPR}, s_{TPR})$  w przestrzeni ROC. Oznacza to, że z jednym punktem w przestrzeni ROC związana jest cała rodzina klasyfikatorów charakteryzujących się daną wrażliwością i specyficznością. Najlepsze, charakteryzujące się stu procentową poprawnością klasyfikacji, są klasyfikatory odpowiadające punktowi  $(0,1)$ . Okolice głównej przekątnej, oznaczonej na rysunku 2.1 linią przerywaną, utożsamiane są z klasyfikatorami, których poprawność klasyfikacji zbliżona jest do klasyfikatorów działających losowo.



Rys. 2.1. Krzywa ROC klasyfikatora dyskretnego  
Fig. 2.1. The ROC curve for a discrete classifier

Aby oszacować jakość klasyfikatora charakteryzowanego przez punkt  $f(s_{FPR}, s_{TPR})$ , oblicza się pole powierzchni obszaru pod krzywą ROC (ang. *Area Under Curve – AUC*). Na rysunku 2.1 obszar ten zaznaczono kolorem szarym. Im większa jest wartość *AUC*, tym klasyfikator oceniany jest wyżej. Dla klasyfikatorów dyskretnych krzywa ROC składa się z dwóch odcinków. Pierwszy z nich definiowany jest przez punkty o współrzędnych  $(0,0)$  i  $(s_{FPR}, s_{TPR})$ , a drugi – przez punkty o współrzędnych  $(s_{FPR}, s_{TPR})$  i  $(1,1)$ . Innymi słowy, dla binarnego klasyfikatora dyskretnego maksymalizacja *AUC* równoznaczna jest z maksymalizacją *Sens + Spec*, czyli po prostu z maksymalizacją wartości miary *SSS* lub *AvgAcc*. Dowód tej zależności polega na wykazaniu, że dla dwóch klasyfikatorów dyskretnych,  $f_1$  i  $f_2$ ,  $(Sens_{f_1} + Spec_{f_1} > (<)Sens_{f_2} + Spec_{f_2}) \Leftrightarrow (AUC_{f_1} > (<)AUC_{f_2})$ , gdzie  $Sens_f$ ,  $Spec_f$ ,  $AUC_f$  są odpowiednio wrażliwością, specyficznością i *AUC* klasyfikatora  $f$ . W dowodzie wykorzystuje się fakt, że *AUC* to suma dwóch trójkątów oraz prostokąta i jest on na tyle prosty, że zostanie pominięty.

W przypadku klasyfikatorów probabilistycznych (a więc takich, które jako wynik klasyfikacji podają prawdopodobieństwo przynależności przykładu do każdej z klas decyzyjnych) wyznaczanie krzywej ROC odbywa się nieco inaczej. Każdy punkt takiej krzywej reprezentuje klasyfikator dyskretny, otrzymany na podstawie klasyfikatora probabilistycznego poprzez zadanie progu decyzji  $th$ . Każdy przykład o prawdopodobieństwie przynależności do klasy  $X_p$  większym lub równym  $th$  zostaje przyporządkowany do tej klasy. Zmieniając wartość  $th$ , otrzymujemy wykres krzywej ROC. Ponieważ w niniejszej pracy stosowane są jedynie klasyfikatory dyskretne, szczegółowe omówienie tematyki analizy krzywej ROC zostanie pominięte. Informacje na ten temat można znaleźć m.in. w [75].

W zadaniach klasyfikacji obejmujących więcej niż dwie klasy decyzyjne do oceny klasyfikatorów stosowana jest średnia dokładność klas decyzyjnych lub *AUC* definiowana dla problemów wieloklasowych [280]. Motywacją zastosowania średniej dokładności klas decyzyjnych jest taka sama jak w przypadku dwuklasowym. Maksymalizując średnią dokładność klas decyzyjnych, maksymalizujemy liczbę poprawnych klasyfikacji, przy czym wkład dokładności każdej z klas decyzyjnych do sumarycznej oceny klasyfikatora jest jednakowy. W szczególności ocena klasyfikatora preferującego klasy większościowe zostanie obniżona, jeśli równocześnie będzie on źle klasyfikował przykłady z klas mniej licznych. Standardowo zastosowanie *AUC* do oceny klasyfikatorów wieloklasowych polega na wyznaczeniu wartości  $AUC(X_v)$  dla każdego z problemów dwuklasowych (klasa  $X_v$  vs. pozostałe klasy) i „zsumowaniu” otrzymanych wyników. Stosowana jest suma ważona, w której wartość  $AUC(X_v)$  mnożona jest przez prawdopodobieństwo zaklasyfikowania

losowo wybranego przykładu do klasy  $X_v$ . Pewien problem może stanowić fakt, iż zdefiniowano kilka metod wyznaczania  $AUC$  dla problemów wieloklasowych. W pracy [130] przedstawiono przegląd i krytyczną analizę oceny klasyfikatorów za pomocą  $AUC$ . Średnia dokładność klas decyzyjnych (2.26) rozważana jest jako uogólnienie  $AUC$  wyznaczanej dla problemów wieloklasowych. Ze względu na czytelną interpretację średniej dokładności klas decyzyjnych w dalszej części monografii będzie ona, wraz z całkowitą dokładnością klasyfikacji, stosowana do oceny klasyfikatorów regułowych.

Ocena jakości zbiorów reguł regresyjnych polega na obliczeniu miar pozwalających zmierzyć wielkość błędu pomiędzy rzeczywistymi a przewidywanymi wartościami zmiennej decyzyjnej. W tym celu stosowane są standardowe miary, używane w analizie regresji i prognozowaniu szeregów czasowych. W literaturze przedmiotu, traktującej o zastosowaniach algorytmów maszynowego uczenia się, do rozwiązywania problemów o charakterze regresyjnym [144, 237, 330], do oceny modelu najczęściej stosuje się: miarę będącą pierwiastkiem błędu średniokwadratowego (2.28) (*RMSE* – ang. *root mean squared error*), średni błąd bezwzględny (2.29) lub względne postaci tych miar (2.30, 2.31). We wzorach (2.30, 2.31) przez  $\bar{d}$  oznaczono średnią wartość atrybutu decyzyjnego w zbiorze treningowym  $U$ .

$$RMSE(f, T) = \sqrt{\frac{1}{|U|} \sum_{x \in U} (f(x) - d(x))^2}, \quad (2.28)$$

$$MAE(f, T) = \frac{1}{|U|} \sum_{x \in U} |f(x) - d(x)|, \quad (2.29)$$

$$rRMSE(f, T) = \sqrt{\frac{\sum_{x \in U} (f(x) - d(x))^2}{\sum_{x \in U} (d(x) - \bar{d})^2}}, \quad (2.30)$$

$$rMAE(f, T) = \frac{\sum_{x \in U} |f(x) - d(x)|}{\sum_{x \in U} |d(x) - \bar{d}|}. \quad (2.31)$$

### 2.6.2. Metody oceny eksperymentalnej

Ocena klasyfikatora na podstawie wyników klasyfikacji przykładów znajdujących się w tablicy treningowej nie odzwierciedla jego rzeczywistych zdolności klasyfikacyjnych, gdyż model danych, będący składnikiem klasyfikatora, utworzono na podstawie analizy przykładów treningowych. Uzyskana w ten sposób ocena będzie zawyżona w stosunku do jej wartości rzeczywistych. Przykładowo, jeśli tablica treningowa jest niesprzeczna, to każdy

algorytm dokonujący indukcji reguł dokładnych otrzyma maksymalne oceny ze względu na każdą z wymienionych w poprzednim rozdziale miar oceny klasyfikatora.

Bardziej wiarygodną ocenę otrzymuje się, weryfikując działanie klasyfikatora na zbiorze przykładów niezależnym od zbioru treningowego. Zbiór taki nazywany jest zbiorem testowym. Ponieważ w praktyce dysponujemy jednym zbiorem danych, przed przystąpieniem do budowy klasyfikatora dostępny zbiór przykładów dzielony jest na części treningową i testową. Na podstawie części treningowej tworzony jest klasyfikator, natomiast jego ocena dokonywana jest na podstawie wyników klasyfikacji zbioru testowego. Najczęściej zbiory treningowy i testowy zawierają odpowiednio 2/3 i 1/3 przykładów należących do wejściowego zbioru. Taka technika eksperymentalnej oceny klasyfikatora znana jest jako *hold-out* i stosuje się ją głównie do zbiorów zawierających co najmniej tysiąc przykładów. Ocena uzyskana za pomocą techniki *hold-out* uzależniona jest od sposobu wylosowania przykładów testowych. Dlatego też jej wyniki mogą być mylące, szczególnie jeśli analizowany zbiór danych charakteryzuje się nierównomiernym rozkładem liczby przykładów reprezentujących klasy decyzyjne. Z tego powodu zdecydowanie częściej do eksperymentalnej oceny klasyfikatorów stosowana jest jedna z odmian metody, znana jako *k-krotna walidacja krzyżowa* (ang. *k-fold cross validation*). Metoda ta polega na podziale dostępnego zbioru przykładów  $U$  na  $k$  – parami rozłącznych i w miarę możliwości równolicznych – podzbiorów danych  $U_1, U_2, \dots, U_k$  i wyznaczeniu, za pomocą ustalonego algorytmu,  $k$  klasyfikatorów. Każdy  $i$ -ty klasyfikator definiowany jest na podstawie zbioru treningowego, złożonego z przykładów należących do zbioru  $U - U_i$ , a oceniany – na podzbiorze przykładów  $U_i$ . Wynikiem  $k$ -krotnej walidacji krzyżowej jest średnia arytmetyczna z wartości wszystkich otrzymanych w ten sposób ocen. Dzięki zastosowaniu walidacji krzyżowej można również obliczyć wariancję ocen, co ma szczególne znaczenie w ocenie stabilności działania klasyfikatora i podczas porównywania różnych klasyfikatorów. W ostatnim czasie, dla zwiększenia wiarygodności i zmniejszenia wariancji wyników,  $k$ -krotną walidację krzyżową powtarza się kilka, np.  $n$ , razy (dla różnych podziałów  $U_1, U_2, \dots, U_k$ ). Jako ostateczny wynik takiej procedury podaje się średnią wartość z  $n$  średnich wartości ocen otrzymanych w ramach każdej pojedynczej  $k$ -krotnej walidacji krzyżowej. Mechanizm tej walidacji jest zazwyczaj uzupełniany o warstwowe losowanie przykładów należących do klas decyzyjnych (ang. *stratified cross validation*). Warstwowa walidacja krzyżowa polega na takim podziale dostępnego zbioru przykładów, aby w podzbiorach  $U_i$ ,  $i \in \{1, 2, \dots, k\}$  proporcje pomiędzy liczbą przykładów reprezentujących kolejne klasy decyzyjne były takie same jak w  $U$ . W eksperymentalnej ocenie klasyfikatorów najczęściej stosowana jest 10-krotna warstwowa walidacja krzyżowa, a jeśli jest ona powtarzana, to liczba powtórzeń zazwyczaj wynosi od 3 do 10. W zasadzie nie ma żadnych

ograniczeń związanych z maksymalną liczebnością zbioru przykładów, do którego można zastosować  $k$ -krotną walidację krzyżową. Najczęściej ta metoda testowania używana jest w przypadku zbiorów danych, złożonych z więcej niż 100 i mniej niż 1000 przykładów. W chwili obecnej, w wielu pracach porównujących działanie klasyfikatorów,  $k$ -krotną walidację krzyżową stosuje się również do dużo większych zbiorów danych.

Inne znane odmiany walidacji krzyżowej to *leave-one-out* oraz *Monte Carlo cross validation*. Metoda *leave-one-out* to  $k$ -krotna walidacja krzyżowa, w której liczba  $k$  jest równa liczbie dostępnych przykładów. Oznacza to, że każdy klasyfikator testowany jest na jednoelementowym zbiorze testowym. Metoda *Monte Carlo cross validation*, znana również jako *random subsampling*, polega na wielokrotnym stosowaniu techniki *hold-out*. W każdym teście *hold-out* zbiory treningowy i testowy losowane są od nowa. Podczas stosowania tej techniki zaleca się co najmniej 30-krotne wykonanie procedury *hold-out*. Metoda *Monte Carlo cross validation* jest metodą oceniającą klasyfikator w sposób bardziej pesymistyczny niż  $k$ -krotna walidacja krzyżowa.

Dotychczas omówione techniki eksperymentalnej oceny klasyfikatorów zakładały, że każdy przykład występuje jednokrotnie albo w zbiorze treningowym, albo w zbiorze testowym. Do oceny klasyfikatorów tworzonych na podstawie małych zbiorów przykładów stosowana jest technika *bootstrap* [163]. Ze zbioru danych, złożonego z  $n$  przykładów, losowanych jest – z powtórzeniami –  $n$  przykładów. Stanowią one zbiór treningowy, a przykłady, których nie wylosowano, tworzą zbiór testowy. W ramach procedury *bootstrap* przeprowadza się  $k$  eksperymentów, a wynikowa ocena klasyfikatora najczęściej wyraża się wzorem  $\text{eval} = \frac{1}{k} \sum_{i=1}^k (0.632\text{eval}_{test-i} + 0.368\text{eval}_{train-i})$ , gdzie  $\text{eval}_{test-i}$ ,  $\text{eval}_{train-i}$  są odpowiednio wynikami  $i$ -tego eksperymentu po zastosowaniu  $i$ -tego klasyfikatora do  $i$ -tego zbioru testowego i zbioru treningowego. Procedura *bootstrap*, przeprowadzana według przedstawionego schematu, nazywana jest *0.632 bootstrap*.

Wymienione metody oceny eksperymentalnej opisane są w wielu pracach przeglądowych. Niewątpliwie dobrym przykładem jest publikacja Wittenego i Franka [333], opisująca m.in. metody oceny eksperymentalnej zawarte w oprogramowaniu Weka.

### 2.6.3. Metody porównywania zdolności klasyfikacyjnych

Techniki oceny eksperymentalnej umożliwiają wyznaczenie wiarygodnych wartości miar odzwierciedlających jakość klasyfikatorów. Umożliwia to porównanie różnych klasyfikatorów. W najprostszym podejściu porównanie polega na zastosowaniu jednej z technik eksperymentalnej oceny klasyfikatorów i porównanie otrzymanych wyników. Za lepszy uznawany jest ten klasyfikator, który charakteryzuje się lepszą oceną (np. wyższą dokładnością lub mniejszym błędem). Bardziej interesujące i wiarygodne jest porównanie

polegające na ocenie prawdopodobieństwa, że jeden z klasyfikatorów jest lepszy od drugiego (lub pozostałych, jeśli ocenianych jest jednocześnie kilka klasyfikatorów). Porównanie takie można przeprowadzić, stosując metody weryfikacji hipotez statystycznych.

Do chwili obecnej trwa dyskusja na temat sposobów porównywania klasyfikatorów. Jak dotąd nie przedstawiono metodyki, która znalazłaby uznanie wśród wszystkich badaczy. Najczęściej stosowana jest metoda przedstawiona w publikacji Demsara [63].

**Porównanie dwóch klasyfikatorów na podstawie jednego zbioru danych** polega na tym, że gdy ocena dwóch klasyfikatorów odbywa się na podstawie serii eksperymentów (np.  $k$ -krotna walidacja krzyżowa) przeprowadzonych na jednym zbiorze danych, do porównania stosowany jest test t lub jego skorygowana postać, znana jako *corrected resampled t-test* [333]. Jeśli porównanie przeprowadzane jest na tych samych zbiorach treningowych i testowych, to stosowany jest test dla zmiennych zależnych. Jeśli dysponujemy jedynie średnimi i wariancjami wyników, to stosowany jest test t dla zmiennych niepowiązanych. Stosowanie tego testu jest obciążone, ponieważ wariancja różnicy średnich może być mocno przeszacowana, konsekwencją czego są szerokie przedziały ufności i, ostatecznie, brak jest podstaw do odrzucenia hipotezy o nieistotności różnic pomiędzy klasyfikatorami.

Obecnie omówiony zostanie test dla zmiennych zależnych, gdyż taki test stosowany jest w części eksperymentów opisanych w dalszej części monografii. Założymy, że dane są dwa klasyfikatory:  $K_1$  i  $K_2$ . Zdefiniujemy hipotezę zerową, określającą, że na poziomie istotności  $p$  różnica pomiędzy średnimi ocenami klasyfikatorów  $K_1$  i  $K_2$  jest równa zero, co oznacza, że klasyfikatory charakteryzują się średnio tą samą oceną. Hipoteza alternatywna może określać, że różnica ta jest różna od zera (wówczas stosujemy test dwustronny) lub że jest większa od zera (wówczas stosujemy test jednostronny).

Przez  $e_1^{K_1}, e_2^{K_1}, \dots, e_k^{K_1}$  oznaczmy oceny empiryczne, uzyskane przez klasyfikator  $K_1$ , a przez  $e_1^{K_2}, e_2^{K_2}, \dots, e_k^{K_2}$  – oceny empiryczne, uzyskane przez klasyfikator  $K_2$ . Do oszacowania wartości średniej  $m_{K_1-K_2}$  i wariancji  $\delta_{K_1-K_2}^2$  rozkładu zmiennej losowej służą wzory  $m_{K_1-K_2} = \frac{1}{k} \sum_{i=1}^k (e_i^{K_1} - e_i^{K_2})$ ,  $\delta_{K_1-K_2}^2 = \frac{1}{k-1} \left( \sum_{i=1}^k (e_i^{K_1} - e_i^{K_2})^2 - \frac{1}{k} \left( \sum_{i=1}^k (e_i^{K_1} - e_i^{K_2}) \right)^2 \right)$ .

Na ich podstawie wyznaczamy wartość statystyki  $t = \frac{m_{K_1-K_2}}{\sqrt{\left(\frac{1}{k}\right) \delta_{K_1-K_2}^2}}$ . Wartość krytyczną

statystyki obliczamy dla rozkładu t Studenta, poziomu istotności  $p$  i  $k-1$  stopni swobody. Porównując wartość krytyczną z obliczoną wartością statystyki  $t$ , podejmujemy decyzję o odrzuceniu lub braku podstaw do odrzucenia hipotezy zerowej. Skorygowana postać testu t bierze pod uwagę fakt, iż kolejne zbiory treningowe i testowe nie są od siebie niezależne,

gdyż część przykładów należących do zbioru treningowego w  $i$ -tym eksperymentie będzie również do niego należeć w  $j$ -tym eksperymentie (podobnie będzie ze zbiorami testowymi).

Po modyfikacji statystyka  $t$  wyraża się wzorem  $t = \frac{m_{K1-K2}}{\sqrt{\left(\frac{1}{k} + \frac{n_2}{n_1}\right)\delta_{K1-K2}^2}}$ , gdzie  $n_1$  i  $n_2$  są

odpowiednio liczbami przykładów treningowych i testowych w każdym z  $k$  eksperymentów. Skorygowana postać testu t stosowana jest w przypadku wielokrotnie powtarzanych procedur  $k$ -krotnej walidacji krzyżowej lub *hold-out*. Dla 10-krotnej walidacji krzyżowej, wykonywanej jednokrotnie, stosowana jest podstawowa wersja testu.

**Porównanie dwóch klasyfikatorów na podstawie wielu zbiorów danych** odbywa się za pomocą nieparametrycznej alternatywy dla testu t, jaką jest test kolejności par Wilcoxona (ang. *Wilcoxon signed-rank test*). Przez  $E_1^{K1}, E_2^{K1}, \dots, E_N^{K1}$  oznaczmy oceny uzyskane przez klasyfikator  $K1$  na kolejnych zbiorach danych, a przez  $E_1^{K2}, E_2^{K2}, \dots, E_N^{K2}$  – oceny uzyskane przez klasyfikator  $K2$ . Przez  $D_i = E_i^{K1} - E_i^{K2}$  oznaczamy różnice pomiędzy ocenami uzyskanymi przez klasyfikatory  $K1$  i  $K2$ . Po uporządkowaniu różnic w ciąg rosnący przypisujemy im rangi. Następnie osobno sumujemy rangi dla różnic dodatnich i ujemnych, otrzymując  $R^+ = \sum_{D_i > 0} rank(D_i) + \frac{1}{2} \sum_{D_i=0} rank(D_i)$  oraz  $R^- = \sum_{D_i < 0} rank(D_i) + \frac{1}{2} \sum_{D_i=0} rank(D_i)$ , gdzie  $rank(D_i)$  oznacza pozycję (rangę), na jakiej w uporządkowanym ciągu znajduje się różnica  $D_i$ . Jeśli liczba zbiorów, dla których różnica  $D_i = 0$ , jest nieparzysta, to jeden z nich jest w obliczeniach pomijany. Do weryfikacji hipotezy zerowej, stanowiącej, że nie ma istotnych różnic w ocenie klasyfikatorów, stosowana jest statystyka  $T = \min\{R^+, R^-\}$ . Przy zadanym poziomie istotności wartość krytyczną statystyki można odczytać z tabel lub wyznaczyć ją za pomocą jednego z pakietów obliczeń statystycznych. Jeśli klasyfikatory porównywane są na  $N$  zbiorach danych oraz  $N > 25$ , można zastosować statystykę  $z = \frac{T - 0.25N(N+1)}{\sqrt{\frac{1}{24}N(N+1)(2N+1)}}$ ,

której rozkład jest w przybliżeniu normalny. Jeśli porównywane wartości ocen są mieralne, to test Wilcoxona jest mocniejszy od testu znaków (ang. *sign test*), dlatego też jest on preferowany podczas porównywania dwóch klasyfikatorów na wielu zbiorach danych [63]. Test znaków stosowany jest wtedy, gdy porównanie klasyfikatorów bazuje jedynie na informacji o liczbie zwycięstw, porażek i remisów pomiędzy klasyfikatorami. Zestawienie informujące o liczbie zwycięstw/remisów/porażek znane jest pod angielską nazwą *win/tie/loss*. W większości publikacji w zestawieniu *win/tie/loss* zawiera się informację o statystycznie istotnych różnicach pomiędzy klasyfikatorami (stosowany jest test t).

**Porównanie wielu klasyfikatorów na postawie wielu zbiorów danych** najczęściej realizowane jest za pomocą testu Friedmana, będącego nieparametrycznym odpowiednikiem jednoczynnikowej analizy wariancji dla pomiarów powtarzanych. W teście tym zakłada się, że oceny uzyskane przez każdy z klasyfikatorów są podstawą do przyznania im odpowiednich rang. Najlepszy klasyfikator na danym zbiorze otrzymuje rangę 1, drugi klasyfikator otrzymuje rangę 2 itd. W przypadku tych samych wartości oceny (remisów) klasyfikatorom przyporządkowane są średnie wartości rang. Niech  $r_i^j$  będzie rangą  $j$ -tego klasyfikatora, uzyskaną przez niego na  $i$ -tym zbiorze danych. Test Friedmana porównuje średnie wartości rang  $R_j = \frac{1}{N} \sum_i r_i^j$ , jakie otrzymały kolejne klasyfikatory. Hipoteza zerowa określa, że nie ma statystycznych różnic pomiędzy klasyfikatorami, czyli że ich średnie rangi są identyczne. Do weryfikacji hipotezy zerowej stosowana jest statystyka

$$\chi_F^2 = \frac{12}{N(N+1)} \left[ \sum_j R_j^2 - \frac{l(l+1)^2}{4} \right],$$

która w przybliżeniu ma rozkład  $\chi^2$ , z  $l-1$  stopniami

swobody, lub jej mniej konserwatywna postać  $F_F = \frac{(N-1)\chi_F^2}{N(l-1)\chi_F^2}$ , która charakteryzuje się rozkładem  $F$ , z  $(l-1)$  i  $(l-1)(N-1)$  stopniami swobody. W obu wzorach  $N$  oznacza liczbę zbiorów danych, na których porównywane są klasyfikatory, a  $l$  oznacza liczbę porównywanych klasyfikatorów. Wartość krytyczną obu statystyk można odczytać z tablic rozkładu lub obliczyć ją w dowolnym pakiecie obliczeń statystycznych. W pracy [63] przedstawiono analizę zależności pomiędzy testem Friedmana a jednoczynnikową analizą wariancji. Przedstawiono tam również argumentację przemawiającą za stosowaniem testu Friedmana do porównywania wielu klasyfikatorów na wielu zbiorach danych.

Jeśli hipoteza zerowa o braku istotnych różnic pomiędzy klasyfikatorami zostanie odrzucona, kolejnym krokiem analizy porównawczej jest wykonanie testu Nemenyi. Zadaniem testu jest sprawdzenie, które z wartości ocen porównywanych klasyfikatorów są statystycznie różne. Test Nemeniego należy do grupy testów *po fakcie* (ang. *post-hoc*), znanych również, jako testy wielokrotnych porównań. Test wykonuje się po odrzuceniu hipotezy o identycznych ocenach klasyfikatorów. Ocena klasyfikatora o numerze  $i1$  jest statystycznie różna od oceny klasyfikatora o numerze  $i2$ , jeśli różnica pomiędzy średnimi wartościami rankingów  $R_{i1}$  i  $R_{i2}$  jest większa od wartości krytycznej

$CD = q_p \sqrt{\frac{l(l+1)}{6N}}$ . Wartość  $q_p$  zależy od poziomu istotności  $p$  oraz liczby porównywanych klasyfikatorów [63]. Do graficznej reprezentacji wyników testu Nemenyi stosowane są tzw. *CD diagramy* (ang. *critical difference diagrams*). Za pomocą *CD diagramów* można graficznie zaprezentować, które spośród porównywanych klasyfikatorów różnią się pomiędzy

sobą. W pracy [95] przeprowadzono krytykę testu Nemeniego, pokazując, że jest on bardzo konserwatywny i rzadko wykazuje statystyczną przewagę jednego klasyfikatora nad pozostałymi. Dlatego też coraz częściej wynik porównania klasyfikatorów, uzyskany za pomocą testów Friedmana i Nemeniego, stanowi jedynie uzupełnienie do porównań wykonywanych parami za pomocą testu kolejności par Wilcoxona.

## 2.7. Rozwój metod indukcji reguł decyzyjnych

Przed omówieniem najważniejszych problemów związanych z indukcją reguł decyzyjnych trzeba jasno podkreślić, że liczba prac prowadzonych w tej dziedzinie jest tak duża, że niemożliwe jest przedstawienie i omówienie wszystkich kierunków badań. Poniżej zostaną omówione najważniejsze z nich (zdaniem autora). Interesujący przegląd tematyki traktującej o indukcji reguł decyzyjnych znajdzie się zapewne w książce Fürnkranza, Lavrac i Gambergera pod tytułem *Rule learning: Essentials of Machine Learning and Relational Learning*, której wydanie planowane pod koniec 2012 roku.

### 2.7.1. Poprawa zdolności klasyfikacyjnych i opisowych

Głównym celem indukcji reguł dla celów klasyfikacyjnych jest utworzenie klasyfikatora o jak najwyższej jakości. Prace mające na celu podniesienie jakości klasyfikatorów budowanych za pomocą algorytmów pokryciowych obejmują między innymi:

- poszukiwanie lepszych kryteriów i strategii stosowanych do oceny reguł w fazach wzrostu i przycinania [42, 143, 253, 261, 267]; w szczególności do ustalenia zakresów prostych warunków elementarnych stosowana jest zasada minimalnej długości kodu [222, 297, 298] lub bieżąca charakterystyka krzywej ROC tworzonego klasyfikatora [76];
- zmianę reprezentacji lub redefinicję zbioru reguł; zmiana reprezentacji polega na zamianie prostych warunków elementarnych w warunki złożone (np. zanegowane, skośne lub zawierające krzywe wyższych stopni [64, 313, 315, 224, 235, 265, 266]), a także na rozmywaniu reguł [135, 249];
- modyfikację schematów stosowanych do rozstrzygania konfliktów klasyfikacji, w szczególności wykorzystanie złożonych systemów głosujących oraz stosowanie technik wygładzania (np. *smoothing* lub *shrinking*) [42, 121, 122, 124, 131, 296, 284]; do rozstrzygania konfliktów klasyfikacji nadaje się również metoda, w której decyzja podejmowana przez klasyfikator wypracowywana jest przez tzw. lokalnych agentów, którymi w szczególnym przypadku mogą być zbiory reguł [324, 325]; dobry eksperymentalny przegląd wpływu różnych metod rozstrzygania konfliktów na wyniki klasyfikacji przedstawiono w pracach [305, 306], a ogólny wniosek wypływający z tych badań jest taki, że do nadzorowania procesu indukcji reguł oraz rozstrzygania konfliktów

klasyfikacji powinny być wykorzystane identyczne kryteria, gdyż zapewnia to otrzymanie klasyfikatorów o wysokiej dokładności;

- stosowanie podwójnej indukcji (ang. *double induction*) [184], czyli ponowne wyznaczanie reguł na podstawie przykładów treningowych, pokrywanych przez reguły klasyfikujące w sposób niejednoznaczny aktualnie klasyfikowany przykład testowy;
- stosowanie konstruktywnej indukcji, a więc rozszerzenie zbioru atrybutów warunkowych o nowe atrybuty, których wartości zależą funkcjonalnie [28] lub logicznie [145, 338] od kombinacji wartości istniejących atrybutów;
- wykorzystanie wiedzy dziedzinowej o analizowanym problemie w celu odpowiedniego ukierunkowania procesu indukcji; wiedza dziedzinowa może być reprezentowana przez zbiór argumentów (ang. *argumentation based learning*) wyjaśniających dlaczego dany przykład jest (bądź nie jest) reprezentantem danej klasy decyzyjnej; przykładami algorytmów realizujących indukcję reguł na podstawie argumentacji są modyfikacje algorytmów CN2 i MODLEM [200, 206].

Poprawę dokładności klasyfikatorów regułowych próbuje się uzyskać, stosując algorytmy indukcji inne niż pokryciowe. Naturalnym rozwiązaniem jest użycie do tego celu odpowiednio zaadaptowanych algorytmów indukcji reguł asocjacyjnych [148, 235]. Inny sposób polega na tworzeniu reguł na podstawie modelu danych, będącego rezultatem zastosowania innej metody maszynowego uczenia się. Najczęściej do tego celu używane są modele o reprezentacji niesymbolicznej, np. sieci neuronowe, algorytmy ewolucyjne lub maszyny wektorów podpierających (ang. *support vector machines – SVM*). W celu ułatwienia wyznaczenia reguł na podstawie sieci neuronowych do algorytmów trenowania sieci wprowadza się wiele modyfikacji, polegających m.in. na eliminowaniu powiązań pomiędzy neuronami o małych wartościach wag lub zamianie wartości wag na wartości logiczne [66, 67, 308, 310]. W metodach posługujących się modelami SVM zidentyfikowane wektory podpierające stanowią podstawę do wyznaczenia reguł, w tym reguł zawierających złożone (np. skośne lub elipsoidalne) warunki elementarne. Model SVM jest także podstawą do wygenerowania przykładów syntetycznych, które zasilając zbiór przykładów treningowych, pozwalają na wyznaczenie reguł o wyższych zdolnościach klasyfikacyjnych [64, 188]. W algorytmach ewolucyjnych zbiory reguł stanowią zakodowaną populację osobników. Ewolucja przebiega w kierunku maksymalizacji zdolności klasyfikacyjnych oraz minimalizacji złożoności osobnika [79, 249]. Zastosowanie metod ewolucyjnych pozwala na umieszczenie w przesłankach reguł dowolnie złożonych warunków elementarnych.

Kolejny etap w rozwoju klasyfikatorów regułowych to definiowanie systemów złożonych, składających się z dwóch lub większej liczby klasyfikatorów. Ze względu na stosunkowo dużą niestabilność pokryciowych algorytmów indukcji reguł klasyfikatory regułowe o wyższej dokładności klasyfikacji otrzymuje się po zastosowaniu metod znanych

jako *bagging*, *boosting* lub *stacking* [333]. Metody *bagging* i *boosting* uznaje się za homogeniczne, gdyż używają one wielu klasyfikatorów tego samego typu (w naszym przypadku klasyfikatorów regułowych). *Stacking* jest metodą heterogeniczną, klasyfikator składa się z wielu modeli danych otrzymanych na podstawie różnych algorytmów.

Zastosowanie w indukcji reguł techniki *bagging* polega na wielokrotnym losowaniu z powtórzeniami zbiorów treningowych o takiej samej liczbie przykładów jak w zbiorze wejściowym. Dla każdego z otrzymanych w ten sposób zbiorów dokonuje się indukcji reguł [289]. Ideą techniki *boosting* jest budowa klasyfikatora złożonego z wielu prostych klasyfikatorów. W algorytmach indukcji reguł stosujących technikę *boosting* każda reguła traktowana jest jako prosty klasyfikator. Podczas indukcji kolejne reguły tworzone są w taki sposób, aby minimalizować wartość pewnej ustalonej funkcji straty. Przez wykorzystanie różniczkowalnych funkcji straty do minimalizacji wartości tych funkcji możliwe jest użycie metod gradientowych. Przykładami algorytmów działających według przedstawionej tutaj w sposób bardzo ogólny zasady są m.in. ENDER [62] oraz RuleFit [84]. Ich cechą charakterystyczną jest generowanie dużej liczby reguł, co znaczowo utrudnia lub wręcz uniemożliwia ich interpretację. W pracy opisującej działanie algorytmu ENDER autorzy przedstawiają kilka propozycji zmniejszenia liczby generowanych reguł. Proponują oni, aby liczba reguł ustalana była arbitralnie lub klasyfikator składał się jedynie z najlepszych spośród wyznaczonych reguł. Efektywność tych propozycji nie jest jednak w sposób wystarczający zweryfikowana eksperymentalnie.

Do metod homogenicznych można także zaliczyć klasyfikatory wielokrotne, w których problem  $n$ -klasowy dekomponowany jest na kilka (zazwyczaj binarnych) problemów klasyfikacji [289]. Przykładem metody homogenicznej jest również stosowanie klasyfikatora hierarchicznego, w którym na wyższych poziomach umieszczone są reguły ogólne, a na niższych – reguły dokładne [272].

Do metod heterogenicznych, łączących klasyfikatory regułowe i inne metody maszynowego uczenia się, można zaliczyć m.in. algorytm RIONA [101], stosujący paradymat leniwego uczenia się (ang. *lazy learning*), w którym reguły wyznaczane są jedynie na podstawie przykładów treningowych, podobnych do klasyfikowanego przykładu testowego. Przykłady podobne identyfikowane są za pomocą metody  $k$ -najbliższych sąsiadów. Heterogeniczny jest również klasyfikator proponowany przez Stefanowskiego i Wilka [286], łączący reprezentację regułową z techniką wnioskowania na podstawie przypadku (ang. *case based reasoning*). W propozycji tej reguły wyznaczane są jedynie na podstawie przykładów typowych. Przykłady testowe, niepokryte przez żadną z wyznaczonych reguł, klasyfikowane są przez najbliższy przykład znajdujący się w zbiorze tzw. wyjątków. Podobną ideę realizuje algorytm BRACID [207], który przeznaczony jest do indukcji reguł w zbiorach o nierównomiernym rozkładzie liczby przykładów

reprezentujących klasy decyzyjne. BRACID nie jest algorytmem pokryciowym i łączy indukcję reguł, wnioskowanie na podstawie przypadku i metodę najbliższych sąsiadów. Reguły tworzone są, zgodnie z metodyką, od szczegółu (pojedynczego przykładu) do ogólnego (reguły). Efektem działania algorytmu jest zbiór reguł i przykładów, których uogólnić się nie dało.

Autor w pracy [263] przedstawił system heterogeniczny, przeznaczony do rozwiązywania problemów regresyjnych. W systemie tym połączono metodę najbliższych sąsiadów z indukcją reguł regresyjnych oraz metodyką ARIMA [35], stosowaną wtedy, gdy analizowane dane mają charakter szeregu czasowego.

Rozwój metod indukcji reguł decyzyjnych, definiowanych dla celów opisowych, ukierunkowany jest na stosowanie zmodyfikowanych algorytmów pokryciowych lub algorytmów indukcji reguł asocjacyjnych. Zaproponowano algorytmy ukierunkowane na poszukiwanie ograniczonego zbioru silnych (ogólnych i dokładnych) reguł, pokrywających rozłączne obszary w przestrzeni atrybutów. Przykładami takich algorytmów są Apriori-SD [157] i CN2-SD [177] (SD – *Subgroup Discovery* [94, 213]), a także propozycja Webba i Zhangi [328], pozwalająca na wyznaczenie tzw. k-optimálnych reguł, spełniających zdefiniowane przez użytkownika wymagania minimalnej jakości. W *subgroup discovery* wymagania dotyczące zdolności klasyfikacyjnych zbioru reguł nie są tak istotne jak liczba, jakość oraz istotność statystyczna reguł. W szczególności w ocenie statystycznej reguł coraz częściej podnoszony jest problem redukcji fałszywych odkryć [329]. W algorytmach Apriori-SD i CN2-SD zastosowano mechanizm przyporządkowania przykładom wag; waga każdego przykładu maleje wraz ze wzrostem liczby pokrywających go reguł. Takie postępowanie, bardziej niż w standardowym algorytmie pokryciowym, ukierunkowuje poszukiwanie reguł na obszary niepokryte przez żadną z dotychczas wyznaczonych reguł. Zbiory reguł, przeznaczone do opisu danych, uzyskuje się również za pomocą metod oferowanych przez teorię zbiorów przybliżonych. U podstaw tej teorii leżała chęć opisu danych. Zastosowanie reguł minimalnych do rozwiązywania problemów klasyfikacyjnych było niejako naturalną konsekwencją teorii i poszukiwaniem dla niej nowych zastosowań. Do opisu danych można użyć zarówno standardowego [220], jak i tolerancyjnego [271, 292, 295] modelu zbiorów przybliżonych. Do tego celu dobrze będą się nadawać również inne uogólnienia teorii, pozwalające na rozluźnienie kryteriów związanych z przynależnością przykładu do przybliżeń opisywanego pojęcia. Do uogólnień tych można zaliczyć m.in. model zmiennej precyzji [346] oraz modele parametryzowane [107, 146, 304].

W celu ułatwienia interpretacji regułowego modelu danych stosowane są różnego rodzaju techniki wizualizacji [77, 92, 213]. Podstawą wizualizacji są diagramy, grafy i histogramy. Najczęściej węzły reprezentują warunki elementarne, a krawędzie grafu informują

o powiązaniu pomiędzy warunkami a decyzjami. Techniki wizualizacji są użyteczne, jeśli zbioru reguł i warunków elementarnych nie są zbyt liczne.

Niezależnie od sposobu wyznaczenia reguł, w celu poprawy ich zdolności opisowych stosowane są różnego rodzaju techniki optymalizacji. Optymalizacja może dotyczyć zarówno pojedynczych reguł, jak i całego ich zbioru. O wybranych technikach optymalizacji będzie mowa w rozdziale 5, tam też przedstawiony zostanie krótki przegląd prac pokrewnych.

Dla wyznaczenia reguł interesujących i użytecznych dla użytkownika niebagatelne znaczenie mają modyfikacje algorytmów indukcji, które pod uwagę biorą preferencje użytkownika, dotyczące budowy reguł, bądź też wykorzystującą wiedzę dziedzinową o analizowanym problemie. O pracach związanych m.in. z użyciem wiedzy dziedzinowej w indukcji reguł powiemy więcej w rozdziale 6, podczas omawiania zastosowań algorytmów indukcji reguł.

Pomimo tak dużej liczby algorytmów i technik mających na celu poprawę zdolności klasyfikacyjnych i opisowych reguł decyzyjnych, nie można wskazać jednej techniki indukcji, która byłaby najskuteczniejsza niezależnie od charakterystyki analizowanego zbioru danych. Teza ta ma swoje uzasadnienie teoretyczne, które manifestuje się w tzw. *no free lunch theorem*, mówiącym o średniej równoważności klasyfikatorów ze względu na ich zdolności klasyfikacyjne [337]. Twierdzenie to nie oznacza oczywiście, że wszystkie prace opisujące coraz doskonalsze klasyfikatory są bezużyteczne. Z punktu widzenia rozwoju klasyfikatorów interesujące są bowiem wszelkie udoskonalenia prowadzące do ich lepszego działania na zbiorach danych, opisujących rzeczywiste i użyteczne problemy klasyfikacyjne.

Prace prowadzone przez różnych badaczy koncentrują się zatem na definiowaniu systemów: działających średnio lepiej w pewnej grupie zbiorów danych, przeznaczonych do konkretnych zastosowań lub dostosowanych do danych o określonej charakterystyce.

Swego rodzaju remedium na ograniczenia wynikające z twierdzenia *no free lunch* może być stosowanie metauczenia (ang. *meta-learning*). Systemy uczące się na poziomie meta dostosowują swoje parametry do charakterystyki analizowanego zbioru danych. Ze zbioru treningowego wydzielana jest tzw. część walidacyjna, na której testowana jest efektywność różnych klasyfikatorów utworzonych na podstawie analizy pozostałych przykładów treningowych. Najlepszy z wyznaczonych klasyfikatorów stanowi model wyjściowy. Uczenie się na poziomie meta obejmuje zatem wybór metody budowy klasyfikatora i dostrojenie jego parametrów. Strategie postępowania, pozwalające na optymalizację tego procesu, opisano m.in. w [68, 142]. Autor prowadził – samodzielnie oraz wspólnie z Wróblem – prace polegające na użyciu metauczenia w celu identyfikacji miary jakości, jaką należy zastosować w algorytmie indukcji reguł [253, 261]. Wybrane wyniki tych prac zaprezentowane zostaną w rozdziale 4.

### 2.7.2. Problemy związane z rozmiarem i niedoskonałością danych

Na zakończenie tego rozdziału zostaną krótko omówione najważniejsze problemy, z którymi muszą borykać się wszystkie algorytmy indukcji reguł. Do problemów tych zaliczamy:

- dużą liczbę przykładów treningowych,
- dużą liczbę atrybutów warunkowych,
- nierównomierny rozkład liczby przykładów reprezentujących klasy decyzyjne,
- nieznane/brakujące wartości atrybutów warunkowych.

Podstawowe techniki umożliwiające indukcję reguł w dużych zbiorach danych obejmują redukcję lub dekompozycję zbioru przykładów, a także inkrementacyjne wyznaczanie reguł. Techniki redukcji polegają głównie na użyciu metody okien (ang. *windowing*) [238]. Rzadziej od metody okien stosowane są metody wyboru przykładów reprezentatywnych dla zbioru treningowego [29, 118, 292]. Redukcja liczby przykładów pozwala obniżyć liczbę generowanych reguł, ale w niektórych sytuacjach może niekorzystanie wpływać na wrażliwość i specyficzność klasyfikatora ze względu na mniej liczne klasy decyzyjne [118].

Techniki dekompozycji polegają na: podziale zbioru przykładów na podzbiory (tzw. wzorce), wyznaczeniu reguł w każdym podzbiorze i traktowaniu otrzymanych w ten sposób klasyfikatorów jak systemu złożonego [174, 210]. Podczas dekompozycji zbiór przykładów dzielony jest na rodzinę parami rozłącznych podzbiorów, w rodzinie takiej każdy podzbiór opisany jest przez pewien wzorzec. W trakcie klasyfikacji przykład testowy najpierw przyporządkowany jest do konkretnego wzorca, a następnie klasyfikowany jest przez zbiór reguł, które wyznaczono na podstawie przykładów pokrywanych przez wzorzec.

Do indukcji reguł w dużych zbiorach danych stosuje się również algorytmy inkrementacyjne [52, 242, 316, 335]. W algorytmach tych każdy kolejny przykład konfrontowany jest ze zbiorem już wyznaczonych reguł. W zależności od tego, czy przykład jest sprzeczny czy niesprzeczny z istniejącym zbiorem reguł, uruchamiana jest procedura polegająca na aktualizacji pokrywających go reguł albo aktualizacji przyporządkowanych mu ocen.

W analizie dużych zbiorów danych można wykorzystywać techniki rozpraszania i/lub zrównoleglenia obliczeń. Pokryciowe algorytmy indukcji reguł dobrze nadają się do rozpraszania i zrównoleglania, które może polegać m.in. na równoczesnej indukcji reguł w klasach decyzyjnych, równoczesnym poszukiwaniu warunków elementarnych itd.

Problem indukcji reguł w tablicach decyzyjnych, zawierających dużą liczbę atrybutów warunkowych rozwiązywany jest za pomocą metod selekcji. Można je podzielić na metody: filtrujące, powłoki (ang. *wrapper approaches*) [164] oraz osadzone (ang. *embedded approaches*). Metody filtrujące dokonują selekcji cech najbardziej istotnych. Najczęściej stosowane miary istotności to różnego rodzaju współczynniki korelacji lub względna entropia

pomiędzy atrybutem(-ami) warunkowym(-ymi) a atrybutem decyzyjnym. W podejściu wstępującym początkowo pusty zbiór atrybutów jest sukcesywnie rozszerzany, w podejściu zstępującym – ze zbioru wszystkich atrybutów usuwane są atrybuty nieistotne.

Metody powłoki są metodami dwustopniowymi, w których algorytm selekcji połączono z algorymem indukcji. Do budowy klasyfikatora algorytm indukcji używa cech wybranych na etapie selekcji. Jakość wytrenowanego klasyfikatora odzwierciedla jakość wybranego zbioru atrybutów. Interesujące podejście do takiego sposobu selekcji atrybutów przedstawiono w [31], gdzie użyto zmodyfikowanej miary konfirmacji, standardowo stosowanej do oceny atrakcyjności reguł. Osadzone metody selekcji atrybutów są metodami będącymi integralną częścią algorytmów indukcji. Dobry przegląd i systematykę metod selekcji atrybutów zawarto w pracach [29, 164, 282, 333].

W przypadku nierównomiernego rozkładu liczby przykładów pomiędzy klasami decyzyjnymi, przed indukcją można stosować metody pozwalające na jego zrównoważenie. Najpopularniejsze z nich to: usuwanie przykładów z większościowej klasy decyzyjnej, zwiększanie liczby przykładów reprezentujących klasę mniejszościową oraz stosowanie obu tych metod jednocześnie. Wprowadzanie nowych przykładów może odbywać się przez: losową replikację istniejących przykładów, zmianę przyporządkowania przykładów z klasy większościowej do klasy mniejszościowej lub generowanie w sposób syntetyczny nowych przykładów, podobnych do przykładów reprezentujących klasę mniejszościową [48]. Podczas usuwania przykładów z większościowej klasy decyzyjnej stosowane są różnego rodzaju techniki, mające na celu identyfikację przykładów leżących „na granicy” pomiędzy klasami decyzyjnymi [175]. Przykłady takie są usuwane z klasy większościowej, gdyż przyczyniają się do indukcji reguł, które błędnie przyporządkowują przykłady z klasy mniejszościowej do klasy większościowej. Inne podejście przedstawiono w [288], gdzie informacje o górnym i dolnym przybliżeniu klas decyzyjnych wykorzystywane są do indukcji reguł z klasy większościowej i zmiany etykiet przykładów należących do klasy mniejszościowej. Poprawę jakości klasyfikatorów regułowych, budowanych na podstawie zbioru przykładów o nierównomiernym rozkładzie, uzyskiwano także przez: modyfikację schematu klasyfikacji, stosowanie różnych algorytmów indukcji w klasach mniejszościowej i większościowej [124], indukcję reguł ukierunkowaną na maksymalizację wartości *AUC*, *SSS* lub *F-measure* [207], a także użycie *boostingu* (np. [62]).

Ostatni z problemów, z jakimi borykają się algorytmy indukcyjnego uczenia się, to problem nieznanych (brakujących) wartości atrybutów warunkowych. Zauważmy, że wszystkie z omawianych do tej pory algorytmów mogą być, przynajmniej teoretycznie, zastosowane bez żadnych modyfikacji do analizy danych wysokowymiarowych i o nierównomiernym rozkładzie. Żaden z nich bez pewnych modyfikacji nie może być stosowany do analizy tablic zawierających nieznane wartości. Wymienia się trzy przyczyny

pojawiania się przykładów z nieznanymi wartościami atrybutów oraz wskazuje się na dwa podejścia do rozwiązania tego problemu. Przyczyną wystąpienia nieznanej wartości może być to, że: jest ona nieznana, gdyż np. nie została zmierzona (ang. *missing value*); nie sposób jej ustalić, gdyż nie ma odpowiedniej wartości dla danego przykładu (ang. *not applicable value*); nie ma ona znaczenia z punktu widzenia przyporządkowania przykładu do danej klasy decyzyjnej (ang. *don't care value*). Metody rozwiązania problemu nieznanych wartości dzielone są na: metody przetwarzania wstępnego, uruchamiane przed algorytmem indukcji, oraz wewnętrzne, działające w obrębie algorytmu indukcji.

Najprostszą i najmniej skuteczną metodą rozwiązania problemu nieznanych wartości jest usunięcie wszystkich zawierających je przykładów. Metodą przeciwną jest podstawienie w miejsce nieznanej wartości każdej z możliwych wartości atrybutu. Takie postępowanie powoduje zwiększenie liczby przykładów treningowych. Kolejne metody polegają na wstawianiu w miejscu nieznanej wartości takiej, która najczęściej występuje w zbiorze przykładów, lub wartości najczęściej występującej w obrębie danej klasy decyzyjnej (w przypadku atrybutów numerycznych jest to wartość średnia). Możliwym rozwiązaniem jest również traktowanie nieznanej wartości (odpowiadającego jej kodu w tablicy decyzyjnej) jako dodatkowej, pełnoprawnej wartości atrybutu, postępowanie takie nie ma jednak uzasadnienia dla nieznanych wartości typu *missing values*.

Efektywną, wewnętrzną strategią rozwiązywania problemu nieznanych wartości jest ich ignorowanie. Dla algorytmów indukcji reguł oznacza to, że przykład z nieznaną wartością atrybutu  $a$  nie pokrywa żadnych reguł z warunkami elementarnymi, zbudowanymi na jego podstawie (zarówno w fazie budowy reguły, jak i klasyfikacji). Podczas testu sprawdzającego, czy na podstawie atrybutu  $a$  można zbudować dobry warunek elementarny, wszystkie przykłady z nieznanymi wartościami atrybutu  $a$  nie są brane pod uwagę. Modyfikacją tego podejścia jest tzw. strategia pesymistyczna (ang. *pessimistic value strategy*), w której zakłada się, że wszystkie przykłady pozytywne, zawierające nieznane wartości aktualnie rozważanego atrybutu, nie spełniają testu, natomiast wszystkie negatywne przykłady z takimi wartościami zawsze test spełniają [334].

Bardziej zaawansowane obliczeniowo strategie próbują metodami maszynowego uczenia przewidzieć, jaka powinna być nieznana wartość atrybutu, lub wykorzystują informację o wartościach przykładów podobnych do tego, który zawiera nieznaną wartość [334]. To drugie podejście było przedmiotem intensywnych badań w teorii zbiorów przybliżonych. W pracach [125, 170, 284, 287] rozważano różne modyfikacje standardowej relacji nieroróżnialności, w których pod uwagę brana jest semantyka brakujących wartości. W szczególności w pracy [125] przedstawiono uogólnienie relacji nieroróżnialności oraz modyfikację algorytmu LEM2, umożliwiającą indukcję reguł z niekompletnych tablic decyzyjnych. W pracy [170] Kryszkiewicz przedstawiła definicję relacji tolerancji dla

niekompletnych tablic decyzyjnych. Umożliwiło to definiowanie przybliżeń zbiorów w niekompletnych tablicach oraz zdefiniowanie tzw. uogólnionych reguł decyzyjnych, wskazujących na przynależność przykładu do jednej z kilku klas decyzyjnych. Stefanowski [284,286] proponuje dwa uogólnienia zbiorów przybliżonych, stosujące tzw. wartościowaną relację tolerancji (dla *missing values*) oraz niesymetryczną relację podobieństwa (dla *not applicable values*).

Porównanie efektywności różnych strategii rozwiązywania problemu brakujących wartości można znaleźć m.in. w przeglądzie prezentowanym przez Wohlraba i Fürnkranza [334].

W dalszej części monografii problemy związane z brakującymi wartościami atrybutów rozwiązywane będą za pomocą strategii polegającej na ich ignorowaniu. Ta niewymagająca dodatkowych nakładów obliczeniowych strategia pozwala na indukcję reguł o jakości porównywalnej z metodami bardziej zaawansowanymi [334].

### **3. OBIEKTYWNE MIARY OCENY REGUŁ I ICH ZBIORÓW**

Ocena reguły polega na przyporządkowaniu jej wartości liczbowej. W zależności od zastosowanej miary wartość ta może być interpretowana na różne sposoby. W niniejszym rozdziale omówione zostaną tzw. obiektywne miary jakości, dokonujące oceny reguł na podstawie informacji o liczbie pokrywanych przez nie przykładów pozytywnych i negatywnych. Zostaną także opisane miary biorące pod uwagę budowę reguł, ich wzajemne podobieństwo oraz równomierność rozłożenia przykładów w pokrywanym przez nie „obszarze”. W rozdziale przedyskutowano również możliwość wielokryterialnej oceny reguł. Poza miarami oceniającymi reguły omówiono również wybrane miary oceniające jakość zbioru reguł.

W rozdziale 2 starano się zaakcentować, że oceny reguł dokonuje się na każdym etapie indukcji, a także podczas stosowania ich w klasyfikacji. Jeśli nie zdefiniowano żadnych dodatkowych kryteriów i preferencji, to w algorytmach pokryciowych ocena reguł ukierunkowana jest na takie sterowanie procesem indukcji, aby wyznaczone reguły pokrywały jak najwięcej przykładów pozytywnych i jak najmniej przykładów negatywnych.

Z punktu widzenia zdolności opisowych oceny reguł dokonuje się w celu wyselekcjonowania spośród nich tych najbardziej interesujących, najbardziej użytecznych, najbardziej unikalnych czy też najbardziej istotnych w sensie statystycznym [83, 106, 128, 132, 151, 179, 187, 215, 223, 240, 257, 300, 307, 329, 341 itd.]. Część miar stosowanych do oceny zdolności opisowych często nie jest odpowiednia do nadzorowania procesu indukcji.

Dla perspektywy opisowej ważna jest także ocena zbioru reguł; w tym przypadku oceniana jest m.in.: liczba reguł, liczba warunków elementarnych występujących w regułach oraz to, w jakim stopniu reguły wzajemnie się pokrywają [342]. Do oceny zdolności klasyfikacyjnych zbiorów reguł stosowane są miary zdefiniowane w rozdziale 2. Dla rozważanych przez nas klasyfikatorów dyskretnych miarami tymi są: całkowita dokładność klasyfikacji (2.13), wrażliwość i specyficzność klasyfikatora ze względu na klasę decyzyjną (2.20,2.21), średnia dokładność klas decyzyjnych (2.26) oraz inne miary, omówione w podrozdziale 2.5.1. Przeglądu podstawowych kryteriów oceny reguł i zbiorów reguł

decyzyjnych dokonują Yao i Zhou [342], wprowadzając pojęcia oceny *mikro*, która dotyczy pojedynczych reguł, i oceny *makro*, która dotyczy zbiorów reguł.

W praktycznych zastosowaniach systemów regułowych często wymagane jest, aby ocena uwzględniała specyficzne oczekiwania użytkownika [345]. Niekiedy możliwe jest wówczas zdefiniowanie miary złożonej, biorącej pod uwagę jednocześnie wszystkie kryteria jakości podane przez użytkownika [254, 257]. Jeśli zdefiniowanie takiego złożonego kryterium jest niemożliwe, ocena wykonywana jest oddziennie, ze względu na każde z kryteriów szczegółowych, a ostateczny ranking reguł otrzymuje się po posortowaniu ich zgodnie z porządkiem leksykograficznym [22,277].

Niezależnie od tego, czy oceniana jest pojedyncza reguła czy zbiór reguł, oceny reguł dokonuje się za pomocą pewnej funkcji przyjmującej wartości w zbiorze liczb rzeczywistych.

**Definicja 3.1.** Niech  $RUL$  będzie zbiorem reguł decyzyjnych, a  $\Gamma$  – rodziną zbiorów reguł decyzyjnych. Funkcję  $q : RUL \rightarrow R$  nazywamy miarą oceniającą jakość reguł, funkcję  $Q : \Gamma \rightarrow R$  nazywamy miarą oceniającą jakość zbioru reguł.

Powyższa definicja jest ogólna i nie mówi wiele o miarach jakości. Zanim przedstawione zostaną miary jakości, dokonujące ocen mikro i makro, konieczne jest wyjaśnienie pewnych kwestii terminologicznych i przyjętych oznaczeń. Po pierwsze, słowo „miara” zostało tutaj użyte niezgodnie z definicją miary, stosowaną w matematyce, gdyż wartości miar rozważanych w dalszej części rozdziału nie zawsze są nieujemne. Nazwa „miara” używana będzie jedynie po to, aby pozostać w zgodzie z ogólnie przyjętym w literaturze przedmiotu angielskim terminem *measure* [106, 132, 187, 240]. Po drugie, nazwa „miary(-a) oceniające(-a) jakość reguł” odnosi się w niniejszej pracy zarówno do miar jakości reguł decyzyjnych (ang. *quality measures*, *evaluation metrics*, *learning heuristics*, *search heuristics*), jak i do miar oceniających atrakcyjność reguł (ang. *attractiveness measure*, *interestingness measure*). W publikacjach te dwie kategorie są rozróżniane, dlatego że *quality masures* używane są do nadzorowania procesu wzrostu i przycinania reguł [8, 40, 90, 251], podczas gdy *attractiveness measures* wykorzystywane są do ustalania rankingu i filtracji reguł już wyznaczonych (w tym reguł asocjacyjnych) [83, 106, 128, 132, 151, 179, 187, 240, 329]. Zbiory *quality masures* i *attractiveness measures* nie są identyczne, mają jednak dużą część wspólną, wiele miar funkcjonujących jako miary atrakcyjności funkcjonuje również jako miary jakości [3,143,132,340]. Ponieważ niniejsza monografia koncentruje się na regułach decyzyjnych, stosowany w niej będzie wspólny termin *miara jakości*, przy czym w razie konieczności akcentowany będzie cel oceny. Miary jakości najczęściej dzielone są na dwie kategorie [128, 132, 240, 187]: miary obiektywne (ang. *objective measures*) oraz miary subiektywne (ang. *subjective measures*). Najwięcej miar obiektywnych definiowanych jest na podstawie tablicy kontyngencji, opisującej regułę w kontekście pokrywanej przez nią liczby

przykładów pozytywnych i negatywnych, oraz ogólnej liczby przykładów pozytywnych i negatywnych, znajdujących się w analizowanym zbiorze danych. Do miar obiektywnych zalicza się również miary oceniające budowę reguł (*ang. form dependent measures*). Zadaniem tego typu miar jest ocena zwięzłości (*ang. conciseness*) i/lub osobliwości (*ang. peculiarity*) wyrażenia reprezentowanego przez regułę. Obiektywne miary oceniające budowę reguł wykorzystują m.in. informacje o: liczbie warunków elementarnych, z jakich zbudowana jest reguła, rozmiarze zakresów warunków elementarnych oraz syntaktycznym podobieństwie reguł [91, 197, 345, 254, 314].

Zadaniem miar subiektywnych jest ocena reguł ze względu na subiektywne preferencje użytkownika. Informacje o przykładach pozytywnych i negatywnych, pokrywanych przez regułę, oraz o rozmiarze klas decyzyjnych stanowią tutaj jedynie uzupełnienie dla innych czynników, które zależą od specyficznej dziedziny zastosowania oraz wiedzy dziedzinowej i każdorazowo mogą być inne. Przykładowo w pracach [117, 257] autor wraz z Grucą zdefiniował miarę złożoną, przeznaczoną do badania jakości reguł i zbiorów reguł stosowanych do funkcjonalnego opisu genów.

Zadaniem miar subiektywnych jest mierzenie takich cech reguły, jak: osobliwość, stopień nieoczekiwania (*ang. unexpectedness, surprisingness*), nowość (*ang. novelty*), użyteczność (*ang. usefulness*) czy też aplikowalność (*ang. actionability*) [132, 218, 345]. Miary oceniające użyteczność oraz możliwość zastosowania reguł zostały zaliczone do grupy tzw. miar semantycznych (*ang. semantic measures*) [132].

Warto także wspomnieć o jeszcze jednej klasyfikacji miar jakości; otóż miary te mogą być tzw. miarami korzyści (*ang. gain-type measures*) lub miarami kosztu (*ang. cost-type measures*). Miary korzyści mają taką cechę, że im wyższa jest ich wartość, tym lepsza jest ocena reguły. W przypadku miar kosztu jest odwrotnie: im wyższa jest wartość miary, tym gorszą ocenę otrzymuje reguła.

### **3.1. Miary definiowane na podstawie tablicy kontyngencji**

Niech dane są: reguła  $r \equiv \varphi \rightarrow \psi$  oraz zbiór przykładów  $E$ . Tabelę 3.1 nazywać będziemy tablicą kontyngencji dla reguły  $r$  oraz zbioru przykładów  $E$ .

Tabela 3.1

Tablica kontyngencji dla reguły  $r \equiv \varphi \rightarrow \psi$ 

	$ [\varphi] $	$ [\neg\varphi] $	
$ [\psi] $	$p$	$P - p$	$P$
$ [\neg\psi] $	$n$	$N - n$	$N$
	$p + n$	$P + N - p - n$	$P + N$

Oznaczenia w tabeli 3.1 są zgodne z oznaczeniami przyjętymi w rozdziale 2:  $[\varphi]$  ( $[\neg\varphi]$ ) to zbiór przykładów pokrywanych (niepokrywanych) przez przesłankę reguły,  $[\psi]$  ( $[\neg\psi]$ ) to zbiór przykładów pokrywanych (niepokrywanych) przez konkluzję reguły,  $p$  ( $n$ ) to liczba przykładów pozytywnych (negatywnych), pokrywanych przez regułę,  $P - p$  ( $N - n$ ) to liczba przykładów pozytywnych (negatywnych), niepokrywanych przez regułę,  $P$  ( $N$ ) to liczba wszystkich przykładów pozytywnych (negatywnych), znajdujących się w zbiorze  $E$ .

W niniejszej publikacji poświęcamy również nieco uwagi regułom regresyjnym, w szczególności takim, w konkluzjach których znajduje się konkretna wartość liczbową ( $\psi \equiv d = v$ ). Zakładając, że zdefiniowane są wartości progowe  $\varepsilon_1, \varepsilon_2$ , dla ustalonej reguły  $r = \varphi \rightarrow \psi$  zbiór  $[\psi]$  zawiera przykłady o wartościach atrybutu decyzyjnego, należących do przedziału  $[v - \varepsilon_1, v + \varepsilon_2]$ . Pozostałe przykłady tworzą zbiór przykładów negatywnych  $[\neg\psi]$ .

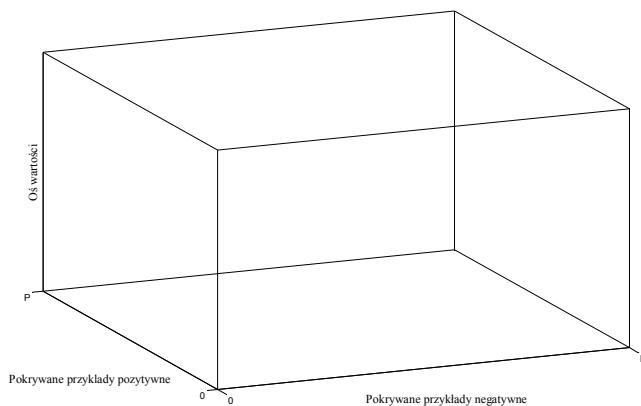
Dla obiektywnych miar jakości, definiowanych na podstawie tablicy kontyngencji, określono wiele własności teoretycznych. Pewna grupa własności, związana z monotonicznością miar, ma szczególne znaczenie z punktu widzenia stosowania miar do nadzorowania procesu indukcji reguł [90, 223, 264]. Więcej własności dedykowanych jest dla miar oceniających zdolności opisowe reguł [43, 73, 106, 113, 128, 132, 240, 300].

Aby zobrazować, w jaki sposób zmieniają się wartości miary wraz ze zmieniającą się liczbą przykładów pokrywających regułę, dokonuje się wizualizacji wartości miary. Zauważmy, że dla miar definiowanych na podstawie tablicy kontyngencji, dla ustalonej reguły  $r$  i odpowiadającej jej tablicy kontyngencji 3.1 zapis  $q(r)$  można umownie zastąpić przez  $q(p, n, P - p, N - n)$ , a ten z kolei można uprościć do postaci  $q(p, n)$ . Jest to możliwe, gdyż znając regułę i zbiór przykładów, na podstawie którego ma zostać wyznaczona wartość miary, można w jednoznaczny sposób określić wartości  $p, n, P - p, N - n$ . Oczywiście działanie odwrotne jest niemożliwe, gdyż z informacji o  $p$  i  $n$  nie można jednoznacznie wnioskować o tym, jaka jest postać reguły  $r$ . Niemożliwe jest także posługiwanie się zapisem  $q(p, n)$ , jeśli znane są jedynie: postać reguły  $r$  oraz wartości  $p$  i  $n$ , a nie jest znana całkowita liczba przykładów pozytywnych i negatywnych. Zapisem  $q(p, n, P - p, N - n)$  będziemy

posługiwać się jedynie wtedy, gdy dla prowadzonych rozważań istotne będą wartości  $p, n, P - p, N - n$ , a nie to, jaka jest konkretna postać reguły.

### 3.1.1. Przestrzeń wartości

Założymy, że dany jest zbiór  $E$ , zawierający przykłady reprezentujące co najmniej dwie klasy decyzyjne. Założymy, że przykłady należące do jednej z tych klas (oznaczmy ją jako  $X$ ) będą traktowane jako przykłady pozytywne, a wszystkie pozostałe – jako przykłady negatywne. Dowolną miarę jakości, definiowaną na podstawie tablicy kontyngencji 3.1 i oceniającą w zbiorze  $E$  reguły opisujące klasę  $X$ , można przedstawić jako funkcję zmiennych  $p$  i  $n$ . Wykres wartości miary rysowany jest w przestrzeni trójwymiarowej. Można także dokonać wizualizacji dwuwymiarowej, rysując poziomice wykresu trójwymiarowego. Na rysunku 3.1 zaznaczony jest obszar, w obrębie którego rysowany jest wykres wartości miary. Obszar ten to przestrzeń wartości miary (lub prostu – przestrzeń wartości).



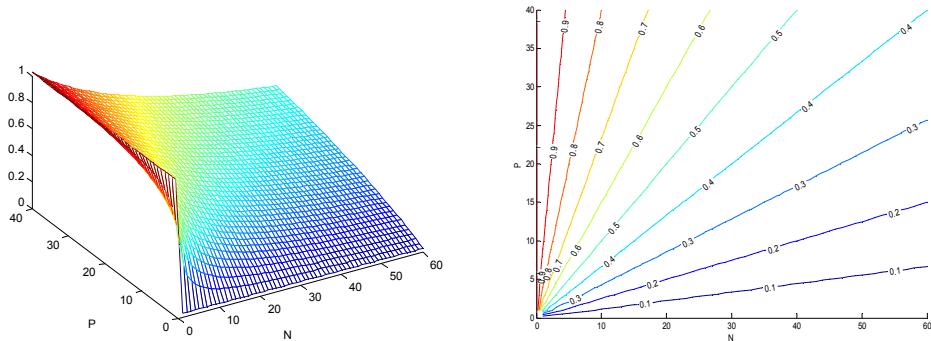
Rys. 3.1. Przestrzeń wartości miary  
Fig. 3.1. The measure values space

W przestrzeni wartości dopuszcza się sytuację, w której  $p=0$ . Jest to założenie czysto teoretyczne, gdyż w praktyce algorytm indukcji reguł decyzyjnych zawsze zwraca regułę pokrywającą przynajmniej jeden przykład pozytywny.

Dwuwymiarowa wizualizacja wartości miary odbywa się za pomocą poziomic. Graficznie poziomice reprezentowane są przez pewną krzywą, która powstała w wyniku przecięcia trójwymiarowego wykresu z płaszczyzną równoległą do osi przykładów.

**Definicja 3.2.** Niech dana jest miara  $q$ , zdefiniowana na podstawie tablicy kontyngencji. Poziomicą miary  $q$  o wartości  $k \in R$  nazywamy krzywą łączącą wartości miary  $q$  dla wszystkich par uporządkowanych  $\langle p, n \rangle$ , takich że  $q(p, n) = k$ .

**Przykład 3.1.** Założymy, że dana jest miara  $q(p, n) = p / (p + n)$ , która w literaturze znana jest jako *precision* [90]. Wykresy miary  $q$  przedstawiono na rysunku 3.2.



Rys. 3.2. Wykresy miary *precision*  
Fig. 3.2. Graphs of the *precision* measure

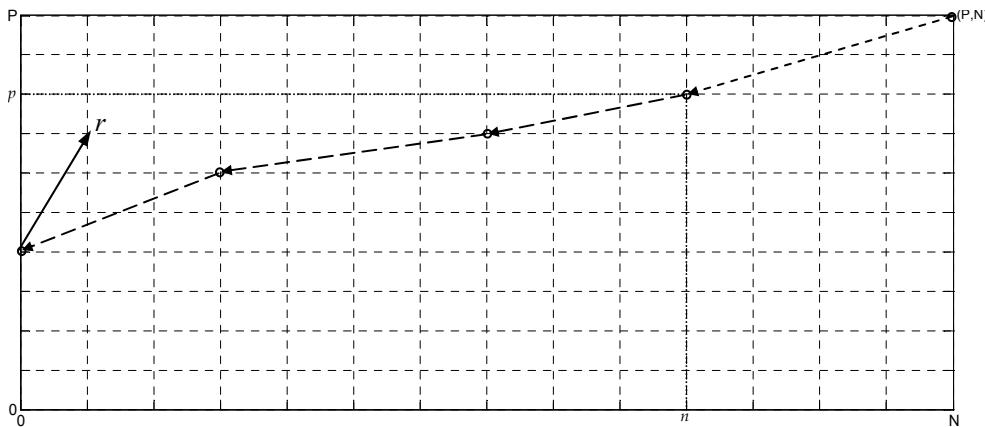
Oba wykresy prezentowane na rysunku 3.2 dają pełny obraz zachowania się miary w przestrzeni wartości. W przypadku wykresu dwuwymiarowego dodatkową informację stanowią wartości umieszczone na poziomikach.

Przedstawiony sposób wizualizacji jest intuicyjny i zaproponowany został jednocześnie przez kilku autorów. Jedną z pierwszych propozycji wizualizacji wartości miar jakości przedstawił Bruha [40]; wizualizacja odbywała się w układzie dwuwymiarowym. Oś odciętych reprezentowała liczbę wszystkich przykładów pokrywanych przez regułę, na osi rzędnych umieszczone zostały wartości miary. Trójwymiarowy sposób wizualizacji autor przedstawił w pracy [244].

Wizualizacja dwuwymiarowa za pomocą poziomików upowszechniła się po artykule Fürnkranza i Flacha [90]. W artykule tym autorzy wprowadzili pojęcie przestrzeni pokrycia (ang. *coverage space*), posługując się analogią do przestrzeni ROC (ang. *ROC space*) [75]. Przestrzeń wartości miary jest przestrzenią pokrycia, rozszerzoną o dodatkowy wymiar, który informuje o wartościach miary.

W przestrzeni wartości i przestrzeni pokrycia można śledzić proces indukcji reguły. Na rysunku 3.3 przedstawiono przestrzeń pokrycia z naniesionymi punktami, odpowiadającymi kolejnym krokom indukcji reguły (faza wzrostu – linia przerywana, faza przycinania – linia ciągła). Zgodnie z tym, co napisano w rozdziale 2, większość pokryciowych algorytmów indukcji reguł działa w ten sposób, że początkowo tworzona reguła pokrywa wszystkie możliwe przykłady. Na rysunku 3.3 reguła taka charakteryzowana jest przez punkt o współrzędnych  $\langle P, N \rangle$ . Następnie, w miarę dodawania kolejnych warunków elementarnych, punkt, którego współrzędne odpowiadają aktualnie pokrywanej liczbie przykładów pozytywnych i negatywnych, zbliża się do osi przykładów pozytywnych  $Y$ . Oznacza to, że reguła jest coraz bardziej specjalizowana i pokrywa coraz mniej przykładów negatywnych. Jeśli algorytm tworzy reguły dokładne, a analizowany zbiór przykładów jest niesprzeczny, wynikowa reguła pokrywa  $p$  przykładów pozytywnych

i żadnego przykładu negatywnego, co odpowiada punktowi o współrzędnych  $\langle p,0 \rangle$ . Idealna – z punktu widzenia klasyfikacji przykładów zawartych w zbiorze treningowym – jest reguła pokrywająca wszystkie przykłady pozytywne i żadnego przykładu negatywnego (punkt  $\langle P,0 \rangle$ ).



Rys. 3.3. Wizualizacja procesu tworzenia reguły w przestrzeni pokrycia  
Fig. 3.3. Visualization of the rule induction process in the coverage space

Po fazie wzrostu realizowana jest faza przycinania, której zadaniem jest poprawa jakości reguły. Ponieważ fazy wzrostu i przycinania realizowane są najczęściej za pomocą heurystycznych procedur przeszukiwania, nie ma gwarancji, że wynikowa reguła, oznaczona na rysunku jako  $r$ , jest regułą o jakości wyższej od reguł utworzonych w kolejnych etapach fazy wzrostu.

### 3.1.2. Własności

W literaturze przedmiotu tematyka własności miar jakości traktowana jest zbiorczo w odniesieniu do reguł asocjacyjnych i decyzyjnych. Jedynie w kilku pracach, [106, 179, 223, 300], przeprowadzono analizę wybranych własności pod kątem ich użyteczności dla miar oceniających jakość reguł decyzyjnych. Do chwili obecnej nie ma opracowania, w którym w uporządkowany i jednolity sposób określono by zestaw własności przeznaczonych dla miar oceniających jakość reguł decyzyjnych.

W dalszej części rozdziału przedstawione zostaną znane z literatury własności miar jakości. Przeprowadzone zostaną analizy równoważności, zależności i ewentualnych sprzeczności pomiędzy własnościami. Na tym tle omówiony zostanie ujednolicony zestaw własności, którymi (zdaniem autora) powinny charakteryzować się miary przeznaczone do oceny reguł decyzyjnych.

Założymy, że dana jest miara jakości  $q$ , definiowana na podstawie tablicy kontyngencji. W dalszej części przedstawiono najważniejsze, znane z literatury własności, definiowane dla miary  $q$ :

- P1  $q(p, n, P - p, N - n) = 0$  jeżeli  $\frac{p}{p + n} = \frac{P}{P + N}$ ;
- P2<sub>1</sub>  $q(p, n, P - p, N - n) < q(p + k, n - k, P - p - k, N - n + k)$ ; miara jest funkcją rosnącą ze względu na  $p$ , przy niezmieniających się wartościach  $p + n, P, N$ ;
- P2<sub>2</sub>  $q(p, n, P - p, N - n) < q(p + k, n, P - p - k, N - n)$ ; miara jest funkcją rosnącą ze względu na  $p$ , przy niezmieniających się wartościach  $n, P, N$ ;
- P2<sub>3</sub>  $q(p, n, P - p, N - n) > q(p, n + k, P - p, N - n - k)$ ; miara jest funkcją malejącą ze względu na  $n$ , przy niezmieniających się wartościach  $p, P, N$ ;
- P3  $q(p, n, P - p, N - n) > q(p, n, P - p + k, N - n - k)$ ; miara jest funkcją malejącą ze względu na  $P$ , przy niezmieniających się wartościach  $p, n, P + N$ ;
- T1  $q(p, n, P - p, N - n) = q(p, P - p, n, N - n)$  wtedy i tylko wtedy, gdy  $q(\varphi \rightarrow \psi) = q(\psi \rightarrow \varphi)$ ; miara jest symetryczna za względu na transpozycję tablicy kontyngencji;
- T2  $q(p, n, P - p, N - n) = q(k_1 p, k_1 n, k_2 (P - p), k_2 (N - n))$  oraz  $q(p, n, P - p, N - n) = q(k_1 p, k_2 n, k_1 (P - p), k_2 (N - n))$ ,  $k_1, k_2 \in R^+$ ; wartość miary nie zmienia się po przeskalowaniu kolumn lub przeskalowaniu wierszy;
- T3<sub>1</sub>  $q(p, n, P - p, N - n) = -q(P - p, N - n, p, n)$  wtedy i tylko wtedy, gdy  $q(\varphi \rightarrow \psi) = -q(\neg\varphi \rightarrow \psi)$ ; zamiana kolumn w tablicy kontyngencji powoduje zmianę znaku wartości miary;
- T3<sub>2</sub>  $q(p, n, P - p, N - n) = -q(n, p, N - n, P - p)$ , wtedy i tylko wtedy, gdy  $q(\varphi \rightarrow \psi) = -q(\varphi \rightarrow \neg\psi)$ ; zamiana wierszy w tablicy kontyngencji powoduje zmianę znaku wartości miary;
- T4  $q(p, n, P - p, N - n) = q(N - n, P - p, n, p)$ , wtedy i tylko wtedy, gdy  $q(\varphi \rightarrow \psi) = q(\neg\varphi \rightarrow \neg\psi)$ ; zamiana kolumn i wierszy w tablicy kontyngencji nie powoduje zmiany wartości miary;
- T5  $q(p, n, P - p, N - n) = q(p, n, P - p, N - n + k)$ ; zwiększenie liczby przykładów negatywnych, niepokrywanych przez regułę, nie wpływa na wartość miary;
- B2  $q(p, n, P - p, N - n) > 0$ , jeżeli  $\frac{p}{p + n} > \frac{P}{P + N}$ ;

- B3  $q(p, n, P - p, N - n) < 0$ , jeżeli  $\frac{p}{p + n} < \frac{P}{P + N}$ ;
- BGK  $q(p, n, P - p, N - n) = const$ , jeżeli  $p = n$ ; wartość miary jest stała, jeśli liczba pokrywających ją przykładów pozytywnych i negatywnych jest identyczna;
- M1  $q(p, n, P - p, N - n) \leq q(p + k, n, P - p, N - n)$ ; miara jest funkcją niemalejącą ze względu na  $p$ , przy niezmieniających się wartościach  $n, P - p, N - n$ ;
- M2  $q(p, n, P - p, N - n) \geq q(p, n, P - p + k, N - n)$ ; miara jest funkcją nierosnącą ze względu na  $(P - p)$ , przy niezmieniających się wartościach  $p, n, N - n$ ;
- M3  $q(p, n, P - p, N - n) \geq q(p, n + k, P - p, N - n)$ ; miara jest funkcją nierosnącą ze względu na  $n$ , przy niezmieniających się wartościach  $p, P - p, N - n$ ;
- M4  $q(p, n, P - p, N - n) \leq q(p, n, P - p, N - n + k)$ ; miara jest funkcją niemalejącą ze względu na  $(N - n)$ , przy niezmieniających się wartościach  $n, p, P$ .

Oznaczenia poszczególnych własności wskazują na nazwiska autorów, którzy je zaproponowali, bądź wskazują na angielską nazwę danej własności. Własności P po raz pierwszy zaproponował Piatetsky-Shapiro [223], a następnie były one przywoływanie i modyfikowane przez różnych autorów [83, 132, 151, 307] jako pożądane dla miar oceniających atrakcyjność reguł. Własności P<sub>21</sub>, P<sub>22</sub>, P<sub>23</sub> opisują monotoniczność miary przy zmieniającej się liczbie przykładów pozytywnych i negatywnych, pokrywanych przez regułę. Własności P<sub>21</sub>, P<sub>22</sub>, P<sub>23</sub> zakładają, że ogólna liczba przykładów oraz proporcja pomiędzy liczbą przykładów pozytywnych i negatywnych nie zmieniają się ( $P/N = const$ ). Własność P<sub>3</sub> mówi o zmianach wartości miary, w sytuacji gdy nie zmienia się liczba przykładów pozytywnych i negatywnych, pokrywanych przez regułę, ale na korzyść  $P$  zmienia się proporcja pomiędzy liczbą przykładów pozytywnych i negatywnych ( $P/N$  rośnie). Własność P<sub>3</sub> definiowana jest – podobnie jak P<sub>21</sub>, P<sub>22</sub>, P<sub>23</sub> – dla niezmieniającej się, ogólnej liczby przykładów.

Własności T przedstawione zostały w pracy [307], w której Tan wraz ze współpracownikami rozszerza zbiór o nowe własności, związane z różnymi operacjami wykonywanymi na tablicy kontyngencji. Własności T<sub>1</sub>, T<sub>31</sub>, T<sub>32</sub>, T<sub>4</sub> mówią o symetrii miary i pokrywają się z własnościami zaproponowanymi wcześniej przez Carnapa [46]. Eells i Fitelson [73] wskazują, które z rodzajów symetrii proponowanych przez Carnapa są odpowiednie dla miar będących miarami konfirmacji (ang. *confirmation measures*). Miarami tymi nazywane są miary mające własności P<sub>1</sub>, B<sub>1</sub> i B<sub>2</sub>. Wyniki prac Eella i Fitelsona wykorzystano w pracach [43, 106, 300], w których przeprowadza się analizę symetrii kilku miar stosowanych do ceny reguł decyzyjnych. Idąc za terminologią przedstawioną w [73] i konsekwentnie stosowaną w [43, 106, 300], własność T<sub>1</sub> to tzw. *commutativity symmetry*,

własności  $T_{3_1}$  i  $T_{3_2}$  to odpowiednio *evidence symmetry* (symetria względem przesłanki) i *hypothesis symmetry* (symetria względem hipotezy lub symetria hipotetyczna), natomiast własność  $T_4$  to *total symmetry*. Własność  $T_4$  wynika z  $T_{3_1}$  i  $T_{3_2}$ , każda miara charakteryzująca się  $T_{3_1}$  i  $T_{3_2}$  ma również własność  $T_4$ .

Poprzez T2 i T5 wyrażono postulat, że wartość miary nie powinna się zmieniać po pomnożeniu przez liczbę dodatnią wybranych kolumn (wartości w kolumnach), wierszy lub komórek w tablicy kontyngencji. W książce pod redakcją Guilleta i Hamiltona [128] oraz w przeglądach [132, 307] dokonano analizy kilkudziesięciu miar atrakcyjności reguł pod kątem posiadania przez nie własności P i T. Własność T2 ma związek z przedstawioną przez Blancharda [27] klasyfikacją miar jakości. W klasyfikacji tej rozróżnia się m.in. miary o naturze opisowej i naturze statystycznej. Wartości miar opisowych nie zmieniają się, jeśli wartości komórek tablicy kontyngencji pomnożone zostaną przez liczbę dodatnią. Odpowiada

to własności T2, w której  $k_1 = k_2$ . Miary o naturze statystycznej nie charakteryzują się własnością T2 (gdy  $k_1 \neq k_2$ ). W szczególności wartości miar o naturze statystycznej zmieniają się wraz ze zmieniającą się liczbą przykładów. Dzieje się tak nawet wtedy, gdy proporcje pomiędzy wartościami komórek tablicy kontyngencji pozostają niezmienione.

Za pomocą P1, B1 i B2 wyrażono postulaty dotyczące wartości miar. Dla dowolnej reguły  $\varphi \rightarrow \psi$  oraz dowolnego zbioru przykładów  $E$  wartość miary powinna być dodatnia, jeśli  $Pr(\psi | \varphi) > Pr(\psi)$ . Innymi słowy, dla reguły  $\varphi \rightarrow \psi$  wartość miary powinna być dodatnia, jeśli w zbiorze  $E$  prawdopodobieństwo przynależności przykładu do zbioru  $[\psi]$ , pod warunkiem jego przynależności do zbioru  $[\varphi]$ , jest większe od prawdopodobieństwa, że przykład wybrany losowo z  $E$  należy do  $[\psi]$ . W sytuacji przeciwej wartość „miary” powinna być liczbą ujemną. Jeśli wspomniane prawdopodobieństwa są sobie równe, to wartość miary powinna być zrówna 0. Jak już wspomniano, miarę mającą jednocześnie własności P1, B1, B2 nazywa się Bayesowską miarą konfirmacji (ang. *Bayesian confirmation measure*) lub po prostu miarą konfirmacji. Tematyka możliwości zastosowania miar konfirmacji do oceny atrakcyjności reguł podejmowana była m.in. w pracach [43, 106], a w szczególności w [300]. Poddano w nich analizie kilka miar atrakcyjności pod kątem posiadania przez nie własności konfirmacji.

Własność BGK mówi o zachowaniu miary wtedy, gdy liczba przykładów negatywnych i pozytywnych, pokrywanych przez regułę, jest identyczna. Blanchard [27] postuluje, aby w takiej sytuacji wartość miary była stała (*eq*). Własność BGK nie bierze pod uwagę rozkładu pomiędzy liczbą przykładów pozytywnych i negatywnych. Miary charakteryzujące się własnością P1 dostarczają innej informacji niż miary mające własność BGK. Założymy, że dana jest reguła  $\varphi \rightarrow \psi$  oraz miary  $q_1$  i  $q_2$ , mające odpowiednio własności P1 i BGK. Jeśli

$q_1(\varphi \rightarrow \psi) > 0$ , to uzyskujemy informację, która można zinterpretować w następujący sposób: jeśli  $\varphi$  jest spełnione, to  $\psi$  jest spełnione częściej niż w sytuacji, w której nie mamy żadnej informacji o spełnieniu  $\varphi$ . Jeśli  $q_2(\varphi \rightarrow \psi) > eq$ , to uzyskujemy informację, którą można zinterpretować jako: jeśli  $\varphi$  jest spełnione, to  $\psi$  najczęściej również jest spełnione.

Własności M1, M2, M3, M4 podane zostały przez Greco, Pawlaka i Słowińskiego, podobnie jak P<sub>21</sub>, P<sub>22</sub>, P<sub>23</sub>, P<sub>3</sub>; związane są one z monotonicznością miar. W przeciwnieństwie do P<sub>21</sub>, P<sub>22</sub>, P<sub>23</sub> monotoniczność definiowana za pomocą M1, M2, M3, M4 nie zakłada, że proporcja pomiędzy liczbą przykładów pozytywnych i negatywnych w zbiorze danych jest stała ( $P/N \neq const$ ). Własności M1, M2, M3, M4 opierają się na założeniu, że o tyle o ile rośnie zmieniająca się wartość jednego z argumentów, o tyle rośnie liczba przykładów w rozważanym zbiorze danych. Można zauważyć, że własności M1, P<sub>21</sub>, P<sub>22</sub> dotyczą monotoniczności miary ze względu na rosnącą wartość  $p$ . Własności M2, P<sub>3</sub> dotyczą monotoniczności ze względu na rosnącą wartość  $P$ , a własności M3, P<sub>23</sub> monotoniczności – ze względu na zwiększające się  $n$ . Własność M4 ma związek z zaproponowaną przez Tana [307] własnością T5, przy czym można zauważyć, że T5 jest szczególnym przypadkiem M4.

W artykule przeglądowym [97] autorzy dodatkowo rozważają pięć własności zaproponowanych przez Lencę i współpracowników [178]. Jedynie dwie z nich mają związek z tablicą kontyngencji. Można je zapisać następująco:

- L1 – wartość miary jest identyczna dla wszystkich reguł, dla których nie istnieją kontrprzykłady;
- L2 – wartość miary rośnie wraz ze wzrostem ogólnej liczby przykładów i przy niezmienionych proporcjach w tablicy kontyngencji.

Geng i Hamilton interpretują L1 w taki sposób, że wartość miary dla wszystkich reguł niepokrywających żadnych przykładów negatywnych powinna być identyczna, niezależnie od tego, ile reguła pokrywa przykładów pozytywnych. Druga interpretacja L1 może być taka, że wartości miary dla wszystkich reguł pokrywających wszystkie przykłady pozytywne i żadnego przykładu negatywnego powinny być identyczne. Własność L2 jest szczególnym przypadkiem T2, gdy  $k_1 = k_2$ . Jeśli  $k_1 = k_2$ , oznacza to, że wartości wszystkich komórek tablicy kontyngencji zostały pomnożone przez liczbę dodatnią. Lenca i współpracownicy [178] argumentują, że reguła wyznaczona z większej liczby przykładów jest lepsza od reguły o podobnej charakterystyce, ale wyznaczonej na podstawie mniejszego zbioru przykładów. Własność L2 wyraża postulat, że do oceny atrakcyjności reguł powinno się używać jedynie miary o naturze statystycznej.

Własność L1 w nieco szerszym kontekście rozważana była w środowisku badaczy koncentrujących się na analizie własności miar konfirmacji. W artykule [59] Crupi wraz ze współpracownikami definiuje własności  $Ex_1$  i L, wiążące wartości miar konfirmacji ( $c$ ) z sytuacjami ekstremalnymi, polegającymi na pokrywaniu przez ocenianą regułę jedynie przykładów pozytywnych lub negatywnych. Wykorzystując pojęcie wynikania logicznego  $\models$ , definiują oni funkcję  $v(\psi, \varphi)$  w taki sposób, że  $v(\psi, \varphi) = V$  (np.  $V=1$ ), jeśli  $\varphi \models \psi$ ,  $v(\psi, \varphi) = -V$ , jeśli  $\varphi \models \neg\psi$ , oraz  $v(\varphi, \psi) = 0$  w każdej innej sytuacji. Własności  $Ex_1$  i L można wtedy zapisać w następujący sposób:

- $Ex_1$  – jeżeli  $v(\psi_1, \varphi_1) > v(\psi_2, \varphi_2)$ , wówczas  $c(\varphi_1 \rightarrow \psi_1) > c(\varphi_2 \rightarrow \psi_2)$ ,
- L – wartość  $c(\varphi \rightarrow \psi)$  jest maksymalna, jeśli  $\varphi \models \psi$ ; wartość  $c(\varphi \rightarrow \psi)$  jest minimalna, jeśli  $\varphi \models \neg\psi$ .

Zgodnie z przyjętą w niniejszej monografii terminologią, miary mające własność  $Ex_1$  przyporządkowują zawsze wyższą ocenę regule  $\varphi_1 \rightarrow \psi_1$ , niepokrywającej żadnego przykładu negatywnego, niż każdej innej regule  $\varphi_2 \rightarrow \psi_2$ , pokrywającej jakieś przykłady negatywne. Zauważmy, że porównywane reguły mogą wskazywać na różne klasy decyzyjne. Miary mające własność  $Ex_1$  przyporządkowują zawsze niższą ocenę regule  $\varphi_1 \rightarrow \psi_1$ , pokrywającej jedynie negatywne przykłady, niż każdej innej regule  $\varphi_2 \rightarrow \psi_2$ , pokrywającej jakiś przykład pozytywny. Własność L określa, że miara powinna przyjmować wartość maksymalną, gdy reguła jest dokładna, a minimalną, gdy nie pokrywa żadnego przykładu pozytywnego. Na L można popatrzać jako na uogólnienie L1, gdyż L1 wymaga jedynie, aby ocena reguł dokładnych była identyczna.

Ponieważ wiele z miar konfirmacji nie spełniało warunku definiowanego przez własność  $Ex_1$ , w pracach [59, 112] przedstawiono propozycje normalizacji kilku miar konfirmacji. Normalizacja powodowała, że miary miały własności  $Ex_1$ . W pracy [112] Greco, Słowiński i Szczęch wykazali, że sposób normalizacji, jaki proponuje Crupi [59], powoduje, że wiele miar konfirmacji zmienia się w miary równoważne ze względu na porządek reguł (o tym typie równoważności będzie mowa w dalszej części rozdziału). Wykorzystując różne sposoby określania maksymalnej i minimalnej wartości miary (tabela 3.2) Greco, Słowiński i Szczęch przedstawiają własne propozycje normalizacji miar konfirmacji, rozszerzając w ten sposób zbiór miar oceny atrakcyjności reguł.

Tabela 3.2  
Metody normalizacji miar konfirmacji [112]

Wartości tablicy kontyngencji po normalizacji	Propozycja Nicoda		Bayesowska		Na podstawie prawdopodobieństwa		Propozycja Crupiego	
	Max.	Min.	Max.	Min.	Max.	Min.	Max.	Min.
$p'$	$p+n$	0	$P$	0	$P+n$	0	$p+n$	0
$n'$	0	$p+n$	0	$N$	0	$p+n$	0	$p+n$
$P-p'$	$P-p$	$P-p$	0	$P$	0	$P+N-p-n$	$P-p-n$	$P$
$N-n'$	$N-n$	$N-n$	$N$	0	$P+N-p-n$	0	$N$	$N-n-p$

W pracy [113] Greco, Słowiński i Szczęch wskazują również na pewne paradoksy  $\text{Ex}_1$  i  $L$ , wynikające z faktu, że własności te pod uwagę biorą jedynie sytuacje ekstremalne (reguły dokładne i całkowicie niedokładne) oraz nie uwzględniają pokrycia reguł. Jako rezultat swoich rozważań autorzy ci proponują słabsze formy  $\text{Ex}_1$  i  $L$ , które można przedstawić w sposób następujący:

- $w\text{Ex}_1$  – jeżeli  $v(\psi_1, \varphi_1) > v(\psi_2, \varphi_2)$  oraz  $v(\psi_1, \neg\varphi_1) < v(\psi_2, \neg\varphi_2)$ , wówczas  $c(\varphi_1 \rightarrow \psi_1) > c(\varphi_2 \rightarrow \psi_2)$ ,
- $wL$  – wartość  $c(\varphi \rightarrow \psi)$  jest maksymalna wtedy i tylko wtedy, gdy  $\varphi \models \psi$  oraz  $\neg\varphi \models \neg\psi$ , wartość  $c(\varphi \rightarrow \psi)$  jest minimalna wtedy i tylko wtedy, gdy  $\varphi \models \neg\psi$  oraz  $\neg\varphi \models \psi$ .

W pracy [113] autorzy piszą, że słaba postać  $\text{Ex}_1$  gwarantuje, że miara konfirmacji nie może osiągnąć wartości maksymalnej, jeśli nie zostaną spełnione warunki  $\varphi \models \psi$  i  $\neg\varphi \models \neg\psi$ . Własność  $w\text{Ex}_1$  gwarantuje również, że miara konfirmacji nie może osiągnąć wartości minimalnej, jeśli nie zostaną spełnione warunki  $\varphi \not\models \psi$  i  $\neg\varphi \models \psi$ . Miary charakteryzujące się własnością  $w\text{Ex}_1$  porządkują reguły w taki sposób, że na początku rankingu znajdują się reguły pokrywające jedynie wszystkie przykłady pozytywne, a na końcu znajdują się reguły pokrywające jedynie wszystkie przykłady negatywne.

Warunki definiujące własność  $w\text{Ex}_1$  nie mówią nic o porównaniu jakości reguł dokładnych, pokrywających wszystkie przykłady pozytywne, oraz reguł dokładnych, ale niepokrywających wszystkich przykładów pozytywnych. Podobnie jest z porównaniem reguł pokrywających jedynie wszystkie przykłady negatywne z regułami pokrywającymi jedynie wybrany podzbiór zbioru przykładów negatywnych. Własność  $w\text{Ex}_1$  można wzmacnić w taki sposób, aby wymuszała ona również porównywanie wymienionych powyżej reguł. Zmodyfikowaną słabą własność  $\text{Ex}_1$  ( $mw\text{Ex}_1$ ) można zapisać jako:

- $mw\text{Ex}_1$  – jeżeli  $(v(\psi_1, \varphi_1) > v(\psi_2, \varphi_2))$  lub  $v(\psi_1, \varphi_1) = v(\psi_2, \varphi_2) = V$ , lub  $v(\psi_1, \varphi_1) = v(\psi_2, \varphi_2) = -V$  i  $v(\psi_1, \neg\varphi_1) < v(\psi_2, \neg\varphi_2)$ , wówczas  $c(\varphi_1 \rightarrow \psi_1) > c(\varphi_2 \rightarrow \psi_2)$ .

Miary charakteryzowane przez  $mwEx_1$  porządkują reguły w taki sposób, że na początku rankingu znajdują się reguły dokładne, pokrywające wszystkie przykłady pozytywne, a na końcu znajdują się reguły całkowicie niedokładne, pokrywające jedynie wszystkie przykłady negatywne. Reguły dokładne, ale niepokrywające wszystkich przykładów pozytywnych, reguły niedokładne, ale pokrywające przynajmniej jeden przykład pozytywny oraz reguły pokrywające jedynie pewien podzbior zbioru przykładów negatywnych znajdują się w środku rankingu. Warunki własności  $mwEx_1$  nie mówią nic na temat porównania jakości takich reguł. Daltego, podobnie jak w przypadku  $wEx_1$ , własność  $mwEx_1$  pozwala uniknąć paradoksów [113], zawartych w definicji  $Ex_1$ .

Zauważmy, że w odniesieniu do reguł decyzyjnych i konkretnego zbioru danych warunek  $(\varphi|=\psi \wedge \neg\varphi|=\neg\psi)$  oznacza, że reguła pokrywa wszystkie przykłady pozytywne i żadnego przykładu negatywnego. Warunek  $(\varphi|\neq\neg\psi \wedge \neg\varphi|=\psi)$  oznacza, że reguła pokrywa wszystkie przykłady negatywne i żadnego przykładu pozytywnego. Warunki te są zgodne również z warunkami definiującymi  $wL$ .

Wszystkie sposoby normalizacji przedstawione w tabeli 3.2 zakładają, że miara osiąga maksimum, jeśli reguła nie pokrywa przykładów negatywnych, a minimum, jeśli reguła nie pokrywa przykładów pozytywnych. Jedynie Bayesowska metoda normalizacji zakłada, że najwyżej powinna być oceniona reguła pokrywająca wszystkie przykłady pozytywne i żadnego przykładu negatywnego, a najniżej – reguła pokrywająca wszystkie przykłady negatywne i żadnego przykładu pozytywnego. Normalizacja Bayesowska wykonywana jest z punktu widzenia konkluzji reguły. W metodzie bazującej na prawdopodobieństwie normalizacja wykonywana jest ze względu na przesłankę reguły. Reguła uzyskuje maksymalną ocenę, jeśli pokrywane przez nią przykłady negatywne staną się pozytywne, a niepokrywane przykłady pozytywne staną się negatywne. W pozostałych metodach zakłada się, że wartość miary powinna osiągnąć maksimum, jeśli zależność reprezentowaną przez regułę możliwa potraktować w kategorii wynikania logicznego. Oznacza to, że maksymalną ocenę uzyskują reguły dokładne, niekoniecznie pokrywające całą klasę decyzyjną. Szczegółową analizę i uzasadnienie metod normalizacji można znaleźć w [112, 113].

Na zakończenie zajmijmy się jeszcze klasyfikacją związaną z różną interpretacją wyrażeń warunkowych  $\varphi \rightarrow \psi$ , przedstawioną przez Blancharda [27]. Z klasyfikacją tą wiążą się pewne dodatkowe własności miar. Standardowo zależność reprezentowana przez regułę traktowana jest jako przybliżona implikacja logiczna, która dopuszcza istnienie kontrprzykładów. Zbiorami przykładów pozytywnych i negatywnych, pokrywanych przez regułę, są odpowiednio  $[\psi] \cap [\varphi]$  i  $[\psi] \cap [\neg\varphi]$ . Klasyfikacja zdefiniowana przez Blancharda zakłada, że reguła  $\varphi \rightarrow \psi$  może być traktowana jako:

- quasi-implikacja, w której jako przykłady pozytywne należy rozpatrywać zbiory  $[\psi] \cap [\varphi]$  oraz  $[\neg\psi] \cap [\neg\varphi]$ , a jako przykłady negatywne – zbiór  $[\neg\psi] \cap [\varphi]$ ; miary użyte do oceny tak interpretowanej reguły powinny spełniać warunek  $q(\varphi \rightarrow \psi) = q(\neg\psi \rightarrow \neg\varphi)$ ;
- quasi-koniunkcja, w której jako przykłady pozytywne należy rozpatrywać zbiór  $[\psi] \cap [\varphi]$ , a jako przykłady negatywne – zbiory  $[\psi] \cap [\neg\varphi]$  oraz  $[\neg\psi] \cap [\varphi]$ ; miary użyte do oceny tak interpretowanej reguły powinny spełniać warunek  $q(\varphi \rightarrow \psi) = q(\psi \rightarrow \varphi)$ ;
- quasi-równoważność, w której jako przykłady pozytywne należy rozpatrywać zbiory  $[\psi] \cap [\varphi]$  oraz  $[\neg\psi] \cap [\neg\varphi]$ , a jako przykłady negatywne – zbiory  $[\psi] \cap [\neg\varphi]$  oraz  $[\neg\psi] \cap [\varphi]$ ; miary użyte do oceny tak interpretowanej reguły powinny spełniać warunek  $q(\varphi \rightarrow \psi) = q(\psi \rightarrow \varphi) = q(\neg\psi \rightarrow \neg\varphi) = q(\neg\varphi \rightarrow \neg\psi)$ .

We wszystkich wymienionych przypadkach im większa liczba przykładów pozytywnych i mniejsza liczba przykładów negatywnych, pokrywanych przez  $\varphi \rightarrow \psi$ , tym quasi-implikacja, quasi-koniunkcja i quasi-równoważność są lepsze. Miary użyte do oceny quasi-implikacji biorą pod uwagę atrakcyjność reguł  $\varphi \rightarrow \psi$  i  $\neg\psi \rightarrow \neg\varphi$ . Miary użyte do oceny quasi-koniunkcji biorą pod uwagę atrakcyjność reguł  $\varphi \rightarrow \psi$  i  $\psi \rightarrow \varphi$ . Wreszcie miary użyte do oceny quasi-równoważności biorą pod uwagę jednocześnie atrakcyjność reguł  $\varphi \rightarrow \psi$ ,  $\psi \rightarrow \varphi$ ,  $\neg\varphi \rightarrow \neg\psi$ ,  $\neg\psi \rightarrow \neg\varphi$ . Jeśli użytkownik zainteresowany jest jedynie oceną reguł o postaci  $\varphi \rightarrow \psi$ , a atrakcyjność reguł  $\psi \rightarrow \varphi$ ,  $\neg\varphi \rightarrow \neg\psi$ ,  $\neg\psi \rightarrow \neg\varphi$  nie jest istotna dla danej dziedziny zastosowania, to zgodnie z postulatami Blancharda użyta miara jakości nie powinna spełniać żadnej z wymienionych powyżej własności.

Przedstawiony spis stanowi uporządkowanie i ujednolicenie zapisu własności definiowanych dla miar atrakcyjności reguł. Większość publikacji traktujących o obiektywnych miarach atrakcyjności reguł koncentruje się na analizie pewnego podzbioru zaprezentowanych własności.

### **3.1.3. Własności wymagane podczas oceny reguł decyzyjnych**

Zanim przejdziemy do przedstawienia minimalnego zbioru własności, pożądanego dla reguł decyzyjnych, zdefiniowane zostaną dwie, ważne z punktu widzenia oceny reguł decyzyjnych, miary jakości. Miarami tymi są dokładność (ang. *precision*) oraz pokrycie (ang. *coverage*). Podane zostaną również trzy stwierdzenia, które będą pomocne w uzasadnieniu wyboru pożądanych własności miar.

**Definicja 3.2.** Niech  $E$  jest zbiorem przykładów, a  $r$  – regułą decyzyjną. Dokładność i pokrycie reguły  $r$  w zbiorze  $E$  definiowane są następująco:

$$\text{precision}(r) = \frac{p}{p+n}, \quad \text{coverage}(r) = \frac{p}{P}.$$

W literaturze związanej z miarami jakości istnieją pewne nieścisłości terminologiczne, związane z miarami *precision* i *coverage*. Miara *precision* w części publikacji funkcjonuje także jako *accuracy*, *consistency* [8,40,42] lub *confidence*, *certainty* [6, 106]. Miara *coverage* występuje też jako *recall* [90]. Część autorów, zwłaszcza zajmujących się indukcją i oceną reguł asocjacyjnych, jako miarę podstawową wymienia wsparcie (ang. *support*) reguły, definiowane po prostu jako  $p$ , lub wsparcie przesłanki, definiowane jako  $p+n$  [15].

Zakresem obu przedstawionych w definicji 3.2 miar jest przedział  $[0,1]$ . Obie miary są miarami korzyści. Pokryciowe algorytmy indukcji dążą do utworzenia dla każdej klasy decyzyjnej jednej, dokładnej reguły, pokrywającej wszystkie przykłady pozytywne, a jeśli jest to niemożliwe, nadzcznąą ideą jest indukcja reguł o jak największej dokładności i jak największym pokryciu. W przypadku algorytmów indukcji reguł decyzyjnych dla celów opisowych mamy do czynienia z bardzo podobną sytuacją. Część z tych algorytmów działa na zasadzie generowania pokrycia zbioru treningowego [177]. Pozostałe algorytmy konstruowane są w taki sposób, aby generować reguły spełniające wymogi minimalnej dokładności i ogólności (np. algorytm Explore [285]).

W języku prawdopodobieństwa *precision* można zapisać jako prawdopodobieństwo warunkowe  $P(\psi | \varphi)$ .

Z badań empirycznych wynika, że zazwyczaj wraz ze wzrostem dokładności reguły maleje jej pokrycie, zachodzi również sytuacja odwrotna. Zależności pomiędzy dokładnością i pokryciem nie można wyrazić analitycznie, gdyż to, w jakim stopniu wzrost dokładności przyczynia się do spadku pokrycia (i odwrotnie), uzależnione jest od analizowanego zbioru danych.

Miary jakości definiowane na podstawie tablicy kontyngencji są wskaźnikami wyrażającymi kompromis pomiędzy oceną dokładności a oceną pokrycia reguły. Dodatkowo, część miar podczas oceny bierze pod uwagę informację o proporcji pomiędzy liczbą wszystkich przykładów pozytywnych i negatywnych.

**Stwierdzenie 3.1.** Niech dany jest zbiór złożony z  $P > 0$  przykładów pozytywnych i  $N > 0$  przykładów negatywnych. Niech  $p, n, P - p, N - n$  są wartościami w tablicy kontyngencji, charakteryzującej regułę  $r_1$ , oraz  $p + k, n, P - p, N - n$ ,  $k \geq 0$  są wartościami w tablicy kontyngencji, charakteryzującej regułę  $r_2$ . Wówczas  $\text{precision}(r_1) \leq \text{precision}(r_2)$  oraz  $\text{coverage}(r_1) \leq \text{coverage}(r_2)$ .

Dowód: wystarczy udowodnić, że  $\frac{p}{p+n} \leq \frac{p+k}{p+k+n}$  oraz  $\frac{p}{P} \leq \frac{p+k}{P+k}$ . Ponieważ wszystkie składniki ułamków są nieujemne, podane nierówności możemy zapisać w postaciach równoważnych,

$p(p+k+n) \leq (p+n)(p+k)$  oraz  $p(P+k) \leq P(p+k)$ , po uproszczeniu uzyskując  $p^2 + pk + pn \leq p^2 + pk + pn + nk$  i  $pP + pk \leq pP + Pk$ . Łatwo zauważać, że wszystkie składniki sum są liczbami nieujemnymi, w szczególności  $P > 0$  oraz  $p \leq P$ , przytoczone nierówności są zatem zawsze spełnione. W przypadku gdy  $n = 0$  i/lub  $k = 0$  dokładność reguł  $r_1, r_2$  będzie identyczna. Jeśli  $p = P$ , identyczne będzie pokrycie reguł  $r_1, r_2$ . ◆

**Stwierdzenie 3.2.** Niech dany jest zbiór złożony z  $P > 0$  przykładów pozytywnych i  $N > 0$  przykładów negatywnych. Niech  $p, n, P-p, N-n$  są wartościami w tablicy kontyngencji, charakteryzującej regułę  $r_1$ , oraz  $p, n+k, P-p, N-n$  są wartościami w tablicy kontyngencji, charakteryzującej regułę  $r_2$ . Wówczas  $\text{precision}(r_1) > \text{precision}(r_2)$ .

Dowód stwierdzenia 3.2 jest analogiczny do dowodu stwierdzenia 3.1, dlatego też zostanie pominięty.

**Stwierdzenie 3.3.** Niech dane są dwie reguły:  $r_1 \equiv \varphi \rightarrow \psi$  oraz  $r_2 \equiv \psi \rightarrow \varphi$ . W zbiorze przykładów  $E$  pomiędzy dokładnością a pokryciem reguł  $r_1$  i  $r_2$  zachodzą następujące związki:  $\text{precision}(r_1) = \text{coverage}(r_2)$  oraz  $\text{precision}(r_2) = \text{coverage}(r_1)$ .

Dowód: klasę decyzyjną, na którą wskazuje reguła  $r_2$ , tworzą wszystkie przykłady spełniające przesłankę  $r_1$ . Aby udowodnić stwierdzenie, wystarczy na podstawie tabeli 3.1 i definicji 3.2 wyznaczyć dokładność i pokrycie reguły  $r_1$ , a następnie transponować macierz przedstawioną w tabeli 3.1 i wyznaczyć dokładność i pokrycie reguły  $r_2$ . Uzyskane wyniki przedstawiają się następująco:  $\text{precision}(r_1) = p/(p+n)$ ,  $\text{coverage}(r_1) = p/P$  oraz  $\text{precision}(r_2) = p/P$ ,  $\text{coverage}(r_2) = p/(p+n)$ . ◆

Poniżej podano dwie grupy własności, jakimi zdaniem autora powinny charakteryzować się miary jakości przeznaczone do oceny reguły decyzyjnych. Podane własności odnoszą się do miar korzyści. Celem autora nie było definiowanie nowych własności, nowych oznaczeń własności ujęto jedynie po to, aby ujednolicić zapis oraz, aby nazwa informowała o tym, czego dana własność dotyczy.

Zakładając, że indukcja oraz porównanie jakości reguł odbywają się na podstawie ustalonego zbioru przykładów, miary nadzorujące proces indukcji i stosowane do porównywania jakości reguł decyzyjnych powinny być monotoniczne w sposób taki, jak definiują to własności P2<sub>2</sub> i P2<sub>3</sub>. Oznaczmy P2<sub>2</sub> jako M<sub>p</sub> oraz P2<sub>3</sub> jako M<sub>n</sub>. Własności M<sub>p</sub> i M<sub>n</sub> stanowią minimalny zbiór własności związanych z monotonicznością miar przeznaczonych do nadzorowania procesu indukcji reguł decyzyjnych.

Własności M1, M2, M3, M4, P3 definiowane są dla zmieniającego się zbioru przykładów. Prześledźmy, jakie są konsekwencje monotoniczności definiowanej przez M1,

M<sub>2</sub>, M<sub>3</sub>, M<sub>4</sub>. W tabeli 3.3 pokazano, jak zmieniają się proporcje pomiędzy liczbą przykładów pozytywnych i negatywnych oraz jaki wpływ na dokładność i pokrycie reguł mają zmieniające się wartości w tablicy kontyngencji (informację tę umieszczono w kolumnie *Kierunek zmian*). W tabeli 3.3 przyjęto następujące oznaczenia:  $D=P/N$  oznacza proporcję pomiędzy liczbą przykładów pozytywnych i negatywnych; zapis  $X^{\uparrow}$  oznacza, że wartość  $X$  rośnie; zapis  $X^{\uparrow\rightarrow}$  oznacza, że wartość  $X$  nie maleje. Analogicznie należy rozumieć zapisy  $X^{\downarrow}$ ,  $X^{\downarrow\rightarrow}$ .

Tabela 3.3

## Własności związane z monotonicznością miar

Nazwa	Opis	Kierunek zmian $D$ , $precision$ , $coverage$
monotoniczność ze względu na p		
M <sub>1</sub>	$q(p, n, P - p, N - n) \leq q(p + k, n, P - p, N - n)$	$D^{\uparrow}$ , $precision^{\uparrow\rightarrow}$ , $coverage^{\uparrow\rightarrow}$
M <sub>p</sub>	$q(p, n, P - p, N - n) < q(p + k, n, P - p - k, N - n)$	$D^{\rightarrow}$ , $precision^{\uparrow\rightarrow}$ , $coverage^{\uparrow}$
monotoniczność ze względu na (P-p)		
M <sub>2</sub>	$q(p, n, P - p, N - n) \geq q(p, n, P - p + k, N - n)$	$D^{\uparrow}$ , $precision^{\rightarrow}$ , $coverage^{\downarrow}$
monotoniczność ze względu na n		
M <sub>3</sub>	$q(p, n, P - p, N - n) \geq q(p, n + k, P - p, N - n)$	$D^{\downarrow}$ , $precision^{\downarrow}$ , $coverage^{\rightarrow}$
M <sub>n</sub>	$q(p, n, P - p, N - n) > q(p, n + k, P - p, N - n - k)$	$D^{\rightarrow}$ , $precision^{\downarrow}$ , $coverage^{\rightarrow}$
monotoniczność ze względu na (N-n)		
M <sub>4</sub>	$q(p, n, P - p, N - n) \leq q(p, n, P - p, N - n + k)$	$D^{\downarrow}$ , $precision^{\rightarrow}$ , $coverage^{\rightarrow}$

Dla własności M<sub>1</sub> oraz M<sub>3</sub> podczas określania kierunku zmian dokładności i pokrycia reguł wykorzystano stwierdzenia 3.1 i 3.2. W przypadku pozostałych własności określenie kierunku zmian jest proste i intuicyjne.

Interpretacja własności M<sub>p</sub> i M<sub>n</sub> jest następująca:

M<sub>p</sub> – miara powinna być funkcją rosnącą dla rosnącego  $p$  i równocześnie niezmieniającego się zbioru przykładów pozytywnych; wartość  $q$  powinna rosnąć, gdyż zwiększa się dokładność i pokrycie reguły lub zwiększa się tylko pokrycie reguły, jeśli reguła przed wzrostem  $p$  była regułą dokładną; do M<sub>p</sub> można podać własność równoważną M<sub>(P-p)</sub>, zdefiniowaną jako  $q(p, n, P - p, N - n) > q(p - k, n, P - p + k, N - n)$ ; dla M<sub>(P-p)</sub> mamy do czynienia z następującymi kierunkami zmian:  $D^{\rightarrow}$ ,  $precision^{\downarrow\rightarrow}$ ,  $coverage^{\downarrow}$ ;

M<sub>n</sub> – miara powinna być funkcją malejącą dla rosnącego  $n$  i równocześnie niezmieniającego się zbioru przykładów negatywnych; wartość  $q$  powinna maleć, gdyż maleje dokładność przy niezmieniającym się pokryciu; do M<sub>n</sub> można podać własność równoważną M<sub>(N-n)</sub>, zdefiniowaną jako  $q(p, n, P - p, N - n) < q(p, n - k, P - p, N - n + k)$ ; dla M<sub>(N-n)</sub> mamy do czynienia z następującymi kierunkami zmian:  $D^{\rightarrow}$ ,  $precision^{\uparrow}$ ,  $coverage^{\rightarrow}$ .

Motywacje zdefiniowania własności  $M_1, M_2, M_3, M_4$  przedstawiły Greco, Pawlak i Śłowiński w pracy [106]. Argumenty przemawiające za wykorzystaniem słabych nierówności (funkcji niemalejących i nierosnących) były takie, że wartość miary może nie zależeć od tej komórki, której wartości się zmieniają. Przywołani autorzy argumentują, że w takich sytuacjach wartości miary mogą pozostać niezmienione.

Zamieniając w  $M_2$  i  $M_4$  nierówności słabe na ostre, otrzymamy funkcje malejącą i rosnącą. Oznaczmy  $M_2$  i  $M_4$ , zawierające ostre nierówności, jako  $M_{(P-p)P}$  (rośnie  $(P-p)$ , a wraz z tym cały zbiór  $P$ ) oraz  $M_{(N-n)N}$  (rośnie  $(N-n)$ , a wraz z tym zbiór  $N$ ). Uzasadnienie dla  $M_{(P-p)P}$  i  $M_{(N-n)N}$  przedstawia się następująco:

$M_{(P-p)P}$  – miara powinna być funkcją malejącą dla rosnącego  $(P-p)$  i równocześnie zwiększającego się zbioru przykładów pozytywnych; wartość  $q$  powinna maleć, gdyż dokładność reguły pozostaje niezmieniona, natomiast spada jej pokrycie w związku ze zwiększającym się zbiorem przykładów pozytywnych;

$M_{(N-n)N}$  – miara powinna być funkcją rosnącą dla rosnącego  $(N-n)$  i równocześnie zwiększającego się zbioru przykładów negatywnych; za wzrostem wartości  $q$  przemawia zmieniająca się proporcja  $D$ , dokładność i pokrycie pozostają bez zmian, ale ponieważ wzrosła liczba przykładów negatywnych, więc przy tej zmienionej proporcji regułę należy ocenić wyżej.

Własność  $M_{(N-n)N}$  stanowi zaprzeczenie T5 [307].  $M_{(N-n)N}$  pozostaje jednak w zgodzie z argumentacją stanowiącą, że wraz ze wzrostem liczby przykładów negatywnych, które nie są pokrywane przez regułę, zaufanie do reguły powinno wzrastać. Własności  $M_{(P-p)P}$  i  $M_{(N-n)N}$  zdefiniowano, zakładając, że wartości miary powinny zależeć od wartości wszystkich komórek tablicy kontyngencji. Ostra nierówność w  $M_{(N-n)N}$  pozwala na przeprowadzenie dowodu stwierdzenia 3.5.

**Stwierdzenie 3.4.** Niech dana jest miara jakości  $q$ , zdefiniowana na podstawie tablicy kontyngencji. Jeśli  $q$  ma własności  $M_p$  i  $M_n$ , to ma również własność  $P2_1$ .

Dowód: własność  $P2_1$  definiowana jest dla niezmieniającego się zbioru przykładów pozytywnych i negatywnych. Z założenia wynika, że miara  $q$  ma własności  $M_p$  i  $M_n$ . Zamiast  $M_n$  możemy posługiwać się równoważną własnością  $M_{(N-n)}$ . Na podstawie  $M_p$ , a następnie  $M_{(N-n)}$ , prawdziwe są następujące nierówności:  $q(p, n, P - p, N - n) < q(p + k, n, P - p - k, N - n) < q(p + k, n - k, P - p - k, N - n + k)$ .

Wynika stąd, że miara ma własność  $P2_1$ . ◆

**Stwierdzenie 3.5.** Niech dana jest miara jakości  $q$ , zdefiniowana na podstawie tablicy kontyngencji. Jeśli  $q$  ma własności  $M_{(P-p)P}$  i  $M_{(N-n)N}$ , to ma również własność  $P3$ .

Dowód: na podstawie  $M_{(P-p)P}$ , a następnie  $M_{(N-n)N}$ , prawdziwe są następujące nierówności:  
 $q(p,n,P-p,N-n) > q(p,n,P-p+k,N-n) > q(p,n,P-p+k,N-n-k)$ .  $\diamond$

**Stwierdzenie 3.6.** Niech dana jest miara jakości  $q$ , zdefiniowana na podstawie tablicy kontyngencji. Jeśli  $q$  ma własności  $M_1$  i  $M_{(P-p)P}$ , to ma również własność  $M_p$ .

Dowód: z własności  $M_1$  otrzymamy  $q(p,n,P-p,N-n) \leq q(p+k,n,P-p,N-n)$ .

Na podstawie  $M_{(P-p)P}$  otrzymamy  $q(p+k,n,P-p-k,N-n) > q(p+k,n,P-p,N-n)$ .

Ostatecznie  $q(p,n,P-p,N-n) \leq q(p+k,n,P-p,N-n) < q(p+k,n,P-p-k,N-n)$ .  $\diamond$

**Stwierdzenie 3.7.** Niech dana jest miara jakości  $q$ , zdefiniowana na podstawie tablicy kontyngencji. Jeśli  $q$  ma własności  $M_3$  i  $M_{(N-n)N}$ , to ma również własność  $M_n$ .

Dowód jest analogiczny do dowodu stwierdzenia 3.6.  $\diamond$

**Uwaga 3.1.** Niech dana jest miara jakości  $q$ , zdefiniowana na podstawie tablicy kontyngencji. Jeśli miara ma własności  $M_p$  i  $M_n$ , to wartość miary osiąga maksimum dla reguły  $r$  takiej i tylko takiej, dla której  $precision(r)=1$  oraz  $coverage(r)=1$ .

Dowód: aby przeprowadzić dowód, wystarczy wykazać sprzeczność dwóch hipotez.

**(hip. h1)** Dla miary mającej własności  $M_p$  i  $M_n$  oraz przyjmującej wartość maksymalną  $max$  dla pewnej reguły  $r_1$  prawdziwe są nierówności  $precision(r_1) \neq 1$  lub  $coverage(r_1) \neq 1$ .

Rozpatrzmy przypadek  $coverage(r_1) \neq 1$ . Oznacza on, że liczba przykładów pozytywnych, pokrywanych przez  $r_1$ , jest mniejsza od  $P$ , czyli  $P = p + k$ , gdzie  $k > 0$ . Stąd na mocy własności  $M_p$  otrzymamy  $q(p,n,P-p,N-n) < q(p+k,n,0,N-n)$ . Oznacza to, że  $max$  nie jest maksymalną wartością miary, co jest sprzeczne z przyjętym założeniem. W przypadku  $precision(r_1) \neq 1$  reguła  $r_1$  pokrywa jakieś przykłady negatywne, a zatem  $n > 0$ . Na podstawie własności  $M_n$  wiadomo, że  $q(p,0,P-p,N) > q(p,n,P-p,N-n)$ . Wynika stąd, że  $max$  nie jest maksymalną wartością miary, co jest sprzeczne z przyjętym założeniem.

**(hip. h2)** Dla miary mającej własności  $M_p$  i  $M_n$  oraz reguły  $r$  o własnościach  $precision(r)=1$  i  $coverage(r)=1$   $max$  nie jest maksymalną wartością miary.

Skoro  $q(r)$  nie jest maksymalną wartością miary, istnieje reguła  $r_1 \neq r$  taka, że  $q(r_1) > q(r)$ . Z założeń o dokładności i pokryciu  $r$  wynika, że dla tej reguły  $p = P$  oraz  $n = 0$ . Przez  $p_1$  i  $n_1$  oznaczmy odpowiednio liczbę przykładów pozytywnych i negatywnych, pokrywanych przez regułę  $r_1$ . Skoro jakość reguł  $r_1$  i  $r$  jest różna, to mamy do czynienia z jednym z trzech przypadków  $p_1 = p + k_p$  i  $n_1 = n$ ,  $p_1 = p$  i  $n_1 = n + k_n$ ,  $p_1 = p + k_p$  i  $n_1 = n + k_n$ , w których  $k_p \neq 0$  i  $k_n \neq 0$ . Zauważmy, że w pierwszym z nich  $k_p$  musi być mniejsze od zera. Na podstawie monotoniczności  $M_p$  oznacza to, że  $q(r_1) < q(r)$ , co jest sprzeczne z przyjętym

założeniem, że  $q(r_1) > q(r)$ . W drugim przypadku  $k_n$  musi być większe od zera. Na podstawie monotoniczności  $M_n$  również to oznacza, że  $q(r_1) < q(r)$  i jest sprzeczne z założeniem, że  $q(r_1) > q(r)$ . Wreszcie dla trzeciej możliwości uzyskamy ciąg nierówności  $q(p_1, n_1, P - p_1, N - n_1) < q(P, n_1, 0, N - n_1) < q(P, 0, 0, N)$ . Pierwsza nierówność wynika z własności  $M_p$ , a druga – z własności  $M_n$ . Z nierówności tych wynika, że  $q(r_1) < q(r)$ , co oczywiście jest sprzeczne z założeniem, że  $q(r_1) > q(r)$ .

Wykazanie sprzeczności hipotez  $h1$  i  $h2$  potwierdza prawdziwość uwagi 3.1. ◆

Powiązanie monotoniczności z dokładnością i pokryciem reguł rozważał już Piatetsky-Shapiro [223]. Zdefiniowane przez niego własności gwarantują, że tezy uwagi 3.1 są spełnione. Dowód przedstawiony powyżej jest formalnym potwierdzeniem rozumowania opisanego w [223].

**Uwaga 3.2.** Niech dana jest miara jakości  $q$ , zdefiniowana na podstawie tablicy kontyngencji, mająca własności  $M_p$  i  $M_n$ . Załączmy, że dana jest reguła  $r \equiv \varphi \rightarrow \psi$  oraz przez  $r_{\neg}$  oznaczmy regułę  $\varphi \rightarrow \neg\psi$ . Wartość miary  $q$  osiąga minimum wtedy i tylko wtedy, gdy  $precision(r_{\neg}) = 1$  oraz  $coverage(r_{\neg}) = 1$ .

Dowód: dowód uwagi 3.2 jest analogiczny do dowodu uwagi 3.1. Dowód uwagi 3.2 można również przeprowadzić, korzystając z równoważności  $precision(r_{\neg}) = 1 \Leftrightarrow precision(r) = 0$  oraz  $coverage(r_{\neg}) = 1 \Leftrightarrow coverage(r) = 0$ . Równoważność  $precision(r_{\neg}) = 1 \Leftrightarrow \neg precision(r) = 0$  wynika z faktu, że  $precision(r_{\neg}) = 1 \Leftrightarrow n_{\varphi\psi} = 0 \Leftrightarrow \neg precision(r) = 0$ . Podobnie  $coverage(r_{\neg}) = 1 \Leftrightarrow coverage(r) = 0$  wynika z faktu, że  $coverage(r_{\neg}) = 1 \Leftrightarrow n_{\varphi\psi} = 0 \Leftrightarrow coverage(r) = 0$ .

Dla miar kosztu warunki monotoniczności będą przeciwnostawne: tam, gdzie miara korzyści ma być miarą rosnącą, tam miara kosztu ma być miarą malejącą itd.

Własności  $M_p$  i  $M_n$  będą również odpowiednie do nadzorowania procesu indukcji regresyjnych. W indukcji reguł regresyjnych, w fazach wzrostu i przycinania, rozmiary zbiorów przykładów pozytywnych i negatywnych mogą się dynamicznie zmieniać. Własność  $M_p$  gwarantuje, że spośród dwóch reguł, do których przypisane są równoliczne zbiory przykładów pozytywnych  $P$  i przykładów negatywnych  $N$  oraz pokrywających identyczną liczbę przykładów negatywnych, lepsza będzie reguła pokrywająca więcej przykładów pozytywnych. Własność  $M_n$  gwarantuje, że spośród dwóch reguł, do których przypisane są równoliczne zbiory przykładów pozytywnych  $P$  i przykładów negatywnych  $N$  oraz pokrywających identyczną liczbę przykładów pozytywnych, lepsza będzie reguła pokrywająca mniej przykładów negatywnych.

Druga grupa własności (tabela 3.4) przeznaczona jest dla miar przeznaczonych do oceny opisowej jakości reguł decyzyjnych. Podane własności dotyczą miar korzyści. W oznaczeniach stosowanych w tabeli 3.4 przyjęto, że  $const_1$  i  $const_2$  są pewnymi stałymi, których wartości są uzależnione od rozważanego zbioru przykładów. Przyjęto także, że  $r \equiv \varphi \rightarrow \psi$ ,  $\neg r \equiv \neg \varphi \rightarrow \psi$  oraz  $r \neg \equiv \varphi \rightarrow \neg \psi$ .

Tabela 3.4

## Własności miar związane z oceną zdolności opisowych reguł decyzyjnych

Nazwa	Opis
$D_{in}$	$q(r)=const_1$ wtedy i tylko wtedy, gdy $precision(r)=P/(P+N)$
$D_>$	$q(r)>const_1$ wtedy i tylko wtedy, gdy $precision(r)>P/(P+N)$
$D_<$	$q(r)<const_1$ wtedy i tylko wtedy, gdy $precision(r)<P/(P+N)$
$D_{eq}$	$q(r)=const_2$ wtedy i tylko wtedy, gdy $p=n$
$D_{ES}$	$q(r)=-q(\neg r)$ ; miara ma własność symetrii względem przesłanki
$D_{HS}$	$q(r)=-q(r \neg)$ ; miara ma własność symetrii względem hipotezy
$D_{wL}$	miara $q$ charakteryzuje się własnością wL
$D_{mwEx1}$	miara $q$ charakteryzuje się własnością mwEx <sub>1</sub>

Jaka była motywacja zaproponowania własności zawartych w tablicy 3.4? Własności  $D_{in}$ ,  $D_>$ ,  $D_<$  związane są z posiadaniem przez miarę cech zbliżonych do miar konfirmacji. Wartość stałej  $const_1$  uzależniona jest od konkretnego zbioru przykładów i ma ścisły związek z proporcją pomiędzy liczbą przykładów pozytywnych i negatywnych. Oczywiście miarę mającą własności  $D_{in}$ ,  $D_>$ ,  $D_<$  można znormalizować, tak aby jej zmieniona postać spełniała warunki konfirmacji (wartość stałej  $const_1$  będzie wtedy równa zero). W przypadku konkretnych miar normalizacja powodować będzie niepotrzebne komplikacje, związane m.in. z interpretacją wzoru definiującego miarę. Wyznaczenie w konkretnym zbiorze przykładów wartości stałej  $const_1$  nie jest trudne, wystarczy we wzorze definiującym miarę dokonać podstawień  $p := P$  oraz  $n := N$ .

Zdaniem autora  $D_{in}$ ,  $D_>$ ,  $D_<$  stanowią minimalny zestaw własności, jakimi powinny się charakteryzować miary przeznaczone do oceny jakości opisowej reguł decyzyjnych. Własność  $D_{in}$  wyraża postulat niezależności, definiowany przez różnych autorów, w tym przez Blancharda [27]. Wartość miary charakteryzującej się własnościami  $D_{in}$ ,  $D_>$ ,  $D_<$  wskazuje jednoznacznie na to, czy oceniana reguła jest potencjalnie interesująca w świetle rozważanego zbioru przykładów.

Własności  $D_{eq}$ ,  $D_{ES}$ ,  $D_{HS}$ ,  $D_{wL}$  oraz  $D_{mwEx1}$  stanowią uzupełnienie zestawu minimalnego, dostarczając dodatkowych informacji o ocenianej regule  $\varphi \rightarrow \psi$ , a także o jakości reguł  $\neg \varphi \rightarrow \psi$  i  $\varphi \rightarrow \neg \psi$ .

Własność  $D_{eq}$  to postulowana przez Blancharda własność BGK.

Analizę miar jakości reguł ze względu na różnego rodzaje symetrii przeprowadzono m.in. w [73, 106], gdzie podano wiele przykładów ilustrujących, że jedynym rodzajem symetrii, pożądanym dla miar traktujących reguły jako przybliżone implikacje logiczne, jest symetria względem hipotezy. Przykłady te przedstawiono w kontekście analizy miar konfirmacji. W pracy [300] autorka analizuje własności symetrii w oderwaniu od tego, czy dana miara ma własność konfirmacji czy nie. Argumentacja przemawiająca za daną własnością lub przeciwko niej jest taka sama jak ta, którą przytaczają Eellsa i Fitelson, a także Greco, Pawlak i Słowiński [106].

W przypadku symetrii  $D_{ES}$  występuje różnica pomiędzy publikacjami Tana [307] oraz Eellsa i Fitelsona [73, 106]. Własność  $D_{ES}$  to własność  $T3_1$ , którą Tan – podobnie jak  $D_{HS}$  ( $T3_2$ ) – nazywa antysymetrią [307]. Tan nie analizuje przydatności bądź nieprzydatności własności  $D_{ES}$  i  $D_{HS}$ , wskazuje jednak, które z popularnych miar charakteryzują się tymi własnościami. Eells i Fitelson przedstawiają natomiast pewne argumenty, które ich zdaniem przemawiają przeciwko  $D_{ES}$ .

W dalszej części przedstawiono argumentację przemawiającą za stosowaniem miar charakteryzujących się własnościami  $D_{ES}$ ,  $D_{HS}$ . Wspomagano się przy tym argumentami przywoływanymi przez Tana i jego współpracowników, a także niektórymi przykładami przedstawianymi przez Eellsa i Fitelsona.

Załóżmy, że dana jest talia kart, z której losowana jest jedna karta. Niech  $\varphi$  oznacza, że *karta jest siódemką pik*, niech  $\psi$  oznacza, że *kolor karty jest czarny*.

*Commutativity symmetry* (T1) nie jest pożądaną własnością dla miar jakości. Niech dane są dwie reguły:  $r_1 \equiv \varphi \rightarrow \psi$  (jeśli karta jest siódemką pik, to kolor karty jest czarny) oraz  $r_2 \equiv \psi \rightarrow \varphi$  (jeśli karta jest koloru czarnego, to karta jest siódemką pik). Autorzy prac [73, 106, 300] argumentują, że  $\psi$  nie potwierdza w tym samym stopniu  $\varphi$ , jak  $\varphi$  potwierdza  $\psi$ . Im wyższa wartość miary, tym silniejsze potwierdzenie. Wniosek z przedstawionego rozumowania jest taki, że reguły  $r_1$ ,  $r_2$  powinny uzyskać różne oceny. Dodatkowym argumentem przemawiającym za poprawnością przytoczonego rozumowania może być analiza dokładności i pokrycia obu reguł. Na podstawie stwierdzenia 3.3 wiadomo, że  $precision(r_1) = coverage(r_2)$  i  $coverage(r_1) = precision(r_2)$ . Zależności te są prawdziwe dla reguł  $r_1$ ,  $r_2$  w dowolnym zbiorze przykładów. Dla miary  $q$ , mającej własność *commutativity symmetry*, prawdziwa jest równość  $q(r_1) = q(r_2)$ . Założmy, że miarę  $q$  da się przedstawić jako funkcję  $f$ , której argumentami są dokładność i pokrycie, wtedy  $f(precision(r_1), coverage(r_1)) = f(precision(r_2), coverage(r_2))$ . Na mocy stwierdzenia 3.3 otrzymujemy  $f(precision(r_1), coverage(r_1)) = f(coverage(r_1), precision(r_1))$ . W szczególności reguła o dokładności 0.9 i pokryciu 0.1 uzyska taką samą ocenę, jak reguła o dokładności 0.1

i pokryciu 0.9. Miara mająca własność *commutativity symmetry* zakłada, że spadek dokładności może być rekompensowany identycznym wzrostem pokrycia. Tymczasem, jak argumentowano w wielu pracach dotyczących indukcji reguł [37,90,327], sposób, w jaki miara podczas oceny reguły uwzględnia informacje o jej dokładności i pokryciu, powinien być również uzależniony od proporcji pomiędzy liczbą przykładów pozytywnych i negatywnych, zawartych w rozważanym zbiorze przykładów.

W pracach Eells i Fitelsona przytaczana jest argumentacja, że symetria ze względu na przesłankę reguły (*evidence symmetry*) nie jest pożądaną własnością dla miar jakości. Argumentację tych badaczy przyjmują również Greco, Pawlak i Słowiński [106] oraz Szczęch [300]. Eells i Fitelson podają przykład dwóch reguł:  $r_1 \equiv \varphi \rightarrow \psi$  (jeśli karta jest siódemką pik, to kolor karty jest czarny) oraz  $r_2 \equiv \neg\varphi \rightarrow \psi$  (jeśli karta nie jest siódemką pik, to kolor karty jest czarny). Argumentują oni, że przesłanka  $\varphi$  silniej potwierdza hipotezę  $\psi$ , niż potwierdza ją  $\neg\varphi$ , gdyż z wiedzy o  $\varphi$  wprost wynika  $\psi$ , natomiast wiedza o  $\neg\varphi$  jest prawie zupełnie nieużyteczna ze względu na potwierdzenie hipotezy reprezentowanej przez  $\psi$ . Dlaczego zatem własność  $D_{ES}$  znalazła się w tabeli 3.4? Zdaniem autora argumentacja przedstawiona przez Eellsa i Fitelsona jest odpowiednia, jeśli na każdą z reguł  $r_1$ ,  $r_2$  patrzy się oddziennie. Popatrzmy na to, jak zmienia się nasza wiedza o konkluzji  $\psi$  na podstawie informacji o przesłance  $\varphi$  i jej negacji  $\neg\varphi$ . Wiedzę tę wyrażamy w języku reguł, a reguły charakteryzowane są przez dwie podstawowe miary, jakimi są dokładność i pokrycie. Reguła  $r_1$  charakteryzuje się dokładnością równą 1 i pokryciem 1/13 (0.0769). Reguła  $r_2$  charakteryzuje się dokładnością 12/51 (0.2352) i pokryciem 12/13 (0.9230). Jeśli regułę  $r_1$  zastąpimy regułą  $r_2$ , to uzyskamy spadek dokładności równy 0.7648 oraz wzrost pokrycia 0.8461. Jeśli regułę  $r_2$  zastąpimy regułą  $r_1$ , to uzyskamy wzrost dokładności równy 0.7648 oraz spadek pokrycia wynoszący 0.8461. Różnice pomiędzy dokładnościami i pokryciami reguł są stałe, ale różnią się znakiem. To spostrzeżenie przemawia, zdaniem autora, na korzyść symetrii ze względu na przesłankę reguły. Symetria ta dostarcza informacji nie tylko o jakości reguły  $\varphi \rightarrow \psi$ , będącej przedmiotem aktualnego zainteresowania, ale również o jakości reguły  $\neg\varphi \rightarrow \psi$ .

Symetria względem hipotezy (*hypothesis symmetry*) jest pożądaną własnością dla miar jakości. Niech dane są reguły  $r_1 \equiv \varphi \rightarrow \psi$  (jeśli karta jest siódemką pik, to kolor karty jest czarny) oraz  $r_2 \equiv \varphi \rightarrow \neg\psi$  (jeśli karta nie jest siódemką pik, to kolor karty nie jest czarny). Eells i Fitelson argumentują, że miara  $q$  powinna z taką samą siłą mierzyć wpływ  $\varphi$  na  $\psi$  jak wpływ  $\varphi$  na  $\neg\psi$ . Różnica w ocenie powinna polegać jedynie na różnych znakach wartości miary ( $q(r_1) = -q(r_2)$ ). Autorzy argumentują, że nie są w stanie przytoczyć żadnego

kontrprzykładu dyskwalifikującego symetrię względem hipotezy. Żadna dodatkowa argumentacja przemawiająca przeciwko temu rodzajowi symetrii nie jest też podawana w pracach [106, 300, 307].

Argumentacja przedstawiona powyżej zakłada, że reguły decyzyjne nie traktujemy jako quasi-koniunkcji ani quasi-równoważności. Gdyby przyjmować jedną z tych interpretacji, to zgodnie z argumentacją Blancharda [27] do oceny reguł należałoby użyć miar charakteryzujących się postulowanymi przez niego rodzajami symetrii.

Na zakończenie pozostaje analiza użyteczności własności  $D_{WL}$  i  $D_{mwEx1}$ . Oczywiście własność  $D_{WL}$  jest użyteczna, w szczególności dla ustalonego zbioru danych własność  $D_{WL}$  gwarantuje, że oceny najwyższą i najniższą uzyskują odpowiednio reguła najlepsza i reguła najgorsza z punktu widzenia analizy ich dokładności i pokrycia. Na podstawie uwag 3.1 oraz 3.2 można łatwo pokazać związek pomiędzy monotonicznościami  $M_p$  i  $M_n$  a własnością  $D_{WL}$ . Z uwagi tych wynika wprost, że z  $M_p$  i  $M_n$  wynika  $D_{WL}$ .

Własność  $D_{mwEx1}$  również jest użyteczna; miary charakteryzowane przez tę własność w sposób poządany porządkują reguły, ponadto uporządkowanie to dotyczy reguł, które mogą wskazywać na różne klasy decyzyjne.

Definiując minimalny zestaw własności pożądanego dla miar oceniających jakość reguł decyzyjnych i definiowanych na podstawie tablicy kontyngencji, zakładaliśmy, że zbiór danych jest ustalony i jego rozmiar nie zmienia się. Na podstawie tego założenia minimalny zbiór własności stanowią:  $M_p$ ,  $M_n$ ,  $D_{in}$ ,  $D_>$ ,  $D_<$ . Wymienione własności są istotne z punktu widzenia oceny zdolności opisowych reguł. Dodatkowo, pierwsze dwie własności są szczególnie istotne z punktu widzenia nadzorowania procesu indukcji reguł.

W przedstawionym zestawie nie ma własności T2, T5, L2, P3, gdyż dotyczą one sytuacji, w których zmienia się rozmiar zbioru danych lub proporcja pomiędzy liczbą przykładów pozytywnych i negatywnych. Z podobnego powodu w rekomendowanym zestawie nie umieszczono M1, M2, M3, M4. Własności M1–M4 opisują zachowanie wartości miary podczas porównywania jakości reguł w różnych zbiorach przykładów. Własności te definiują monotoniczność w sposób bardziej ogólny niż  $M_p$  i  $M_n$ ; dla naszych rozważań  $M_p$  i  $M_n$  są jednak wystarczające. Poprzez stwierdzenia 3.6 i 3.7 pokazano, że zamiana słabej nierówności w M2 i M4 na nierówność ostrą pozwala na powiązanie  $M_p$ ,  $M_n$  z M1–M4. Zamiana M2 (nierówność słaba) na  $M_{(P-p)P}$  (nierówność ostra) nie powoduje żadnych komplikacji związanych pojawiением się sprzeczności pomiędzy  $M_{(P-p)P}$  oraz innymi własnościami, a przytoczona przez autora argumentacja przemawiająca za stosowaniem  $M_{(P-p)P}$  jest prawidłowa.

Jako pożądanej dla reguł decyzyjnych nie wymieniono również własności L1, gdyż własność ta jest m.in. sprzeczna z zasadą mówiącą o tym, że spośród reguł o identycznej dokładności lepsza jest ta, która charakteryzuje się większym pokryciem.

Pomimo tego, że w zbiorze reguł regresyjnych trudno jest mówić o regułowym opisie klasy decyzyjnej, przedstawiony zestaw własności można również rekomendować dla miar definiowanych na podstawie tablicy kontyngencji i stosowanych do oceny reguł regresyjnych. Z zestawu tego należałoby usunąć symetrię  $D_{HS}$ , gdyż reguły regresyjne z konkluzjami  $d \neq v$ , gdzie  $v$  jest wartością ciągłego atrybutu decyzyjnego, zdaniem autora nie zawierają zbyt użytecznych informacji.

### **3.1.4. Równoważność i podobieństwo ze względu na porządek reguł**

Założmy, że  $M$  jest zbiorem obiektywnych miar jakości, zdefiniowanych na podstawie tablicy kontyngencji. W zbiorze tym możliwe jest rozważanie identyczności, równoważności oraz podobieństwa miar, przy czym identyczność, równoważność i podobieństwo mogą być badane *ze względu na ustanawiany przez nie porządek reguł* (krócej – *porządek reguł*), a także *ze względu na sposób rozstrzygania konfliktów podczas klasyfikacji*. Fürnkranz i Flach [90] definiują identyczność i równoważność miar ze względu na porządek reguł, przy czym definiowane przez nich identyczność i równoważność analizowane są w obrębie reguł wskazujących na identyczną klasę decyzyjną. Założmy, że zbiór  $RUL_X$  złożony jest z reguł wskazujących na dowolną klasę decyzyjną  $X$ .

**Definicja 3.3.** Dowolne miary  $q_1, q_2 \in M$  są identyczne ze względu na klasę decyzyjną  $X$  wtedy i tylko wtedy, gdy  $\bigvee_{RUL, RUL_X \subseteq RUL} \bigvee_{r_1, r_2 \in RUL_X} \bigvee_{(q_1(r_1) = q_1(r_2) \Leftrightarrow q_2(r_1) = q_2(r_2))} (q_1(r_1) = q_1(r_2) \Leftrightarrow q_2(r_1) = q_2(r_2))$ .

**Definicja 3.4.** (Równoważność ze względu na porządek reguł).

Dowolne miary  $q_1, q_2 \in M$  są równoważne ze względu na klasę decyzyjną  $X$  wtedy i tylko wtedy, gdy są miarami kompatybilnymi lub antagonistycznymi ze względu na  $X$ .

Dowolne miary korzyści  $q_1, q_2 \in M$  są równoważne ze względu na porządek reguł wskazujących na  $X$  wtedy i tylko wtedy, gdy są kompatybilne ze względu na  $X$ .

Dowolne miary  $q_1, q_2 \in M$ , z których jedna jest miarą korzyści, a druga jest miarą kosztu, są równoważne ze względu na porządek reguł wskazujących na  $X$  wtedy i tylko wtedy, gdy są antagonistyczne ze względu na  $X$ .

**Definicja 3.5.** Dowolne miary  $q_1, q_2 \in M$  są:

- kompatybilne ze względu na  $X$  wtedy i tylko wtedy, gdy:

$$\bigvee_{RUL, RUL_X \subseteq RUL, r_1, r_2 \in RUL_X} \left( q_1(r_1) > q_1(r_2) \Leftrightarrow q_2(r_1) > q_2(r_2) \right),$$

- antagonistyczne ze względu na  $X$  wtedy i tylko wtedy, gdy:

$$\bigvee_{RUL, RUL_X \subseteq RUL, r_1, r_2 \in RUL_X} \left( q_1(r_1) > q_1(r_2) \Leftrightarrow q_2(r_1) < q_2(r_2) \right).$$

Miary antagonistyczne można zamienić w miary kompatybilne, mnożąc wartości jednej z nich przez -1.

W definicjach 3.3, 3.4 i 3.5 identyczność, kompatybilność i antagonistyczność miar związane są z porównywaniem jakości reguł wskazujących na identyczną klasę decyzyjną. Dla celów indukcji reguł decyzyjnych i wyboru z opisu każdej klasy decyzyjnej reguł najbardziej interesujących takie porównania są wystarczające. Zauważmy jednak, że w przypadku porównywania jakości reguł asocjacyjnych przedstawione definicje identyczności i równoważności nie są odpowiednie, gdyż konkluzje reguł asocjacyjnych nie wskazują na klasy decyzyjne. W szczególności może zdarzyć się tak, że w zbiorze reguł asocjacyjnych konkluzja każdej z reguł będzie inna. Porównania jakości reguł asocjacyjnych mogą oczywiście dotyczyć podzbiorów reguł, których konkluzje są identyczne, ale częściej dotyczą reguł zbudowanych z różnych przesłanek i konkluzji. Chcąc badać identyczność i równoważność miar oceny atrakcyjności reguł asocjacyjnych, w definicjach 3.3, 3.4 i 3.5 możemy zrezygnować z warunku mówiącego o tym, że porównania dotyczą jedynie reguł o identycznych konkluzjach. W ten sposób uzyskamy nowe definicje: *bezwzględnej* identyczności, kompatybilności, antagonistyczności i równoważności miar. Ponieważ w obrębie naszego zainteresowania znajdują się przede wszystkim reguły decyzyjne, w dalszej części monografii w większości przypadków będą nas interesować identyczność, kompatybilność, antagonistyczność i równoważność miar *względem klasy decyzyjnej*.

Prostym wnioskiem, wynikającym z sygnalizowanej definicji *bezwzględnej* identyczności i równoważności miar, jest to, że każde miary identyczne lub równoważne w sposób bezwzględny będą identyczne lub równoważne ze względu na dowolną klasę decyzyjną. Łatwo również zauważyc, że definicje kompatybilności i antagonistyczności miar sprowadzają się do wymagania, aby wraz z monotonicznie zmieniającymi się wartościami jednej z miar, monotonicznie zmieniały się również wartości drugiej z nich.

W dalszej części, dla uproszczenia zapisu, będziemy posługiwać się sformułowaniami: *miary równoważne* (czyli kompatybilne lub antagonistyczne), *miary równoważne ze względu na porządek reguł* (czyli kompatybilne), *miary kompatybilne*, *miary antagonistyczne* – każdorazowo mając na myśli równoważność, kompatybilność i antagonistyczność względem klasy decyzyjnej. Jeśli w jakiejś części monografii będziemy odnosić się do równoważności, kompatybilności, antagonistyczności bezwzględnej, zostanie to wyraźnie zaakcentowane.

W dalszej części rozdziału skoncentrujemy się na miarach korzyści. Relację równoważności miar ze względu na porządek reguł oznaczmy przez  $Eq$ .

**Stwierdzenie 3.8.** Relacja równoważności miar jest relacją równoważności. Relacja  $Eq$  równoważności miar ze względu na porządek reguł jest relacją równoważności.

Dowód: zwrotność i symetryczność relacji równoważności miar oraz relacji  $Eq$  wynikają wprost z definicji identyczności, kompatybilności i antagonistyczności. W dowodzie przechodniości należy rozważyć trzy sytuacje (w dowodzie przechodniości relacji  $Eq$  wystarczy rozważyć pierwszą z nich). 1. Pary miar  $q_1, q_2$  i  $q_2, q_3$  są kompatybilne, wówczas na mocy równoważności zawartej w definicji kompatybilności uzyskuje się kompatybilność miar  $q_1, q_3$ . 2. Pary miar  $q_1, q_2$  i  $q_2, q_3$  są antagonistyczne, czyli  $q_1(r_1) > q_1(r_2) \Leftrightarrow q_2(r_1) < q_2(r_2)$  i  $q_2(r_1) < q_2(r_2) \Leftrightarrow q_3(r_1) > q_3(r_2)$ , a zatem  $q_1(r_1) > q_1(r_2) \Leftrightarrow q_3(r_1) > q_3(r_2)$ , co oznacza, że miary  $q_1, q_3$  są kompatybilne. 3. Para miar  $q_1, q_2$  jest kompatybilna, a para  $q_2, q_3$  jest antagonistyczna, wówczas zgodnie z definicją 3.5  $q_1(r_1) > q_1(r_2) \Leftrightarrow q_2(r_1) > q_2(r_2)$  i  $q_2(r_1) > q_2(r_2) \Leftrightarrow q_3(r_1) < q_3(r_2)$ , a zatem  $q_1(r_1) > q_1(r_2) \Leftrightarrow q_3(r_1) < q_3(r_2)$ , co oznacza, że miary  $q_1, q_3$  są antagonistyczne. ◆

Ze stwierdzenia 3.8 wynika, że równoważność  $Eq$  jest szczególnym przypadkiem równoważności miar. Miary, pomiędzy którymi zachodzi relacja  $Eq$ , są miarami równoważnymi. Natomiast nie każde miary równoważne będą takie w sensie relacji  $Eq$ . Dzięki stwierdzeniu 3.8 badanie równoważności miar ze względu na porządek reguł można sprowadzić do wyznaczenia klas abstrakcji relacji  $Eq$ .

Równoważność miar ze względu na porządek reguł pozwala wnioskować o monotoniczności jednej z miar na podstawie monotoniczności drugiej.

**Stwierdzenie 3.9.** Niech dane są dwie miary jakości,  $q_1, q_2 \in M$ , równoważne ze względu na porządek reguł. Jeśli jedna z tych miar charakteryzuje się jakimkolwiek rodzajem monotoniczności prezentowanym w tablicy 3.3, to druga miara jest monotoniczna w ten sam sposób.

Dowód: Dowód stwierdzenia wynika z definicji 3.4, określającej równoważność miar, oraz z dowodu Tw.2.6, przytoczonego w artykule Fürnkranza i Flacha [90]. Obecnie przytoczymy podobny dowód, rozszerzając rozumowanie również na monotoniczności M1–M4,  $M_{(P-p)P}$  oraz  $M_{(N-n)N}$ . Z definicji kompatybilności wynika, że miary są kompatybilne wtedy i tylko wtedy, gdy  $\forall_{RUL, RUL_X \subseteq RUL} \forall_{r_1, r_2 \in RUL_X} (q_1(r_1) < q_1(r_2) \Leftrightarrow q_2(r_1) < q_2(r_2))$ . Ponadto, definicja

kompatybilności nie wymaga, aby reguły należące do zbioru RUL były wyznaczone na podstawie identycznego zbioru przykładów. Dla każdej reguły decyzyjnej i dowolnego

zbioru przykładów można zdefiniować tablicę kontyngencji, charakteryzującą regułę w rozważanym zbiorze przykładów. W tabeli 3.3 mamy do czynienia z funkcjami rosnącymi (malejącymi) oraz niemalejącymi (nierośnającymi).

Rozważmy funkcje rosnące (malejące). Założymy, że miara  $q_1$  jest monotoniczna, czyli przy nałożeniu odpowiednich warunków na  $p_i, n_i, P_i, N_i$ ,  $i \in \{1,2\}$  spełniony jest warunek  $q_1(p_1, n_1, P_1, N_1) < (>) q_1(p_2, n_2, P_2, N_2)$ . W ten sposób otrzymamy wszystkie przypadki funkcji rosnących (malejących), rozważane w tabeli 3.3.

Jeśli miary  $q_1$  i  $q_2$  są kompatybilne, to na podstawie kompatybilności otrzymamy również  $q_2(p_1, n_1, P_1, N_1) < (>) q_2(p_2, n_2, P_2, N_2)$ , co oznacza, że miara  $q_2$  jest monotoniczna w taki sam sposób jak miara  $q_1$ . Podobnie można udowodnić, że z monotoniczności  $q_2$  wynika monotoniczność  $q_1$ .

Dla monotoniczności związanej z funkcjami niemalejącymi (nierośnającymi) miara  $q_1$  spełnia jeden z warunków  $q_1(p_1, n_1, P_1, N_1) \leq (\geq) q_1(p_2, n_2, P_2, N_2)$ . Ostrą nierówność już rozpatrywaliśmy, zatem do zakończenia dowodu wystarczy udowodnić implikację: jeśli  $q_1(p_1, n_1, P_1, N_1) = q_1(p_2, n_2, P_2, N_2)$ , to  $q_2(p_1, n_1, P_1, N_1) = q_2(p_2, n_2, P_2, N_2)$ . Założymy, że implikacja ta nie jest prawdziwa; przy tym założeniu jej konkluzja przyjmuje postać  $q_2(p_1, n_1, P_1, N_1) \neq q_2(p_2, n_2, P_2, N_2)$ . Zatem  $q_2(p_1, n_1, P_1, N_1) > q_2(p_2, n_2, P_2, N_2)$  albo  $q_2(p_1, n_1, P_1, N_1) < q_2(p_2, n_2, P_2, N_2)$ . W świetle przyjętego założenia, że miary  $q_1, q_2$  są kompatybilne, oznacza to, że  $q_1(p_1, n_1, P_1, N_1) < q_1(p_2, n_2, P_2, N_2)$  albo  $q_1(p_1, n_1, P_1, N_1) > q_1(p_2, n_2, P_2, N_2)$ , co jest sprzeczne z założeniem, że  $q_1(p_1, n_1, P_1, N_1) = q_1(p_2, n_2, P_2, N_2)$ . W ten sposób otrzymaliśmy potwierdzenie, że na podstawie monotoniczności miary  $q_1$  możemy wnioskować o identycznym typie monotoniczności miary  $q_2$ . Podobne rozumowanie można przeprowadzić, wnioskując o monotoniczności  $q_1$  na podstawie monotoniczności  $q_2$ . ◆

**Uwaga 3.3.** Niech dane są dwie antagonistyczne miary jakości  $q_1, q_2 \in M$ . Jeśli jedna z tych miar charakteryzuje się wybranym rodzajem monotoniczności, prezentowanym w tabeli 3.3, to druga miara nie charakteryzuje się tym typem monotoniczności.

Dowód uwagi 3.2 jest analogiczny do dowodów stwierdzenia 3.9, zamiast kompatybilności należy wykorzystać fakt, że miary są antagonistyczne.

**Uwaga 3.4.** Jeśli miara  $q \in M$  charakteryzuje się własnościami  $D_{in}$  i  $M_p$ , to ma również własności  $D_>$  i  $D_<$ .

Dowód: udowodnijmy, że reguła charakteryzuje się własnością  $D_>$ . Dla każdej reguły  $r_>$ , pokrywającej  $p_>$  przykładów pozytywnych i  $n_>$  negatywnych, takich

że  $(p_>/(p_> + n_>)) > (P/(P+N))$ , musi istnieć reguła  $r$ , pokrywająca tyle samo przykładów negatywnych  $n_>$  oraz mniej niż  $p_>$  przykładów pozytywnych, taka że  $(p/(p+n_>)) = (P/(P+N))$ . Zgodnie z definicją monotoniczności  $M_p$  oznacza to, że  $q(r_>) > q(r) = const_1$ . Podobne rozumowanie można przeprowadzić, dowodząc własności  $D_<$  dla reguły  $r_<$ , spełniającej warunek  $(p_</(p_< + n_<)) < (P/(P+N))$ .

**Uwaga 3.5.** Jeśli miary  $q_1, q_2 \in M$  są równoważne ze względu na porządek reguł oraz miara  $q_1$  charakteryzuje się własnościami  $D_{in}, D_>, D_<, D_{eq}$ , to miara  $q_2$  również charakteryzuje się własnościami  $D_{in}, D_>, D_<, D_{eq}$ .

Dowód uwagi wynika wprost z definicji równoważności ze względu na porządek reguł (kompatybilności) oraz stwierdzenia 3.9.

Zastosowanie miar identycznych lub równoważnych w pokryciowych algorytmach indukcji reguł skutkuje wyznaczeniem identycznych zbiorów reguł. Zakładamy tutaj, że miary stosowane są w fazach wzrostu i przycinania reguły, a wszystkie pozostałe parametry programu są identyczne. W szczególności kryterium stopu, decydujące o zakończeniu procesu indukcji, jest dostosowane do wartości, jakie przyjmuje każda z miar. Przykładowo, jeśli miary  $q_1$  i  $q_2$  są równoważne ze względu na porządek reguł oraz algorytm używa miary  $q_1$  i dokonuje indukcji reguł o jakości  $q_1(r) \geq v1$ , to algorytm stosujący  $q_2$  dokonuje indukcji reguł o jakości  $q_2(r) \geq v2$ . Wartość  $v2$  dobrana jest w taki sposób, aby spełniona była zależność  $\forall r \ q_1(r) \geq v1 \Leftrightarrow q_2(r) \geq v2$ .

Zauważmy, że klasyfikatory złożone z identycznych zbiorów reguł, ale stosujące różne schematy klasyfikacji, mogą się charakteryzować różnymi zdolnościami klasyfikacji. Zastosowanie miar równoważnych w procedurze rozstrzygania konfliktów klasyfikacji będzie prowadziło do identycznych wyników jedynie w specyficznych sytuacjach. Na podstawie pewnego zbioru reguł utwórzmy dwie listy:  $L_1$  i  $L_2$ . Pozycja reguły  $r$  na liście  $L_1$  uzależniona jest od jej oceny, uzyskanej na podstawie miary  $q_1$ . Pozycja reguły  $r$  na liście  $L_2$  uzależniona jest od jej oceny, uzyskanej na podstawie miary  $q_2$ . Jeśli  $q_1$  i  $q_2$  są identyczne lub równoważne ze względu na porządek reguł, to otrzymamy identyczne listy, a wyniki klasyfikacji realizowanej za ich pomocą będą również identyczne (przyjmujemy dodatkowe założenie, że jeśli w zbiorze wystąpią reguły o identycznej ocenie, to ich kolejność na obu listach będzie taka sama). Z taką sytuacją mamy do czynienia w algorytmie RIPPER, stosowanym m.in. przez Fürnkranza. Zapewne z tego też powodu Fürnkranz bada [90,143] jedynie identyczność i równoważność miar ze względu na porządek reguł. Można zauważać, że dla innych schematów klasyfikacji użycie miar równoważnych ze względu na porządek reguł może prowadzić do różnych wyników. Zanim zajmiemy się równoważnością

miar ze względu na sposób rozstrzygania konfliktów klasyfikacji, poświęcimy nieco miejsca problemowi badania podobieństwa pomiędzy porządkami reguł, ustanawianymi przez różne miary jakości.

Analizując różne zbiory danych, można zauważyc, że poza miarami identycznymi i równoważnymi istnieje całkiem spora grupa miar, która porządkuje zbiory reguł w bardzo podobny sposób. Autor badał wiele miar jakości, wykazując, że dla pewnej grupy miar różnice w porządkach reguł otrzymanych na ich podstawie dotyczą często pojedynczych reguł. Wskazanie miar porządkujących reguły w podobny sposób pozwoli na określenie grup miar podobnych oraz wybór w obrębie każdej grupy tych miar, które charakteryzują się najbardziej interesującymi własnościami. Poniżej w sposób formalny przedstawiono definicje porządku i reprezentacji porządku reguł.

Oceniając reguły należące do zbioru  $RUL$  za pomocą miary  $q_1$ , możemy uporządkować ten zbiór ze względu na nierosnącą (miary korzyści) lub niemalejącą (miary kosztu) wartość miary. W ten sposób otrzymujemy porządek, na czele którego znajduje się reguła(-y) najlepsza(-e), a na końcu – reguła(-y) najgorsza(-e). Ponieważ relacje  $\leq$  i  $\geq$  są relacjami słabego częściowego porządku, w ocenianym zbiorze mogą znaleźć się reguły charakteryzowane przez identyczną wartość miary. Reguły takie są między sobą nierozróżnialne, a przez to – nieporównywane. Aby móc je porównać, należy wprowadzić dodatkowe kryterium oceniające, w szczególności może to być inna miara jakości. Oczywiście zastosowanie kolejnego kryterium jakości  $q_2$ , odróżniającego reguły nieodróżnialne ze względu na  $q_1$ , w dalszym ciągu nie gwarantuje, że możliwe będzie porównanie każdych dwóch reguł. Jako przykład wystarczy podać dwie różne reguły pokrywające te same zbiory przykładów pozytywnych i negatywnych. Za pomocą miar definiowanych na podstawie tablicy kontyngencji nie da się tych reguł odróżnić.

**Definicja 3.6.** Niech dane są: zbiór reguł decyzyjnych  $RUL$  oraz zbiór miar  $M$ . Relacje zdefiniowaną w zbiorze  $RUL^2$  jako  $IND(RUL, B) = \{ \langle r_i, r_j \rangle \in RUL \times RUL : \forall q \in B \quad q(r_i) = q(r_j) \}$ , nazywamy relacją nierozróżnialności reguł należących do zbioru  $RUL$  ze względu na wartości miar należących do zbioru  $B \subseteq M$ .

Relacja  $IND(RUL, B)$  jest relacją równoważności. W dowodzie tej własności można wprost skorzystać z faktu, że relacja  $IND(RUL, B)$  definiowana jest w podobny sposób jak relacja nierozróżnialności w teorii zbiorów przybliżonych [220]. Do klas abstrakcji relacji  $IND(RUL, B)$  należą wszystkie reguły nierozróżnialne ze względu na oceny uzyskane przez miary należące do zbioru  $B$ .

**Definicja 3.7.** Niech dane są: zbiór reguł decyzyjnych  $RUL_X$ , zawierający reguły o konkluzjach wskazujących na klasę decyzyjną  $X$ , oraz zbiór miar  $B = \{q_1, q_2, \dots, q_k\}$ , na podstawie którego utworzono ciąg  $\langle q_{i1}, q_{i2}, \dots, q_{ik} \rangle$ . Ciąg klas abstrakcji  $\langle \underbrace{[ ]^{q_{i1}, q_{i2}, \dots, q_{ik}}}_{IND(RUL_X, B)}, \underbrace{[ ]^{q_{i1}, q_{i2}, \dots, q_{ik}}}_{IND(RUL_X, B)}, \dots, \underbrace{[ ]^{q_{i1}, q_{i2}, \dots, q_{ik}}}_{IND(RUL_X, B)} \rangle$  relacji  $IND(RUL_X, B)$  o własności (3.1):

$$\forall \underset{\substack{n < m \\ n, m \in \{1, 2, \dots, J\}}}{\bigwedge} \underset{s \in \underset{m}{\bigcup} \{ [ ]^{q_{i1}, q_{i2}, \dots, q_{ik}} \}_{IND(RUL_X, B)}}{\bigwedge} \exists \underset{z \in \{1, 2, \dots, k\}}{\bigwedge} \left( \left( \forall q_{il}(r) = q_{il}(s) \right) \wedge (q_{iz}(r) > q_{iz}(s)) \right) \quad (3.1)$$

nazywamy porządkiem reguł wskazujących na klasę decyzyjną  $X$ , utworzonym na podstawie ciągu miar  $\langle q_{i1}, q_{i2}, \dots, q_{ik} \rangle$ . Dodatkowo, liczbę  $i$ , identyfikującą kolejność klasy abstrakcji w porządku reguł, nazywać będziemy numerem klasy abstrakcji.

Porządek reguł wskazujących na klasę decyzyjną  $X$  ze względu na ciąg miar  $c = \langle q_{i1}, q_{i2}, \dots, q_{ik} \rangle$  oznaczany będzie przez  $ord_{RUL_X}^c$ .

Kolejność klas abstrakcji w porządku uzależniona jest od kolejności miar tworzących ciąg  $\langle q_{i1}, q_{i2}, \dots, q_{ik} \rangle$ . Jeśli  $c = \langle q \rangle$ , to kolejne klasy abstrakcji, będące składnikami porządku  $ord_{RUL_X}^c$ , złożone są z reguł charakteryzowanych przez coraz mniejsze wartości miary  $q$ .

**Definicja 3.8.** Niech dane są: zbiory reguł decyzyjnych  $RUL_{Xj}$ , zawierające reguły wskazujące na klasy decyzyjne  $X_j$ ,  $j = 1, 2, \dots, v$ , oraz ciąg miar jakości  $c = \langle q_{i1}, q_{i2}, \dots, q_{ik} \rangle$ . Porządkiem reguł należących do zbioru  $RUL$  ze względu na ciąg miar  $c$  nazywamy konkatenację porządków  $ord_{RUL_{X1}}^c, ord_{RUL_{X2}}^c, \dots, ord_{RUL_{Xv}}^c$ .

**Przykład 3.2.** Niech dane są zbiory  $RUL = \{r_1, r_2, r_3, r_4\}$  oraz  $B = \{q_1, q_2\}$ . Ponadto w zbiorze  $RUL$  znajdują się reguły o identycznej konkluzji. Wiadomo także, że  $q_1(r_1) = 0.1$ ,  $q_1(r_2) = 0.2$ ,  $q_1(r_3) = 0.3$ ,  $q_1(r_4) = 0.4$ ,  $q_2(r_1) = 0.1$ ,  $q_2(r_2) = 0.1$ ,  $q_2(r_3) = 0.05$ ,  $q_2(r_4) = 0.4$ . Klasy abstrakcji relacji  $IND(RUL, \{q_1\})$  są jednoelementowe, gdyż wartości miary  $q_1$  dla wszystkich reguł ze zbioru  $RUL$  są różne. Porządek reguł ze względu na miarę  $q_1$  będzie miał zatem postać  $\langle \{r_4\}, \{r_3\}, \{r_2\}, \{r_1\} \rangle$ . Dla miary  $q_2$  porządek reguł będzie miał postać  $\langle \{r_4\}, \{r_1, r_2\}, \{r_3\} \rangle$ , gdyż reguły  $r_1, r_2$  są nieroróżnicalne ze względu na wartość miary  $q_2$ . Porządek reguł ze względu na ciąg miar  $\langle q_1, q_2 \rangle$  ma postać  $\langle \{r_4\}, \{r_3\}, \{r_2\}, \{r_1\} \rangle$ , natomiast porządek reguł ze względu na ciąg miar  $\langle q_2, q_1 \rangle$  ma postać  $\langle \{r_4\}, \{r_2\}, \{r_1\}, \{r_3\} \rangle$ .

Przedstawiona definicja porządku reguł zakłada, że powstaje on z połączenia porządków ustalonych oddziennie w obrębie zbiorów reguł wskazujących na kolejne klasy decyzyjne. Kolejność (numeracja) klas decyzyjnych jest dowolna, ważne jest, aby podczas porównywania porządków otrzymanych na podstawie różnych ciągów miar kolejność klas była identyczna.

Zauważmy, że nie ma przeszkód, aby zgodnie z ideą definicji 3.7 określić porządek, który nie bierze pod uwagę przynależności reguł do klas decyzyjnych.

Im mniejsza będzie liczba par klas abstrakcji, które przez ciągi miar  $c_1$  i  $c_2$  oceniane są w sposób antagonistyczny, tym większe będzie podobieństwo porządków reguł generowanych przez  $c_1$  i  $c_2$ . Tak rozumiane podobieństwo musi być badane w kontekście konkretnego zbioru reguł.

**Definicja 3.9.** Niech dany jest zbiór reguł decyzyjnych  $RUL_X$ , wskazujących na klasę decyzyjną  $X$ . Reguły z tego zbioru ułożono w pewnej dowolnej kolejności, uzyskując ciąg  $\langle r_1, r_2, \dots, r_n \rangle$ . Niech dany jest ciąg miar jakości  $c$ . Na podstawie  $RUL_X$  i ciągu  $c$  otrzymano porządek  $ord_{RUL_X}^c$ . Ciąg liczb  $rep_{RUL_X}^c$ , złożony z  $n$  liczb naturalnych, w którym wartość  $i$ -tego wyrazu jest równa numerowi klasy abstrakcji w porządku  $ord_{RUL_X}^c$ , do której należy reguła  $r_i$ , nazywać będziemy reprezentacją porządku reguł  $ord_{RUL_X}^c$ .

Reprezentacja porządku zbioru reguł, złożonego z opisów kilku klas decyzyjnych ( $rep_{RUL}^c$ ), jest konkatenacją reprezentacji porządków utworzonych dla każdej klasy decyzyjnej  $rep_{RUL_{Xi}}^c$ . Aby zachować możliwość badania podobieństwa reprezentacji porządków  $rep_{RUL}^{c1}$  i  $rep_{RUL}^{c2}$ , zakłada się, że reprezentacje te dołączane są do  $rep_{RUL}^{c1}$  i  $rep_{RUL}^{c2}$  w tej samej kolejności. Nie ma znaczenia kolejność, w jakiej ustawione będą reguły należące do  $RUL_{Xi}$ , ważne jest, aby podczas tworzenia reprezentacji  $rep_{RUL_{Xi}}^{c1}$  i  $rep_{RUL_{Xi}}^{c2}$  kolejność ta nie zmieniała się. Wprowadzenie pojęcia reprezentacji porządku ostatecznie prowadzi do podania następującej definicji podobieństwa ciągów miar ze względu na zbiór reguł  $RUL$ .

**Definicja 3.10.** Niech dane są reprezentacje  $rep_{RUL}^{c1}$ ,  $rep_{RUL}^{c2}$  porządków  $ord_{RUL}^{c1}$ ,  $ord_{RUL}^{c2}$ . Przez  $S(c_1, c_2, RUL)$  oznaczymy wartość współczynnika mierzącego podobieństwo pomiędzy reprezentacjami  $rep_{RUL}^{c1}$ ,  $rep_{RUL}^{c2}$ . Jeżeli  $S(c_1, c_2, RUL) > th$ , to ciągi miar  $c_1$  i  $c_2$  są podobne w stopniu  $th$  ze względu na miarę podobieństwa  $S$  oraz sposób uporządkowania reguł należących do zbioru  $RUL$ .

**Przykład 3.3.** Przykład jest kontynuacją przykładu 3.2. Zakładając, że reguły z przykładu 3.2 ustawiono w porządku  $\langle r_1, r_2, r_3, r_4 \rangle$ , otrzymamy następujące reprezentacje:  $rep_{RUL}^{q_2} = \langle 2, 2, 3, 1 \rangle$ ,  $rep_{RUL}^{\langle q_1, q_2 \rangle} = \langle 4, 3, 2, 1 \rangle$ .

W definicji 3.10 do mierzenia podobieństwa reprezentacji porządków reguł można użyć współczynnika korelacji  $\tau$  Kendala, współczynnika korelacji rang R Spearmana lub dowolnej innej miary, przeznaczonej do określania podobieństwa pomiędzy ciągami liczb lub rang. Zastosowana miara powinna osiągać maksymalną wartość, jeśli reprezentacje porządków będą identyczne, oraz wartość minimalną, jeśli ciągi miar  $c_1$  i  $c_2$  porządkują reguły w sposób antagonistyczny.

Z definicji 3.10 wynika, że o podobieństwie miar można mówić jedynie w kontekście ustalonej miary podobieństwa i ustalonego zbioru reguł. Poniżej zdefiniowano podobieństwo ze względu na rodzinę zbiorów reguł.

**Definicja 3.11.** Niech dana jest rodzina zbiorów reguł  $\mathfrak{R}$ . Dla każdego zbioru  $RUL \in \mathfrak{R}$  dane są reprezentacje  $rep_{RUL}^{c_1}$ ,  $rep_{RUL}^{c_2}$  porządków  $ord_{RUL}^{c_1}$ ,  $ord_{RUL}^{c_2}$ . Jeżeli dla każdego  $RUL \in \mathfrak{R}$ ,  $S(c_1, c_2, RUL) > th$ , to ciągi miar  $c_1$  i  $c_2$  są podobne w stopniu  $th$  ze względu na miarę podobieństwa  $S$  oraz sposób uporządkowania reguł należących do zbiorów tworzących rodzinę  $\mathfrak{R}$ .

### 3.1.5. Równoważność ze względu na rozstrzyganie konfliktów klasyfikacji

Do tej pory rozważano równoważność i podobieństwo miar ze względu na porządek reguł, obecnie rozważana będzie równoważność miar ze względu na sposób rozstrzygania konfliktów klasyfikacji. W rozdziale 2 omówione zostały trzy schematy klasyfikacji. W klasyfikatorach rozstrzygających konflikty według zasady największego zaufania lub przez głosowanie (które w naszym przypadku utożsamiane jest z ważonym głosowaniem za pomocą wartości miary) użycie miar równoważnych ze względu na porządek reguł nie musi prowadzić do identycznych wyników.

**Przykład 3.4.** Dane są trzy reguły:  $r_1, r_2, r_3$ . Reguła  $r_3$  należy do opisu klasy decyzyjnej  $X_1$ , a pozostałe reguły są częścią opisu klasy  $X_2$ . Dane są dwie kompatybilne miary jakości:  $q_1, q_2$  oraz  $q_1(r_1) = 0.2$ ,  $q_1(r_2) = 0.3$ ,  $q_1(r_3) = 0.4$ ,  $q_2(r_1) = 1$ ,  $q_2(r_2) = 2$ ,  $q_2(r_3) = 2.5$ . Niech przykład testowy  $u$  pokrywa wszystkie trzy reguły. Użycie miary  $q_1$  do obliczenia siły głosu reguł spowoduje przyporządkowanie  $u$  do klasy  $X_1$ , użycie miary  $q_2$  zmieni to przyporządkowanie na  $X_2$ .

**Definicja 3.12.** Dowolne miary  $q_1, q_2 \in M$  są równoważne ze względu na sposób rozstrzygania konfliktów klasyfikacji w schemacie największego zaufania, jeżeli dla dowolnego zbioru reguł decyzyjnych  $RUL$  spełniony jest następujący warunek:

$$\forall_{rul \subseteq RUL} \exists_{r_i} q_1(r_i) = \max \{q_1(r_j) : r_j \in rul\} \Leftrightarrow q_2(r_i) = \max \{q_2(r_j) : r_j \in rul\}.$$

**Definicja 3.13.** Dowolne miary  $q_1, q_2 \in M$  są równoważne ze względu na sposób rozstrzygania konfliktów klasyfikacji przez głosowanie, jeżeli dla dowolnego zbioru reguł decyzyjnych  $RUL$  spełniony jest następujący warunek:

$$\forall_{rul1, rul2 \subseteq RUL} \sum_{r \in rul1} q_1(r) > \sum_{r \in rul2} q_1(r) \Leftrightarrow \sum_{r \in rul1} q_2(r) > \sum_{r \in rul2} q_2(r).$$

**Uwaga 3.6.** Dowolne miary jakości  $q_1, q_2 \in M$ , równoważne w sposób taki jak określa to definicja 3.12 lub 3.13, są bezwzględnie równoważne ze względu na porządek reguł.

Chcąc przeprowadzić dowód uwagi 3.6, wystarczy zauważyć, że ograniczając w definicji 3.12 zbiór  $rul$  do dwóch reguł, otrzymamy definicję kompatybilności miar ze względu na porządek reguł, z której usunięto założenie, że reguły wskazują na identyczną klasę decyzyjną. Ograniczając zbiór  $rul$  do dwóch reguł wskazujących na identyczną klasę decyzyjną, otrzymamy definicję kompatybilności miar ze względu na klasę decyzyjną. W definicji 3.13, rozważając jednoelementowe zbiory  $rul1$  i  $rul2$ , otrzymamy definicję kompatybilności miar ze względu na porządek reguł, z której usunięto założenie, że reguły wskazują na identyczną klasę decyzyjną. Ograniczając się do reguł wskazujących na identyczną klasę decyzyjną, uzyskamy definicję kompatybilności miar ze względu na klasę decyzyjną. Zgodnie z definicją 3.4 kompatybilność miar jest równoznaczna z ich równoważnością ze względu na porządek reguł.

**Uwaga 3.7.** Dowolne miary bezwzględnie równoważne ze względu na porządek reguł są równoważne ze względu na sposób rozstrzygania konfliktów klasyfikacji w schemacie największego zaufania.

Dowód: założymy, że miary  $q_1, q_2$  są bezwzględnie równoważne, a więc kompatybilne. Jeżeli  $\forall rul q_1(r_i) = \max \{q_1(r_j) : r_j \in rul\}$ , to  $\forall r_j \in rul$ , takiej że  $q_1(r_i) > q_1(r_j)$ ; prawdą jest również to, że  $q_2(r_i) > q_2(r_j)$ , a więc  $\forall rul q_2(r_i) = \max \{q_2(r_j) : r_j \in rul\}$ . Podobne rozumowanie można przedstawić, rozpoczynając dowód od miary  $q_2$ .

**Stwierdzenie 3.10.** Niech dane są miary  $q_1, q_2 \in M$ . Jeżeli istnieje funkcja rosnąca  $f$ , taka że dla dowolnej reguły  $r$   $f(q_1(r)) = q_2(r)$ , to miary  $q_1, q_2$  są równoważne ze względu na sposób rozstrzygania konfliktów klasyfikacji w schemacie największego zaufania.

Dowód: do wykazania równoważności miar konieczne jest pokazanie, że miary są równoważne tak, jak określa to definicja 3.12. Najpierw udowodnimy implikację ( $\Rightarrow$ ). Niech dla dowolnego  $rul \subseteq RUL$  (w szczególności zawierającego reguły wskazujące na identyczną klasę decyzyjną)  $q_1(r) = \max\{q_1(r_j) : r_j \in rul\}$ . Oznacza to, że dla każdej reguły  $r_j \in rul$   $q_1(r) \geq q_1(r_j)$ . Na podstawie założenia, że funkcja  $f$  jest rosnąca, otrzymamy, że dla każdego  $r_j \in rul$   $f(q_1(r)) \geq f(q_1(r_j))$ , a zatem  $q_2(r) \geq q_2(r_j)$ , czyli  $q_2(r) = \max\{q_2(r_j) : r_j \in rul\}$ . Ponieważ funkcja  $f$  jest rosnąca, funkcja  $f^{-1}$  również jest funkcją rosnącą, co pozwoli nam w analogiczny sposób dowieść implikacji ( $\Leftarrow$ ). ◆

**Stwierdzenie 3.11.** Niech dane są miary  $q_1, q_2 \in M$ . Jeżeli istnieje liczba rzeczywista dodatnia  $a$ , taka że dla dowolnej reguły  $r$   $q_1(r) = aq_2(r)$ , to miary  $q_1, q_2$  są równoważne ze względu na sposób rozstrzygania konfliktów klasyfikacji przez głosowanie.

Dowód: do wykazania równoważności miar konieczne jest przeprowadzenie dowodu równoważności, zawartego w definicji 3.13. Najpierw udowodnimy implikację ( $\Rightarrow$ ). Niech dla dowolnych podzbiorów  $rul1, rul2$ , złożonych odpowiednio z  $l$  i  $s$  reguł, prawdziwa jest nierówność  $q_1(r_{i1}) + q_1(r_{i2}) + \dots + q_1(r_{il}) > q_1(r_{j1}) + q_1(r_{j2}) + \dots + q_1(r_{js})$ .

Wykorzystując założenie dowodzonego stwierdzenia, nierówność tę można zapisać jako  $a(q_2(r_{i1}) + q_2(r_{i2}) + \dots + q_2(r_{il})) > a(q_2(r_{j1}) + q_2(r_{j2}) + \dots + q_2(r_{js}))$ . Ponieważ stała  $a$  jest liczbą rzeczywistą dodatnią, nierówność  $q_2(r_{i1}) + q_2(r_{i2}) + \dots + q_2(r_{il}) > q_2(r_{j1}) + q_2(r_{j2}) + \dots + q_2(r_{js})$  jest również prawdziwa. Dowód implikacji ( $\Leftarrow$ ) można przeprowadzić w sposób analogiczny. ◆

Badanie podobieństwa miar ze względu na sposób rozstrzygania konfliktów klasyfikacji musi być rozważane, podobnie jak podobieństwo ze względu na porządek reguł, w kontekście konkretnych zbiorów danych i konkretnej miary ocenającej efektywność klasyfikatora. Dwie miary będą podobne ze względu na sposób rozstrzygania konfliktów klasyfikacji, jeśli dla pewnej rodziny zbiorów różnice w klasyfikacji tych zbiorów będą mniejsze od zadanego progu lub będą statystycznie nieistotne.

Zauważmy, że zastosowanie w algorytmie pokryciowym, w fazach wzrostu i przycinania, miar równoważnych ze względu na sposób rozstrzygania konfliktów klasyfikacji gwarantuje otrzymanie identycznych zbiorów reguł, o identycznych zdolnościach klasyfikacyjnych (wynika to z uwagi 3.6). Podobieństwo miar ze względu na porządek reguł nie musi przekładać się na podobieństwo w sposobie rozstrzygania konfliktów klasyfikacji. Jednak podobieństwo w sposobie rozstrzygania konfliktów klasyfikacji nie musi przekładać się na podobieństwo w porządkowaniu zbioru reguł.

W rozdziale 4 badane będzie podobieństwo klasyfikatorów regułowych, otrzymanych przez użycie różnych miar jakości. Przedstawione zostaną również wyniki badania podobieństwa miar ze względu na porządek reguł.

### ***3.1.6. Wybrane miary jakości***

W literaturze przedmiotu, traktującej o miarach jakości reguł decyzyjnych i miarach oceny atrakcyjności reguł, można spotkać ponad 50 obiektywnych miar definiowanych na podstawie tablicy kontyngencji. Miary te proponowane były na przestrzeni ostatnich lat przez różnych autorów. Miary związane stricte z oceną reguł decyzyjnych zebrano w pracach [40, 42, 90, 143, 176, 264, 341]. Prace Bruhy [40, 42] koncentrują się głównie na badaniu wpływu miar na dokładność klasyfikatora regułowego, w szczególności na stosowaniu miar w procesie głosowania. Fürnkranz przedstawia wyniki badań związane z użyciem miar w procesie wzrostu i przycinania reguł [90, 143], dodatkowo analizując równoważność wybranych miar ze względu na porządek reguł [90]. Prace Yao i Lavrac koncentrują się na opisaniu miar w języku prawdopodobieństwa [176, 341]. Interesującym spostrzeżeniem jest to, że badania prowadzone przez wymienionych autorów w dużej części dotyczą rozłącznych zbiorów miar.

Miary przeznaczone do oceny atrakcyjności reguł asocjacyjnych analizowano m.in. w pracach przeglądowych [128, 132, 179, 215, 307]. Analiza tych publikacji pozwala na wyciągnięcie wniosku, że duża grupa miar, stosowana do oceny atrakcyjności reguł, używana jest również do oceny jakości reguł decyzyjnych (w tym do nadzorowania procesu indukcji).

Autor w swoich pracach badał efektywność większości miar analizowanych przez Fürnkranza, Bruhę i Yao. Efektywność miar badano w kontekście dokładności klasyfikacji i zdolności opisowych otrzymanych na ich podstawie zbiorów reguł. Reguły wyznaczano za pomocą algorytmów q-ModLEM i RMatrix [246, 248, 253, 254, 261, 264].

Przez długi czas starano się zdefiniować najskuteczniejszą miarę jakości, która zastosowana w pokryciowym algorytmie indukcji prowadziłaby do otrzymania najlepszych wyników, niezależnie od analizowanego zbioru danych. Głębsza analiza tego zagadnienia wymaga ustalenia tego, co oznacza sformułowanie „otrzymania najlepszych wyników”. Zakładając, że ostateczna ocena dotyczy nie tylko pojedynczych reguł, ale również całego ich zbioru, kryteria jakości mogą być różne. W indukcji reguł dla celów klasyfikacji podstawowymi kryteriami będą dokładność klasyfikacji oraz średnia dokładność klas decyzyjnych (lub inne kryteria, o których wspominano w rozdziale 2). Gdy ważny jest również aspekt opisowy, znaczenie będą miały liczba i budowa wyznaczonych reguł. Jak wykazują badania eksperymentalne, w tym również prowadzone przez autora [8, 42, 144, 246, 248, 253, 254], nie można wskazać jednej miary jakości, która jest najlepsza niezależnie

od charakterystyki analizowanego zbioru danych i zdefiniowanego kryterium jakości zbioru reguł.

W tabeli 3.5 przedstawiono 35 obiektywnych miar jakości, definiowanych na podstawie tablicy kontyngencji. Pośród prezentowanych miar większość z nich stosowana jest do oceny reguł decyzyjnych; jest tam także pewna liczba miar funkcjonujących dotychczas jedynie jako miary oceniające atrakcyjność reguł.

W tabeli 3.5 nie zawarto oczywiście wszystkich możliwych miar. Byłoby to trudne, gdyż każdy algorytm indukcji może posługiwać się swoją własną, unikalną miarą jakości. Spośród znanych miar w tablicy 3.5 nie zawarto m.in. parametryzowanej miary *Laplace'a* oraz miar proponowanych przez: Brazdila i Torgo [37l] (miara *product*), Michalskiego [156, 189] (miara *weighted sum*) oraz Bratko i Kononenkę [165] (miara *information score*). W tabeli 3.5 nie ma również miar *costs* i *relative costs*, proponowanych przez Fürnkranza [90, 143], ani miar *Coleman* i  $\chi^2$  [40]. Decyzję o ich pominięciu podjęto arbitralnie, m.in. na podstawie analizy wyników dokładności klasyfikatorów regułowych, trenowanych na wielu benchmarkowych zbiorach danych. Zamiast parametryzowanej miary *Laplace'a*, w której wartość parametru wskazuje na liczbę klas decyzyjnych, w tabeli 3.5 umieszczono postacie tej miary: standardową (*Laplace*) i ważoną (*wLap*). Ponadto, w trakcie badań stosowano również uogólnienie parametryzowanej miary *Laplace'a*, jakim jest *m-estymata*. Miary *product*, *information score* i  $\chi^2$  osiągają niską dokładność klasyfikacji. Stosowanie miar *produkt* i *information score* skutkuje także wyznaczaniem bardzo dużej liczby reguł. Zamiast *weighted sum* wykorzystano miarę *YAILS*, w której (w przeciwieństwie do *weighted sum*) nie występują żadne parametry. Miara *Coleman* jest składnikiem miar *C1* i *C2*, które są udoskonalonymi przez Bruhé [40] uogólnieniami tej miary. Miary *cost* i *relative costs*, przy założeniu identycznego kosztu błędnej klasyfikacji przykładów pozytywnych i negatywnych, przyjmują postać znajdujących się w tabeli 3.5 miar *accuracy* i *WRA*.

Tabela 3.5  
Wybrane miary oceniające jakość reguł decyzyjnych

Nazwa	Wzór	Nazwa	Wzór
<i>accuracy</i>	$p - n$	<i>coverage</i>	$\frac{p}{P}$
<i>full coverage</i>	$\frac{p + n}{P + N}$	<i>precision</i>	$\frac{p}{p + n}$
<i>f-Bayesian confirmation (f)</i>	$\frac{pN - nP}{pN + nP}$	<i>Laplace</i>	$\frac{p + 1}{p + n + 2}$
<i>Weighted Laplace (wLap)</i>	$\frac{(p + 1)(P + N)}{(p + n + 2)P}$	<i>RIPPER</i>	$\frac{p - n}{p + n}$

<i>Lift (Measure of independence)</i>	$\frac{p(P+N)}{(p+n)P}$	<i>Dependency Factor Pawlak - (DF) Popper- (<math>\eta</math>)</i>	$\frac{p(P+N)-P(p+n)}{p(P+N)+P(p+n)}$
<i>g-measure (g)</i>	$\frac{p}{p+n+g}$	<i>One way support (OWS)</i>	$\frac{p}{p+n} \ln \left( \frac{p(P+N)}{(p+n)P} \right)$
<i>Logical sufficiency (LS)</i>	$\frac{pN}{nP}$	<i>Rule specificity and sensitivity (RSS)</i>	$\frac{p}{P} - \frac{n}{N}$
<i>Weighted relative accuracy (WRA)</i>	$\frac{p+n}{P+N} \left( \frac{p}{p+n} - \frac{P}{P+N} \right)$	<i>Novelty, Leverage (Novelty)</i>	$\frac{p}{P+N} - \left[ \frac{P(p+n)}{(P+N)^2} \right]$
<i>Rule interest (RI)</i>	$\frac{p(P+N)-(p+n)P}{P+N}$	<i>m-estimate (m)</i>	$\frac{p+m}{p+n+m} \frac{P}{P+N}$
<i>Cohen, Kappa <math>\kappa</math> (Cohen)</i>	$\frac{(P+N) \left( \frac{p}{p+n} \right) - P}{\left( \frac{P+N}{2} \right) \left( \frac{p+n+P}{p+n} \right) - P}$	<i>Mutual support (MS)</i>	$\frac{p}{n+P}$
<i>F-measure (F)</i>	$\frac{\left( \beta^2 + 1 \right) \left( \frac{p}{p+n} \right) \left( \frac{p}{P} \right)}{\beta^2 \left( \frac{p}{p+n} \right) + \frac{p}{P}}$	<i>CI</i>	$\left( \frac{Np-Pn}{N(p+n)} \right) \left( \frac{2+Cohen}{3} \right)$
<i>C2</i>	$\left( \frac{Np-Pn}{N(p+n)} \right) \left( \frac{P+p}{2P} \right)$	<i>Klösgen</i>	$\left( \frac{p+n}{P+N} \right)^\omega \left( \frac{p}{p+n} - \frac{P}{P+N} \right)$
<i>Correlation, <math>\phi</math>-coefficient (Corr)</i>	$\frac{pN-Pn}{\sqrt{PN(p+n)(P-p+N-n)}}$	<i>s-Bayesian confirmation (s)</i>	$\frac{p}{p+n} - \frac{P-p}{P-p+N-n}$
<i>YAILS</i>	$w_1 \frac{p}{p+n} + w_2 \frac{p}{P}$ $w_1 = 0.5 + 0.25q_{prec}(p,n)$ $w_2 = 0.5 - 0.25q_{prec}(p,n)$	<i>Odds ratio (Odds)</i>	$\frac{p(N-n)}{n(P-p)}$
<i>Relative risk (RR)</i>	$\frac{p}{p+n} \left( \frac{P+N-p-n}{P-p} \right)$	<i>Q2</i>	$\left( \frac{p}{P} - \frac{n}{N} \right) \left( 1 - \frac{n}{N} \right)$
<i>Two way support (TWS)</i>	$\frac{p}{P+N} \ln \left( \frac{p(P+N)}{(p+n)P} \right)$	<i>cFoil</i>	$p \left( \log_2 \left( \frac{p}{p+n} \right) - \log_2 \left( \frac{P}{P+N} \right) \right)$
<i>J-measure (J)</i>	$\frac{1}{P+N} \left( p \ln \left( \frac{p(P+N)}{(p+n)P} \right) + n \ln \left( \frac{n(P+N)}{(p+n)N} \right) \right)$	<i>CN2 Significance, G2-likelihood statistic (CN2)</i>	$2 \left( p \ln \left( \frac{p}{(p+n) \frac{P}{P+N}} \right) + n \ln \left( \frac{n}{(p+n) \frac{N}{P+N}} \right) \right)$
<i>Info gain (Gain)</i>	$Info(P,N) - Info_{pn}(P,N), \quad Info(P,N) = - \left[ \frac{P}{P+N} \log_2 \frac{P}{P+N} + \frac{N}{P+N} \log_2 \frac{N}{P+N} \right]$ $Info_{pn}(P,N) = \frac{p+n}{P+N} Info(p,n) + \frac{P+N-p-n}{P+N} Info(P-p,N-n)$		

Wzory analityczne, definiujące miary, podano w oryginalnej (w miarę możliwości nieuproszczonej) postaci, co w przypadku kilku miar pozwala pokazać, że łączą one oceny dokładności i pokrycia.

Geneza pochodzenia miar jakości jest różna. Część z nich ma charakter empiryczny i została zdefiniowana przez różnych autorów na podstawie ich doświadczenia (np. g, *YAILS*, *C1*, *C2*). Istnieje grupa miar mająca genezę w statystyce matematycznej, w szczególności w dwuwymiarowej analizie dyskretnej dającej podstawy teoretyczne do analizy tablic kontyngencji (np. *Cohen*, *Coleman*, *RSS*, *WRA*). Również teoria informacji, w szczególności zasada minimalnej długości opisu, jest źródłem miar jakości (np. *J*, *CN2*, *LS*). Miary jakości adaptowane są również z innych obszarów informatyki, przykładem może być wyszukiwanie informacji i wywodząca się z niego miara *F*.

Miary zawarte w tabeli 3.5 charakteryzują się różnymi własnościami teoretycznymi. Część z nich przebadano pod kątem posiadania własności P i T [128, 132, 307] lub B i M [43, 106, 264, 300]. O własnościach wybranych miar będzie mowa w rozdziale 4. W dodatku A zamieszczono trój- i dwuwymiarowe wykresy wszystkich miar. Na podstawie analizy wykresów można drogą kontrprzykładu wnioskować, które z miar nie są monotoniczne (własności  $M_p$  i  $M_n$ ). Analiza wykresów może również dostarczyć informacji o nieposiadaniu przez miarę własności  $D_{in}$  i  $D_{eq}$ . Jeśli na wykresie poziomie punkty  $(0,0)$  i  $(P,N)$  nie są połączone odcinkiem będącym poziomą wykresu wartości miary, to miara nie charakteryzuje się własnością  $D_{in}$ . Zakładając, że  $P > N$ , jeśli na wykresie poziomie punkty  $(0,0)$  i  $(P,P)$  nie są połączone odcinkiem, to miara nie ma własności  $D_{eq}$ .

Część z rozważanych miar jest sparametryzowana i wykorzystanie ich do oceny reguł wymaga ustalenia wartości parametrów. W naszych rozważaniach ograniczamy się do miar mających maksymalnie jeden parametr. W zależności od wartości parametru charakterystyka miary (w szczególności jej wykres) ulega zmianie. Dla określonych wartości parametrów i specyficznych proporcji pomiędzy liczbą przykładów pozytywnych i negatywnych miary zawierające parametr(-y) mogą stawać się coraz bardziej podobne do innych miar. Dla równomiernego rozkładu przykładów i  $m=2$  miary  $m$  i *Laplace* są identyczne. Dla wartości parametru  $\omega=1$  miara *Klösgen* staje się miarą *WRA*. W zależności od kierunku zmian parametru  $m$  wykres miary  $m$  staje się coraz bardziej podobny do miary *precision* ( $m \rightarrow 0$ ) lub do *WRA* ( $m \rightarrow \infty$ ). Miara *F* dla  $\beta \rightarrow 0$  zmierza w kierunku *precision*, a dla  $\beta \rightarrow \infty$  w kierunku *coverage*. Wreszcie miara *Klösgen* dla  $\omega \rightarrow \infty$  upodabnia się coraz bardziej do zamieszczonej w tabeli 3.5 *full coverage*, określającej całkowite pokrycie reguły.

Problemem doboru uniwersalnych i specyficznych (dopasowanych do zbioru danych) wartości parametrów zajmowali się Janssen i Fürnkranz [143]. Na podstawie analizy około 40 zbiorów danych zaproponowali wartości parametrów dla miar:  $m$ ,  $F$  oraz miary

zdefiniowanej przez Klösgena [162]. Wartości te wynoszą odpowiednio: 22.466, 0.5 i 0.4323. Przedstawione w dodatku A wykresy miar  $m$ ,  $F$ ,  $Klösgen$  narysowano właśnie dla tych wartości parametrów.

Część z prezentowanych miar wymaga niewielkiej modyfikacji przed zastosowaniem ich w algorytmie indukcji reguł. Dotyczy to miar:  $Corr$ ,  $s$ ,  $LS$ ,  $Odds$ ,  $RR$ ,  $J$ ,  $CN2$ . Dla pewnych szczególnych wartości  $p$  i  $n$  miary te są nieokreślone. W przypadku  $Corr$  i  $s$  dotyczy to sytuacji, w której reguła pokrywa wszystkie możliwe przykłady, czyli gdy  $P=p$  oraz  $N=n$ . Wartości miar  $LS$ ,  $J$ ,  $Odds$ ,  $RR$ ,  $CN2$  są nieokreślone dla reguł dokładnych, czyli gdy  $n=0$ . Wartości  $Odds$  i  $RR$  są również nieokreślone, jeśli oceniana reguła pokrywa wszystkie przykłady pozytywne, czyli gdy  $p=P$ . Aby móc przebadać efektywność miar  $Corr$ ,  $s$ ,  $LS$ ,  $Odds$ ,  $RR$ ,  $J$ ,  $CN2$ , wprowadzono do nich pewne modyfikacje, polegające albo na arbitralnym przypisaniu wartości miary, gdy jest ona nieokreślona, albo na niewielkiej modyfikacji wzoru definiującego miarę. I tak, wartości  $Corr$  i  $s$  dla reguł pokrywających wszystkie przykłady ustalono na 0. Miarę  $LS$  zmodyfikowano w taki sposób, że do mianownika dodano 1. We wzorach definiujących  $J$  i  $CN2$ , w liczniku drugiego składnika sumy, w miejsce  $n$  wstawiono  $(n+1)$ , podobną zmianę wprowadzono do mianownika miary  $Odds$ . Dodatkowo we wzorach definiujących  $Odds$  i  $RR$  wprowadzono modyfikację polegającą na zamianie  $(P-p)$  na  $(P+1-p)$ .

W rozdziale 4 zawarto rezultaty badań nad miarami zamieszczonymi w tablicy 3.5. Badania empiryczne wiążąły się z analizą podobieństwa rankingów reguł utworzonych za pomocą różnych miar oraz weryfikacją efektywności stosowania miar do nadzorowania procesu indukcji reguł. Efektywność rozważana była z punktu widzenia trzech kryteriów biorących pod uwagę zdolności klasyfikacyjne i opisowe wyznaczonych reguł. Ilustracją wyników empirycznych są: analiza własności miar oraz omówienie miar, które okazały się najbardziej efektywne.

### **3.1.7. Zastosowanie miar do parametryzacji modelu zbiorów przybliżonych**

Miarę jakości można użyć do pomiaru stopnia, w jakim zbiór przykładów spełniających przesłankę reguły  $[\varphi]$  zawiera się w klasie decyzyjnej, wskazywanej w konkluzji reguły  $[\psi]$ . Większość miar jakości w tym pomiarze bierze pod uwagę nie tylko stopień, w jakim  $[\varphi] \subseteq [\psi]$ , ale również stopień, w jakim  $[\psi] \subseteq [\varphi]$ .

W teorii zbiorów przybliżonych funkcjonuje pojęcie inkluzyji przybliżonej, określającej stopień zawierania się jednego zbioru w drugim. Standardowo w zbiorze przykładów  $U$  inkluzyja przybliżona  $\mu : 2^U \times 2^U \rightarrow [0,1]$  jest miarą w sensie matematycznym, charakteryzującą się następującymi własnościami [98, 227]:

$$\begin{aligned} \forall X, Y \subseteq U \quad (\mu(X, Y) = 1 \Leftrightarrow X \subseteq Y), \\ \forall X, Y, Z \subseteq U \quad (\mu(Y, Z) = 1 \Rightarrow \mu(X, Y) \leq \mu(X, Z)). \end{aligned}$$

Poza tymi własnościami istnieje jeszcze wiele innych postulowanych własności, m.in.:

$$\begin{aligned} \forall X, Y \subseteq U, X \neq \emptyset \quad (\mu(X, Y) = 0 \Leftrightarrow X \cap Y = \emptyset), \\ \forall X, Y \subseteq U, X \neq \emptyset \quad (\mu(X, Y) + \mu(X, U - Y) = 1). \end{aligned}$$

Standardowa inkluзja przybliżona jest miarą zdefiniowaną zgodnie ze wzorem (3.2):

$$\mu(X, Y) = \frac{|X \cap Y|}{|X|}. \quad (3.2)$$

Inkluzja przybliżona jest uogólnieniem przybliżonej funkcji przynależności, definiowanej dla przykładu  $x \in U$  i zbioru  $Y \subseteq U$  jako (3.3):

$$\mu_{Y, B}(x) = \frac{|I_B(x) \cap Y|}{|I_B(x)|}. \quad (3.3)$$

We wzorze (3.3)  $B \subseteq A$  jest zbiorem atrybutów, a  $I_B(x)$  jest zbiorem przykładów nieroróżnicjalnych lub podobnych do  $x$  ze względu na  $B$ . W standardowym modelu zbiorów przybliżonych do  $I_B(x)$  należą przykłady nieroróżnicjalne od  $x$  ze względu na zbiór atrybutów  $B$ . W tolerancyjnym modelu zbiorów przybliżonych do  $I_B(x)$  należą przykłady podobne do  $x$  ze względu na  $B$  [210, 271, 290, 292, 295].

Większość miar stosowanych do oceny reguł decyzyjnych nie ma cech inkluзji przybliżonej, gdyż nie spełniają one warunków (1)–(3). Zauważmy jednak pewne analogie. Jeśli przyjmiemy, że  $r \equiv \varphi \rightarrow \psi$ ,  $X = [\varphi]$ ,  $Y = [\psi]$  oraz  $\mu(X, Y)$ , zdefiniowana zgodnie ze wzorem (3.2), to  $\mu(X, Y) = \text{precision}(r)$ ;  $\mu(Y, X) = \text{coverage}(r)$ . Miary o własności inkluзji przybliżonej oceniają głównie dokładność lub niedokładność reguły (3.4):

$$e(X, Y) = \frac{|X \cap \neg Y|}{|\neg Y|}, \quad e^*(X, Y) = \frac{|X \cap \neg Y|}{|Y|}. \quad (3.4)$$

Miar (3.3) i (3.4) użyto w pokryciowym algorytmie indukcji reguł VC-DomLEM [30]. W algorytmie tym poszukuje się zbioru reguł, pokrywającego dolne przybliżenia klas decyzyjnych. Zauważmy, że żadna z miar  $\mu(X, Y)$ ,  $e(X, Y)$ ,  $e^*(X, Y)$  nie ocenia pokrycia reguły. Użycie tych miar do nadzorowania procesu indukcji może powodować wyznaczanie dużej liczby reguł nadmiernie dopasowanych do danych treningowych. Pokrycie reguł bierze pod uwagę miara (3.6), która wywodzi się z tzw. względnej przybliżonej funkcji przynależności [111] (3.5):

$$\bar{\mu}_{Y, B}(x) = \frac{|I_B(x) \cap Y|}{|I_B(x)|} - \frac{|(U - I_B(x)) \cap Y|}{|U - I_B(x)|}, \quad (3.5)$$

$$\bar{\mu}(X, Y) = \frac{|X \cap Y|}{|X|} - \frac{|(U - X) \cap Y|}{|U - X|}. \quad (3.6)$$

Miara (3.6) ma własności (1), (2), ale nie ma własności (3) i (4). Nietrudno zauważyć, że (3.6) to miara  $s$  z tabeli 3.5. Ma ona własność  $D_{HS}$  ( $const_1 = 0$ ) i dlatego nie spełnia własności (4). W szczególności każda miara jakości mająca własność  $D_{HS}$  ( $const_1 \neq 1$ ) nie będzie spełniać warunku (4).

Założymy, że  $B \subseteq A$  jest zbiorem atrybutów oraz  $X \subseteq U$  jest zbiorem przykładów. W tolerancyjnym modelu zbiorów przybliżonych dolne i górne przybliżenia zbiorów (w szczególności klas decyzyjnych) definiowane są w sposób następujący (3.7):

$$\begin{aligned} \underline{B}X &= \{x \in U : \mu(I_B(x), X) = 1\}, \\ \overline{B}X &= \{x \in U : \mu(I_B(x), X) > 0\}. \end{aligned} \quad (3.7)$$

Jeśli  $I_B(x)$  jest klasą abstrakcji relacji nieroróżnialności, uzyskamy standardowy model zbiorów przybliżonych. W pracy [346] Ziarko przedstawił model zmiennej precyzji (3.8), w którym przynależności do przybliżeń górnego i dolnego można regulować za pomocą parametrów  $0 \leq q \leq t \leq 1$ :

$$\begin{aligned} \underline{B}_t X &= \{x \in U : \mu(I_B(x), X) \geq t\}, \\ \overline{B}_q X &= \{x \in U : \mu(I_B(x), X) > q\}. \end{aligned} \quad (3.8)$$

W 2005 roku, Ślęzak [304] zaproponował alternatywną postać parametryzowanego modelu zbiorów przybliżonych, w którym wzięto pod uwagę informację o podstawowej dokładności klas decyzyjnych (3.9). Model ten nazwano Bayesowskim modelem zbiorów przybliżonych. Model Bayesowski, podobnie jak model zmiennej precyzji, wymaga ustalenia wartości dwóch parametrów  $0 \leq \varepsilon_q \leq \varepsilon_t \leq 1$ :

$$\begin{aligned} \underline{B}_{\varepsilon_t} X &= \left\{ x \in U : \mu(I_B(x), X) \geq \varepsilon_t \frac{|X|}{|U|} \right\}, \\ \overline{B}_{\varepsilon_q} X &= \left\{ x \in U : \mu(I_B(x), X) > \varepsilon_q \frac{|X|}{|U|} \right\}. \end{aligned} \quad (3.9)$$

Bardziej ogólny model zbiorów przybliżonych zaproponowali Greco, Matarazzo i Słowiński [111]. O przynależności przykładu do dolnego i górnego przybliżenia zbioru decyduje w nim nie tylko wartość inkluzji przybliżonej, ale także wartość miary konfirmacji (3.10). Taka definicja przybliżeń wymaga podania wartości czterech parametrów:  $0 \leq q \leq t \leq 1$ ,  $-1 \leq \alpha \leq \beta \leq 1$ . Jak pokazano w [111], przy szczególnych wartościach parametrów model (3.10) można sprowadzić do każdego z modeli (3.7)–(3.9):

$$\begin{aligned}\underline{B}_{t,\alpha}X &= \{x \in U : \mu(I_B(x), X) \geq t \text{ oraz } c(I_B(x), X) \geq \alpha\}, \\ \overline{B}_{q,\beta}X &= \{x \in U : \mu(I_B(x), X) > q \text{ lub } c(I_B(x), X) > \beta\}. \end{aligned}\quad (3.10)$$

We wzorach (3.10)  $c(I_B(x), X)$  to wartość jednej z sześciu rozważanych w [111] miar konfirmacji, mierzących względne zawieranie się  $I_B(x)$  w  $X$ . Wśród wymienianych w [111] miar znajdują się umieszczone w tabeli 3.5 miary  $f$  i  $s$ .

Ograniczenia nakładane w modelach (3.7)–(3.10) na wartość  $\mu(I_B(x), X)$  determinują to, że reguły generowane na podstawie reduktów charakteryzują się pewną minimalną dokładnością. Dodatkowo w modelu (3.10) wyznaczone reguły będą spełniały również wymagania minimalnej ogólności.

Przyjmijmy obecnie następujące oznaczenia:  $r_{I_B(x),X}$  jest regułą, której konkluzja wskazuje na klasę decyzyjną  $X$ . Zbiór warunków elementarnych reguły  $r_{I_B(x),X}$  zbudowany jest z atrybutów należących do zbioru  $B$ , a zakres dowolnego warunku elementarnego, zbudowanego na podstawie atrybutu  $a \in B$ , określany jest na podstawie zbioru  $I_a(x)$  (mamy zatem analogię do tolerancyjnego modelu zbiorów przybliżonych [271, 292, 295]). Wówczas można przedstawić model zbiorów przybliżonych (3.11), w którym o przynależności przykładu  $x$  do przybliżeń B-dolnego i B-górnego decyduje jakość utworzonej na jego podstawie reguły  $r_{I_B(x),X}$ :

$$\begin{aligned}\underline{B}_{\alpha,q}X &= \{x \in U : q(r_{I_B(x),X}) \geq \alpha_X\}, \\ \overline{B}_{\beta,q}X &= \{x \in U : q(r_{I_B(x),X}) > \beta_X\}. \end{aligned}\quad (3.11)$$

Propozycja (3.11) zachowuje podstawową własność przybliżeń, tzn.  $\underline{B}_{\alpha,q}X \subseteq \overline{B}_{\beta,q}X$ . Reszta własności uzależniona jest od użytej miary jakości. W szczególności dla wybranych miar może nie być spełniona własność  $\underline{B} \cup \{a\}_{\alpha,q} \subseteq \underline{B}_{\alpha,q}$ , gdyż rozszerzenie zbioru atrybutów może powodować nie tylko udokładnienie reguł, ale także spadek ich pokrycia.

Jeśli  $q = precision$ , to  $I_B(x)$  będzie klasą abstrakcji relacji nieroróżnicialności, wyznaczonej przez zbiór atrybutów  $B$ . Dodatkowo, jeśli niezależnie od rozważanego zbioru  $X$  wartości  $\alpha_X$  i  $\beta_X$  będą równe odpowiednio  $t$  i  $q$ , to otrzymamy model zmiennej precyzji. Dla  $\alpha_X = 1$  i  $\beta_X = 0$  otrzymamy standardowy model zbiorów przybliżonych. Z modelu (3.11) można uzyskać model (3.10) jedynie dla tych miar jakości  $q$ , które będą miarami konfirmacji, wymienionymi w pracy [111], oraz będą równoważne z miarą *precision* ze względu na porządek reguł. Jedynie dla takich miar możliwe jest bowiem ustalenie wartości

parametrów  $\alpha_X$  i  $\beta_X$  w taki sposób, aby dla ustalonych  $t$  i  $q \quad \forall X \subseteq U$   $q_{precision}(r_{I_B(x),X}) \geq t \Leftrightarrow q(r_{I_B(x),X}) \geq \alpha_X$  oraz  $q_{precision}(r_{I_B(x),X}) > q \Leftrightarrow q(r_{I_B(x),X}) > \beta_X$ .

Umożliwia to sprowadzenie (3.11) do (3.10).

Dolne przybliżenia klas decyzyjnych są podstawą do wyznaczenia minimalnych reguł decyzyjnych. Reguły minimalne generowane są na podstawie lokalnych reduktów względnych, wyznaczanych oddziennie dla każdego przykładu. Przedstawiony w rozdziale 2 algorytm RMatrix również realizuje ideę lokalnej indukcji reguł, ale nie używa do tego celu reduktów względnych. Indukcja prowadzona jest w taki sposób, aby reguły charakteryzowały się maksymalnie wysoką jakością, przy czym w fazie wzrostu procedura doboru atrybutów, z których zbudowana jest przesłanka, jest heurystyczna i podobna do procedury wyznaczania quasi-najkrótszego reduktu względnego [208].

**Stwierdzenie 3.12.** Założmy, że zbiór atrybutów ustawiono w ciąg  $ai = \langle a_{i1}, a_{i2}, \dots, a_{in} \rangle$ , w którym pierwszych  $s$  atrybutów tworzy zbiór  $S = \{a_{i1}, \dots, a_{is}\}$ , będący lokalnym reduktem względnym dla przykładu  $x \in X \subseteq U$ , wyznaczonym zgodnie z wymaganiami modelu (3.7). Założmy również, że w ciągu atrybutów  $aj = \langle a_{j1}, a_{j2}, \dots, a_{jn} \rangle$  pierwszych  $l$  atrybutów tworzy zbiór  $B = \{a_{j1}, \dots, a_{jl}\}$ , będący najmniejszym (w sensie inkluzji) zbiorem atrybutów spełniających warunek  $q(r_{I_B(x),X}) \geq \alpha_X$ . Jeśli miara  $q$  ocenia zarówno dokładność, jak i pokrycie reguły  $r$  oraz podstawą do tworzenia warunków elementarnych dla algorytmu RMatrix był ciąg  $ai$ , to reguła  $r$  ma nie więcej niż  $s$  warunków elementarnych. Jeśli miara  $q$  ocenia zarówno dokładność, jak i pokrycie reguły  $r$  oraz podstawą do tworzenia warunków elementarnych dla algorytmu RMatrix był ciąg  $aj$ , to reguła  $r$  ma nie mniej niż  $l$  warunków elementarnych.

Dowód: dowód wynika wprost ze specyfiki algorytmu RMatrix. W algorytmie tym wyjściową regułą jest ta, która maksymalizuje wartość miary  $q$ . W standardowym modelu zbiorów przybliżonych (3.7) wyznaczenie reguły  $r_{I_s(x),X}$  na podstawie reduku względnego gwarantuje, że w rozważanym zbiorze przykładów nie można bardziej udokładnić tej reguły. Dodanie do  $r_{I_s(x),X}$  kolejnych warunków, zbudowanych na podstawie atrybutów  $\{a_{i(s+1)}, \dots, a_{in}\}$ , spowoduje jedynie spadek jej dokładności, a więc spadek wartości miary  $q$ . W przypadku ciągu  $aj$  pierwszych  $l$  występujących w nim atrybutów zapewnia, że reguła  $r_{I_B(x),X}$  charakteryzuje się jakością nie mniejszą niż  $\alpha_X$ . Nie można wykluczyć, że dodanie do jej przesłanki kolejnych warunków elementarnych, zbudowanych na podstawie atrybutów  $\{a_{j(l+1)}, \dots, a_{jn}\}$ , spowoduje wzrost wartości miary  $q$ . Oznacza to, że reguła  $r_{I_B(x),X}$  zawiera co najmniej  $l$  warunków elementarnych. ◆

Odpowiednio zaadaptowane miary jakości mogą również być użyte do nadzorowania procesu poszukiwania progów tolerancji. Dla zadanego zbioru atrybutów  $B \subseteq A$  i przykładu  $x \in U$  zbiór tolerancji  $I_B(x)$  może być definiowany m.in. w następujący sposób (3.12):

$$y \in I_B(x) \Leftrightarrow \forall a \in B (\delta_a(a(x), a(y)) < \varepsilon_a). \quad (3.12)$$

We wzorze (3.12)  $\delta_a$  jest pewną miarą odległości, natomiast  $\varepsilon_a$  to próg tolerancji. Dla danego zbioru przykładów i zbioru atrybutów wyboru miary odległości dokonuje się zazwyczaj arbitralnie, natomiast progi tolerancji powinny być takie, aby:

- maksymalizować w zbiorze  $I_B(x)$  liczbę przykładów, które należą do tej samej klasy decyzyjnej co  $x$ ,
- minimalizować w zbiorze  $I_B(x)$  liczbę przykładów nienależących do klasy decyzyjnej, która reprezentuje  $x$ .

Przedstawione wymagania co do wartości progu tolerancji dotyczą pojedynczego przykładu. Problem doboru wartości progów tolerancji może być rozważany także w obrębie całej klasy decyzyjnej lub globalnie, w całym zbiorze przykładów [210]. W podejściu globalnym progi dobierane są w taki sposób, aby jak największa liczba par przykładów uznanych za podobne należała do tych samych klas decyzyjnych oraz aby maksymalnie wiele par przykładów nienależących do tych samych klas decyzyjnych zostało uznanych za niepodobne. Założymy, że dany jest zbiór przykładów  $\mathbf{DT} = (U, A \cup \{d\})$ . Zdefiniujmy dwa zbioru przykładów oraz wektora progów tolerancji  $\langle \varepsilon_{a1}, \varepsilon_{a2}, \dots, \varepsilon_{an} \rangle$  można zdefiniować tablicę kontyngencji (tabela 3.6).

Tabela 3.6  
Tablica kontyngencji dla wektora progów tolerancji

$n_{R_d R_{IA}}$	$n_{R_d \neg R_{IA}}$	$n_{R_d}$
$n_{\neg R_d R_{IA}}$	$n_{\neg R_d \neg R_{IA}}$	$n_{\neg R_d}$
<hr/>		
$n_{R_{IA}}$	$n_{\neg R_{IA}}$	

W tabeli 3.6  $n_{R_d} = n_{R_d R_{IA}} + n_{R_d \neg R_{IA}}$  oznacza liczbę par przykładów mających tę samą wartość atrybutu decyzyjnego;  $n_{\neg R_d} = n_{\neg R_d R_{IA}} + n_{\neg R_d \neg R_{IA}}$  oznacza liczbę par przykładów mających różne wartości atrybutu decyzyjnego;  $n_{R_{IA}} = n_{R_d R_{IA}} + n_{\neg R_d R_{IA}}$  to liczba par przykładów  $\langle x, y \rangle \in U^2$ , dla których  $y \in I_A(x)$  i  $x \in I_A(y)$ ;  $n_{\neg R_{IA}} = n_{R_d \neg R_{IA}} + n_{\neg R_d \neg R_{IA}}$  oznacza liczbę par

przykładów  $\langle x, y \rangle \in U^2$ , dla których  $y \notin I_A(x)$  i  $x \notin I_A(y)$ ;  $n_{R_d R_{IA}}$  to liczba par przykładów mających identyczną wartość atrybutu decyzyjnego oraz takich, że  $y \in I_A(x)$  i  $x \in I_A(y)$ . Znaczenie wartości  $n_{\neg R_d R_{IA}}, n_{R_d \neg R_{IA}}, n_{\neg R_d \neg R_{IA}}$  można wywieść przez analogię do  $n_{R_d R_{IA}}$ .

Zauważmy, że wobec przyjętych założeń każda z miar definiowanych na podstawie tablicy kontyngencji może zostać zaadaptowana do oceny wektora progów tolerancji. Od wartości progów tolerancji zależą zakresy warunków elementarnych reguł wyznaczanych w tolerancyjnym modelu zbiorów przybliżonych. Dla ustalonego przykładu  $x \in U$  oraz atrybutu  $a \in A$  zakres  $Za$  w warunku elementarnym  $a \in Za$  uzależniony jest od wartości progu tolerancji  $\varepsilon_a$  i funkcji odległości  $\delta_a$ . Przykładowo dla atrybutów typu symbolicznego  $Za = \{a(y) \in Va : y \in I_a(x)\}$ , natomiast dla atrybutu numerycznego  $Za = [v_{min}, v_{max}]$ , gdzie  $v_{min} = \min_{y \in U} \{a(y) \in Va : y \in I_a(x)\}$ ,  $v_{max} = \max_{y \in U} \{a(y) \in Va : y \in I_a(x)\}$ . Od wartości progów tolerancji zależą zatem „szerokości” zakresów warunków elementarnych, a to bezpośrednio wpływa na dokładność i pokrycie reguł.

Po niewielkiej modyfikacji tabela 3.6 może opisywać problem doboru progów tolerancji dla konkretnej klasy decyzyjnej (zamiast rozważać pary należące do  $U \times U$ , rozważamy  $U \times X$ ) lub dla konkretnego przykładu (rozważamy jedynie pary  $U \times \{x\}$ ).

Wyniki badań związanych z analizą efektywności algorytmu RMatrix oraz przedstawiony sposób oceny jakości wektora progów tolerancji opisano m.in. w [254].

### 3.1.8. Ocena statystycznej istotności reguł decyzyjnych

Poziom statystycznej istotności reguł ma duże znaczenie w ocenie reguł asocjacyjnych, gdzie często stanowi on podstawę do selekcji reguł najbardziej interesujących [329]. Statystycznej oceny reguł dokonują także niektóre algorytmy pokryciowe. W opisywanym w rozdziale 2 algorytmie CN2 statystyczna ocena kompleksów stanowi zabezpieczenie przed indukcją zbyt specyficznych reguł. W indukcji reguł decyzyjnych dla celów opisowych [94, 213] poziom statystycznej istotności stanowi jedną z podstawowych miar charakteryzujących wyznaczone reguły.

W tablicy kontyngencji charakteryzującej regułę zawarte są informacje, które można wykorzystać do weryfikacji następującej hipotezy  $H_0$ :

*przyporządkowanie t przykładów pokrywanych przez regułę do klasy decyzyjnej wskazywanej w jej konkluzji jest równoważne z przypadkowym (losowym) przyporządkowaniem dowolnych t przykładów do tej klasy decyzyjnej.*

Tablica kontyngencji związana z hipotezą  $H_0$  jest tablicą dwudzielczą, opisującą zachowanie zmiennych jakościowych. Najbardziej rozpowszechnionym testem istotności dla zmiennych jakościowych jest test  $\chi^2$  Pearsona. Wartość testu rośnie (test staje się istotny)

w miarę jak liczebności oczekiwane (losowe przyporządkowanie przykładów do rozważanej klasy decyzyjnej) zaczynają się coraz bardziej różnić od liczebności obserwowanych (przyporządkowanie przykładów pokrywanych przez regułę do rozważanej klasy decyzyjnej). Wartość testu zależy od rozmiaru próbki (liczby przykładów) oraz liczebności charakteryzujących komórki tablicy kontyngencji. Dla małych liczebności oczekiwanych wartość testu może być wysoce nieprecyzyjna. Ponadto, jeśli rozmiar próbki jest niewielki, to stosunkowo małe odchylenia liczebności obserwowanych od liczebności oczekiwanych mogą okazać się istotne. Dlatego też stosowane są różnego rodzaju poprawki testu  $\chi^2$  (np. poprawka *Yatesa*). Dla tabel o wymiarach  $2 \times 2$  możliwe jest obliczenie dokładnego prawdopodobieństwa otrzymania tabeli o liczebnościami obserwowanych. Umożliwia to dokładny test *Fishera* [80], w którym prawdopodobieństwo to wylicza się zgodnie ze wzorem (3.13):

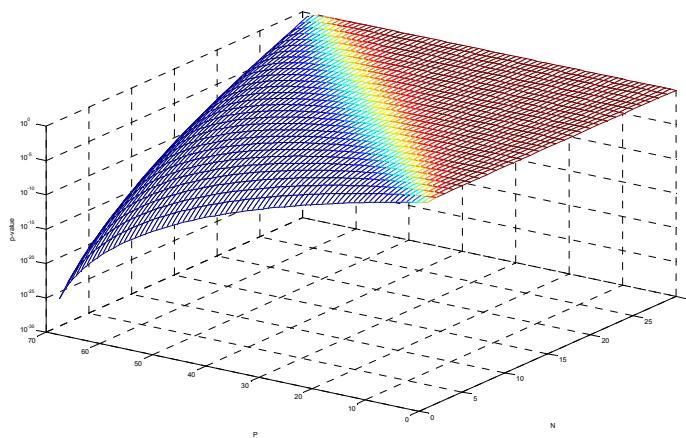
$$Pr(p, n) = \frac{\binom{P}{p} \binom{N}{n}}{\binom{P+N}{p+n}}. \quad (3.13)$$

We wzorze (3.13) zastosowano notację przyjętą dla tablicy kontyngencji 3.1 i reguł decyzyjnych.

Dokładny test Fishera bazuje na rozkładzie hipergeometrycznym, który jest rozkładem dyskretnym. Dla ustalonego zbioru przykładów i ustalonej reguły, aby obliczyć p-wartość testu (p-wartość reguły), sumujemy wartości prawdopodobieństw reguł pokrywających taką samą liczbę przykładów jak reguła  $r$  i charakteryzujących się dokładnością identyczną lub większą od  $r$  (3.14) [7]. Jest to równoznaczne z zastosowaniem testu prawostronnego:

$$p_{val}(p, n) = \sum_{k=0}^{\min\{P-p, n\}} Pr(p+k, n-k). \quad (3.14)$$

W dalszej części monografii statystyczna ocena reguł będzie właśnie bazować na prawostronnym dokładnym teście Fishera. Taki sposób statystycznej oceny reguł bardzo często stosowany jest w bioinformatyce [225]. Dokładny test Fishera używany jest do statystycznej oceny reguł w serwisach internetowych Genecodis [45] i RuleGO [116]. Trójwymiarowy wykres  $p_{val}$  zaprezentowano na rysunku 3.4.



Rys. 3.4. Wykres miary  $p_{val}$  (skala logarytmiczna)  
Fig. 3.4. Graph of the  $p_{val}$  measure (logarithmic scale)

Ze statystyczną oceną wiąże się problem tzw. fałszywych odkryć. Dla pojedynczego testu statystycznego poziom istotności  $\alpha$  określa prawdopodobieństwo popełnienia błędu I rodzaju. Błąd ten polega na odrzuceniu hipotezy zerowej, podczas gdy jest ona prawdziwa. W naszym przypadku będzie to oznaczało uznanie reguły za statystycznie istotną, gdy w rzeczywistości tak nie jest. Dla dużej liczby reguł (dużej liczby testów), których istotność statystyczna weryfikowana jest na poziomie istotności  $\alpha$ , wzrasta prawdopodobieństwo, że poziom odrzuconych fałszywych hipotez zerowych będzie wyższy od  $\alpha$ . Jeśli testów takich jest  $m$ , to prawdopodobieństwo popełnienia co najmniej jednego błędu I rodzaju wynosi  $\pi = 1 - (1 - \alpha)^m$ . W statystyce matematycznej problematyka związana z definiowaniem procedur umożliwiających wyznaczanie dla każdego pojedynczego testu takiego poziomu istotności, aby w ramach całego eksperymentu nie przekroczyć założonego poziomu frakcji fałszywych odkryć, nazywa się problemem testowania wielokrotnego [71, 232]. Najbardziej popularne procedury kontroli poziomu fałszywych odkryć to: korekcja Bonferroniego [100], korekcja Holma [134], metoda permutacji Westwalla i Younga oraz kontrola współczynnika fałszywych odkryć (FDR – ang. *false discovery rate*), zaproponowana przez Benjamina i Hochberga [20]. W literaturze przedmiotu można znaleźć różnego rodzaju modyfikacje tych metod.

W praktycznych zastosowaniach najczęściej używane są: korekcja Bonferroniego oraz metoda FDR. W metodzie Bonferroniego skorygowana p-wartość każdej reguły równa się jej p-wartości bez korekcji, pomnożonej przez ogólną liczbę reguł. Taki sposób korekcji powoduje, że metoda jest bardzo restrykcyjna, co znacząco wpływa na moc testu. Moc testu związana jest z prawdopodobieństwem popełnienia błędu II rodzaju, czyli nieodrzuceniem hipotezy zerowej, podczas gdy jest ona fałszywa. W naszym przypadku oznacza to uznanie istotnej statystycznie reguły za regułę nieistotną. Metoda Bonferroniego kontroluje prawdopodobieństwo popełnienia co najmniej jednego błędu I rodzaju (ang. *family wise*

*terror rate* – FWER) na poziomie całego zbioru reguł. Przyjmując FWER=0.05, spodziewamy się, z prawdopodobieństwem 0.95, że wszystkie reguły, których skorygowana p-wartość jest mniejsza lub równa 0.05, są rzeczywiście statystycznie istotne.

Procedura kontroli FDR przebiega według następującego schematu. Założymy, że dany jest zbiór  $RUL$ , złożony z  $m$  reguł. Dla czytelności zapisu przyjmijmy również, że przez  $p_{val}(r)$  będziemy rozumieli p-wartość reguły  $r$ , obliczoną zgodnie ze wzorem (3.14). Przez  $p_{val}^{RUL}(r)$  oznaczymy skorygowaną p-wartość reguły  $r$  w zbiorze reguł  $RUL$ .

1. Sortując reguły ze zbioru  $RUL$  zgodnie z rosnącymi wartościami  $p_{val}(r)$  otrzymamy ciąg  $\langle r_1, r_2, \dots, r_m \rangle$ , w którym ostatnia reguła charakteryzuje się najwyższą (czyli najgorszą) p-wartością.
2. Dla reguły  $r_m$   $p_{val}^{RUL}(r_m) = p_{val}(r_m)$ .
3. Dla każdego  $i \in \{1, 2, \dots, m-1\}$  skorygowane p-wartości pozostałych reguł oblicza się na podstawie wzoru  $p_{val}^{RUL}(r_{m-i}) = p_{val}(r_{m-i})(m/(m-i))$ .

Na podstawie skorygowanych p-wartości możemy ponownie dokonać weryfikacji tego, które z reguł przestały spełniać minimalne wymagania statystycznej istotności. Przyjmując FDR=0.05, spodziewamy się, że w zbiorze reguł, których skorygowana p-wartość jest mniejsza lub równa 0.05, 5% reguł jest statystycznie nieistotnych.

Problem eliminacji fałszywych odkryć dotyczy zbiorów reguł otrzymanych za pomocą algorytmów, które na podstawie dostępnego zbioru przykładów wyznaczają wszystkie możliwe reguły spełniające zadane przez użytkownika ograniczenia [329]. Tematyka ta poruszana jest m.in. w pracach [96, 179, 329]. W szczególności Webb [329] prezentuje dwie metody korekcji. Pierwsza stosuje korekcję Bonferroniego, nie jest wykonywana na podstawie informacji o liczbie wyznaczonych reguł, ale na podstawie teoretycznej liczby wszystkich możliwych reguł, jakie można wyznaczyć w analizowanym zbiorze danych (metoda dotyczy jedynie atrybutów dyskretnych). Druga metoda korekcji stosuje opisane w rozdziale 2 podejście *hold-out*. Dostępny zbiór danych dzielony jest na dwa rozłączne podzbiory, na jednym z nich dokonywana jest indukcja reguł, a na podstawie drugiego następuje weryfikacja ich statystycznej istotności.

Trudno jednoznacznie określić, czy problem korygowania statystycznej istotności dotyczy reguł generowanych przez algorytmy pokryciowe czy tworzonych na podstawie najkrótszych bądź quasi-najkrótszych lokalnych reductów względnych. Obie metody nie generują wszystkich możliwych reguł, a jedynie pewien ich podzbiór. Jednak liczba tworzonych reguł może być duża, więc prawdopodobieństwo popełnienia co najmniej jednego błędu I rodzaju również będzie duże (dla 100 reguł i  $\alpha = 0.05$  wynosi ono 0.995). W pracach dotyczących indukcji reguł dla celów klasyfikacji aspekt oceny statystycznej

istotności reguł jest często pomijany. Gdy reguły definiowane są dla celów opisowych, podawany jest zazwyczaj nieskorygowany poziom istotności [94, 213]. W dalszej części, podczas omawiania wyników niektórych eksperymentów, przywoływany będzie zarówno podstawowy, jak i skorygowany (kontrola FDR) poziom statystycznej istotności.

### 3.1.9. Pomiar sily interwencji reprezentowanej przez regułę

Dotychczas zakładaliśmy, że wyznaczone reguły stosowane są do klasyfikacji przykładów i/lub do ich opisu. Na reguły decyzyjne można spojrzeć również jako na strategię, zwaną strategią interwencji [108, 109, 277], która wskazuje, jakie działania należy podjąć, aby przykłady obecnie przyporządkowane do określonych klas decyzyjnych stały się reprezentantami innych klas. Podejmowane działania polegają na zmianie wartości atrybutów warunkowych.

W pracy [108] Greco i inni przedstawiają następujący przykład:

*Załóżmy, że dana jest reguła: „jeśli składnik  $\alpha$  występuje w składzie krwi pacjenta, to pacjent jest zdrowy”. Reguła ta sugeruje strategię postępowania, polegającą na wstrzyknięciu składnika  $\alpha$  do krwi chorych pacjentów.*

Oczywiście składnik  $\alpha$  wstrzykiwany jest jedynie tym chorym pacjentom, u których on nie występuje.

Opisy klas decyzyjnych zazwyczaj składają się z więcej niż jednej reguły. Oznacza to, że istnieje wiele potencjalnych strategii interwencji. Ponadto, większość reguł zbudowana jest z więcej niż jednego warunku elementarnego. Realizacja strategii interwencji, wskazywanych przez takie reguły, wymagać będzie podjęcia większej liczby działań. Nawiązując do podanego przykładu, warto zauważyć, że w praktyce część chorych pacjentów nie będzie spełniać jedynie kilku z warunków znajdujących się w przesłance reguły stanowiącej podstawę do interwencji. W niniejszym rozdziale skoncentrujemy się na zaprezentowaniu najważniejszych wskaźników pozwalających oszacować potencjalną efektywność strategii interwencji, sugerowanej przez regułę [108].

Załóżmy, że dane są dwie tablice decyzyjne:  $T = (U, A \cup \{d\})$ ,  $T' = (U', A \cup \{d\})$ . Na podstawie tablicy  $T$  wyznaczono regułę  $r \equiv \varphi \rightarrow \psi$ , w której  $\varphi \equiv w_1 \wedge w_2 \wedge \dots \wedge w_n$ . Reguła  $r$  stanowi podstawę do eksperymentu polegającego na tym, że wartości atrybutów wszystkich przykładów z  $U'$  niespełniających przesłanki  $\varphi$  i nienależących do  $\psi$  zmieniane są w taki sposób, aby przykłady te spełniały  $\varphi$ . Po zastosowaniu interwencji sugerowanej przez regułę  $r$  zbiór  $U'$  przekształca się w  $U''$ . W zbiorze  $U''$  możemy ostatecznie zweryfikować efektywność podjętych działań.

Przez  $precision(r, U)$  oznaczmy dokładność reguły  $r$  w zbiorze przykładów  $U$ . Przez  $[\zeta]$  oznaczmy zbiór przykładów z  $U$  pokrywanych przez  $\zeta$ . Przez  $[\zeta]'$  oznaczmy zbiór przykładów z  $U'$  pokrywanych przez  $\zeta$ .

Definiując wskaźniki efektywności interwencji, Greco i inni [108] zakładają, że tablice  $T$  i  $T'$  są jednorodne. Oznacza to, że dysponując regułą  $r$  o dokładności  $precision(r, U)$ , oczekujemy, że zmieniając warunek  $\neg\varphi$  na  $\varphi$  w zbiorze  $[\neg\varphi]'\cap[\neg\psi]', precision(r, U)\cdot|[\neg\varphi]'\cap[\neg\psi]'|$  przykładów zmieni przyporządkowanie klasy decyzyjnej na  $\psi$ . Zauważmy, że  $\neg\varphi$ , zgodnie z przyjętymi założeniami, można równoważnie zapisać jako  $\neg w_1 \vee \neg w_2 \vee \dots \vee \neg w_n$ , wynika stąd, że dla konkretnego przykładu zmiana  $\neg\varphi$  na  $\varphi$  nie musi wiązać się ze zmianą wartości wszystkich atrybutów, na podstawie których zbudowano  $w_1, w_2, \dots, w_n$ .

Oczekiwany względny wzrost liczby przykładów przyporządkowanych w zbiorze  $U''$  do klasy decyzyjnej  $\psi$ , po zastosowaniu do przykładów z  $U'$  interwencji sugerowanej przez regułę  $r \equiv \varphi \rightarrow \psi$ , można wyrazić za pomocą dwóch równoważnych wyrażeń (3.15), (3.16):

$$\delta(\psi) = \frac{|[\neg\varphi]'\cap[\neg\psi]'|}{|[\neg\psi]'|} \cdot \frac{|[\neg\psi]'|}{|U'|} \cdot \frac{|[\varphi]\cap[\psi]|}{|[\varphi]|}, \quad (3.15)$$

$$\delta(\psi) = \frac{|[\neg\varphi]'\cap[\neg\psi]'|}{|[\neg\varphi]'|} \cdot \frac{|[\neg\varphi]'|}{|U'|} \cdot \frac{|[\varphi]\cap[\psi]|}{|[\varphi]|}. \quad (3.16)$$

W powyższych wzorach pierwsze i trzecie składniki iloczynów są identyczne. Przez  $s$  oznaczmy regułę  $\neg\psi \rightarrow \neg\varphi$ , a przez  $t$  regułę  $\neg\varphi \rightarrow \neg\psi$ . Wskaźnik  $\delta$  można zapisać jako (3.17) lub (3.18):

$$\delta(\psi) = precision(r, U) \cdot precision(s, U') \cdot \frac{|[\neg\psi]'|}{|U'|}, \quad (3.17)$$

$$\delta(\psi) = precision(r, U) \cdot precision(t, U') \cdot \frac{|[\neg\varphi]'|}{|U'|}. \quad (3.18)$$

Zauważmy, że zgodnie ze stwierdzeniem 3.3 wyrażenia (3.17), (3.18) można zapisać równoważnie jako (3.19), (3.20):

$$\delta(\psi) = precision(r, U) \cdot coverage(t, U') \cdot \frac{|[\neg\psi]'|}{|U'|}, \quad (3.19)$$

$$\delta(\psi) = precision(r, U) \cdot coverage(s, U') \cdot \frac{|[\neg\varphi]'|}{|U'|}. \quad (3.20)$$

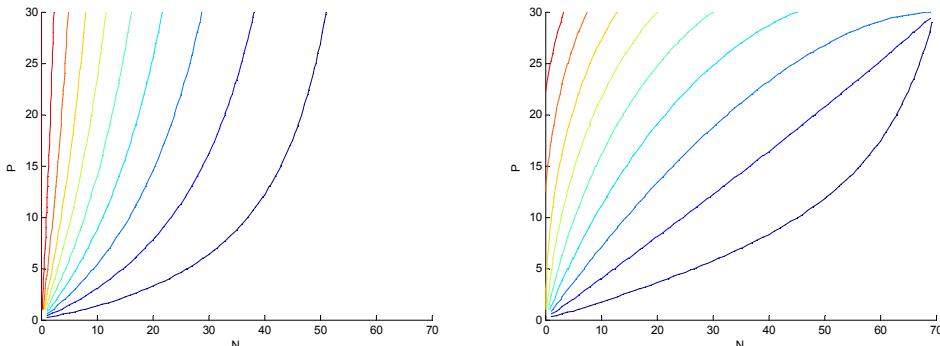
Greco i inni [108] pierwsze dwa składniki iloczynu (3.17) oznaczają jako  $E^{\psi}(r, U, U')$  i nazywają wskaźnikiem efektywności interwencji ze względu na konkluzję reguły  $r$ . Podobnie pierwsze dwa składniki iloczynu (3.18) autorzy ci oznaczają jako  $E^{\varphi}(r, U, U')$  i nazywają wskaźnikiem efektywności interwencji ze względu na przesłankę reguły  $r$ .

Zakładając, że  $U = U'$ , wskaźniki  $E^{\psi}$  i  $E^{\varphi}$  można traktować jako miary jakości reguł decyzyjnych, definiowane na podstawie tablicy kontyngencji. Stosując notację z tablicy kontyngencji (tabela 3.1), miary te można przedstawić za pomocą wzorów (3.21) i (3.22):

$$E^{\psi}(r) = \frac{p}{p+n} \cdot \frac{N-n}{N}, \quad (3.21)$$

$$E^{\varphi}(r) = \frac{p}{p+n} \cdot \frac{N-n}{P+N-p-n}. \quad (3.22)$$

Na rysunku 3.5 zaprezentowano poziomice wykresów miar  $E^{\psi}$  i  $E^{\varphi}$ . Miary  $E^{\psi}$  i  $E^{\varphi}$  nie są równoważne ze względu na porządek reguł. Na wykresach widać wyraźnie, że  $E^{\varphi}$  większą wagę niż  $E^{\psi}$  przywiązuje do oceny pokrycia reguł.



Rys. 3.5. Wykresy poziomic miar  $E^{\psi}$  (wykres lewy) i  $E^{\varphi}$  (wykres prawy)  
Fig. 3.5. Graphs of contour lines of  $E^{\psi}$  (left graph) and  $E^{\varphi}$  (right graph)

Można pokazać, że miara  $E^{\psi}$  spełnia warunek symetrii  $E^{\psi}(\varphi \rightarrow \psi) = E^{\psi}(\neg\psi \rightarrow \neg\varphi)$ . Ponadto, im większe są zbiory  $[\psi] \cap [\varphi]$ ,  $[\neg\psi] \cap [\neg\varphi]$  oraz im mniejszy jest zbiór  $[\neg\psi] \cap [\varphi]$ , tym wartość miary będzie większa. Oznacza to, że  $E^{\psi}$  można wykorzystać do oceny reguł interpretowanych jako quasi-implikacje. Dodatkowo, dla miary  $E^{\psi}$  prawdziwa jest zależność:  $E^{\psi}(\psi \rightarrow \varphi) = E^{\psi}(\neg\varphi \rightarrow \neg\psi) = \text{coverage}(\varphi \rightarrow \psi) \cdot \frac{N-n}{P+N-p-n}$ .

Miara  $E^{\varphi}$  charakteryzuje się symetrią  $E^{\varphi}(\varphi \rightarrow \psi) = E^{\varphi}(\neg\varphi \rightarrow \neg\psi)$ , ponadto  $E^{\varphi}(\psi \rightarrow \varphi) = E^{\varphi}(\neg\psi \rightarrow \neg\varphi) = \text{coverage}(\varphi \rightarrow \psi) \cdot \frac{N-n}{N}$ .

Z punktu widzenia efektywności interwencji najlepszą strategię (abstrahujemy od liczby działań, jakie należy podjąć, aby strategię tę zrealizować) będzie wskazywać reguła  $\varphi \rightarrow \psi$ ,

która w zbiorze  $U$  jest dokładna, a w zbiorze  $U'$  maksymalizuje liczebność zbioru  $[\neg\psi]'\cap[\neg\varphi]'$ . Potencjalnie regułę taką można zastosować do największej liczby przykładów w  $U'$ . Jednak reguła dokładna może być bardzo specyficzna. Z punktu widzenia odkrywania wiedzy interesujące są reguły (wzorce), które są nie tylko dokładne, ale również ogólne. Oznacza to, że im mniejsza jest liczebność zbiorów  $[\neg\psi]\cap[\varphi]$  (wpływ na dokładność reguły  $\varphi\rightarrow\psi$  w  $U$ ) i  $[\psi]\cap[\neg\varphi]$  (wpływ na pokrycie reguły  $\varphi\rightarrow\psi$  w  $U$ ) oraz  $[\neg\psi]'\cap[\varphi]'$  (wpływ na pokrycie reguły  $\neg\varphi\rightarrow\neg\psi$  i dokładność reguły  $\neg\psi\rightarrow\neg\varphi$  w  $U'$ ) i  $[\psi]'\cap[\neg\varphi]'$  (wpływ na dokładność reguły  $\neg\varphi\rightarrow\neg\psi$  i pokrycie reguły  $\neg\psi\rightarrow\neg\varphi$  w  $U'$ ), tym interwencja sugerowana przez  $\varphi\rightarrow\psi$  powinna być efektywniejsza. Zakładając, że  $U=U'$ , bliższa tej idei jest miara  $E^\varphi$ , która w ocenie reguły wykorzystuje informacje ze wszystkich czterech komórek tablicy kontyngencji.

Zakładając, że  $U=U'$ , z powyższego rozumowania wynika, że zbiory przykładów pozytywnych dla reguły  $\varphi\rightarrow\psi$  to  $[\psi]\cap[\varphi]$  i  $[\neg\psi]\cap[\neg\varphi]$ , natomiast zbiory przykładów negatywnych to  $[\neg\psi]\cap[\varphi]$  i  $[\psi]\cap[\neg\varphi]$ . Oznacza to, że regułę interpretujemy jako quasi-równoważność. Blanchard [27] sugeruje, że miary oceny quasi-równoważności powinny w identyczny sposób oceniać reguły  $\varphi\rightarrow\psi$ ,  $\psi\rightarrow\varphi$ ,  $\neg\varphi\rightarrow\neg\psi$ ,  $\neg\psi\rightarrow\neg\varphi$ . W naszym przypadku jedynie reguły  $\varphi\rightarrow\psi$  i  $\neg\varphi\rightarrow\neg\psi$  można traktować jako wskazówki dla interwencji, miara oceniająca siłę interwencji powinna oceniać je w identyczny sposób. Warunek ten spełnia  $E^\varphi$  oraz wszystkie miary przeznaczone do oceny quasi-równoważności (np. *Corr*).

Miara  $E^\varphi$  spełnia warunek  $E^\varphi(\varphi\rightarrow\psi)=E^\varphi(\neg\psi\rightarrow\neg\varphi)$  i zgodnie z (3.19) dokonuje oceny dokładności reguł  $\varphi\rightarrow\psi$  i pokrycia reguły  $\neg\varphi\rightarrow\neg\psi$ . Miara  $E^\varphi$  spełnia warunek  $E^\varphi(\varphi\rightarrow\psi)=E^\varphi(\neg\varphi\rightarrow\neg\psi)$  i zgodnie z (3.18) dokonuje oceny dokładności reguł  $\varphi\rightarrow\psi$  i  $\neg\varphi\rightarrow\neg\psi$ . Miara *Corr* w ocenie bierze pod uwagę dokładność i pokrycie reguły  $\varphi\rightarrow\psi$  oraz dokładność i pokrycie reguły  $\neg\varphi\rightarrow\neg\psi$ . Miara ta może stanowić jeszcze jeden wskaźnik informujący o efektywności interwencji. Autor nie stawia tezy, że wskaźnik ten może zastąpić  $E^\varphi$  i  $E^\varphi$ , z pewnością jednak może być pomocny w tych sytuacjach, w których  $E^\varphi$  i  $E^\varphi$  oceniają reguły w sposób antagonistyczny.

W praktyce część przykładów należących do  $[\neg\psi]'$  nie będzie spełniała jedynie wybranych warunków zdefiniowanych w przesłance  $r\equiv\varphi\rightarrow\psi$ ,  $\varphi\equiv w_1\wedge w_2\wedge\dots\wedge w_n$ . Oznacza to, że aby zrealizować strategię sugerowaną przez  $r$ , wystarczy przykładom tym zmienić wartości jedynie wybranych atrybutów [108]. Przez  $Ec(r)$  oznaczmy zbiór wszystkich warunków elementarnych, tworzących przesłankę reguły  $r$ . Przez  $W\subseteq Ec(r)$

oznaczmy zbiór warunków elementarnych, których nie spełniają przykłady należące do zbioru  $[\neg\psi]'$ . Wówczas wskaźnik  $\delta_W$ , oznaczający oczekiwany wzrost liczby przykładów przyporządkowanych w zbiorze  $U''$  do klasy decyzyjnej  $\psi$  po odpowiedniej zmianie wartości atrybutów tworzących warunki ze zbioru  $W$ , definiowany jest jako (3.23) lub, równoważnie, (3.24):

$$\delta_W(\psi) = \frac{\left| \bigcap_{w \in W} [\neg w]' \cap \bigcap_{w \notin W} [w]' \cap [\neg\psi]' \right|}{|[\neg\psi]'|} \cdot \frac{|[\neg\psi]'|}{|U'|} \cdot \frac{|[\varphi] \cap [\psi]|}{|[\psi]|}, \quad (3.23)$$

$$\delta_W(\psi) = \frac{\left| \bigcap_{w \in W} [\neg w]' \cap \bigcap_{w \notin W} [w]' \cap [\neg\psi]' \right|}{\left| \bigcap_{w \in W} [\neg w]' \cap \bigcap_{w \notin W} [w]' \right|} \cdot \frac{\left| \bigcap_{w \in W} [\neg w]' \cap \bigcap_{w \notin W} [w]' \right|}{|U'|} \cdot \frac{|[\varphi] \cap [\psi]|}{|[\psi]|}. \quad (3.24)$$

Na podstawie warunków elementarnych, należących do  $W$ , określa się reguły  $s_w \equiv \neg\psi \rightarrow (\bigwedge_{w \in W} \neg w) \wedge (\bigwedge_{w \notin W} w)$ ,  $t_w (\bigwedge_{w \in W} \neg w) \wedge (\bigwedge_{w \notin W} w) \rightarrow \neg\psi$  oraz wskaźniki  $E_W^\psi(r, U, U')$ ,  $E_p^\psi(r, U, U')$ . Za pomocą  $\delta_W(\psi)$  można zdefiniować  $\delta(\psi)$  (3.25):

$$\delta(\psi) = \sum_{W \subseteq Ec(r)} \delta_W(\psi). \quad (3.25)$$

## 3.2. Miary niedefiniowane bezpośrednio na podstawie tablicy kontyngencji

Założymy, że dane są: tablica decyzyjna  $\mathbf{DT} = (U, A \cup \{d\})$  oraz zbiór reguł decyzyjnych  $RUL$ , których ocena wykonywana będzie na podstawie  $\mathbf{DT}$ . W dalszej części rozdziału zostaną omówione wybrane aspekty obiektywnej oceny reguł za pomocą miar niedefiniowanych bezpośrednio na podstawie informacji zawartych w tablicy kontyngencji 3.1.

### 3.2.1. Długość reguły

Najczęściej przywoływananą miarą oceny reguł, która w żaden sposób nie wykorzystuje informacji zawartych w tablicy kontyngencji 3.1, jest miara  $L$  (ang. *Length*), informująca o liczbie warunków elementarnych, z jakich zbudowana jest przesłanka [143, 254, 342]. Przez  $Ec(r) = \{w : w \in r\}$  oznaczmy zbiór zawierający wszystkie warunki elementarne tworzące przesłankę reguły  $r$ ; wówczas  $L(r) = |Ec(r)|$ . W zależności od preferencji użytkownika miarę  $L$  można traktować jako miarę kosztu lub miarę korzyści. Najczęściej  $L$  traktuje się jako miarę kosztu, słusznie zauważając, że mniejsza liczba warunków umożliwia łatwiejszą interpretację reguły. Mogą jednak zdarzyć się takie zastosowania, w których

użytkownikowi zależy na wyznaczaniu reguł jak najdłuższych. Miara  $L$  będzie wtedy miarą korzyści. Autor zetknął się z taką interpretacją miary  $L$  podczas indukcji reguł przeznaczonych do funkcjonalnego opisu genów [255, 257] (patrz rozdział 6).

Miarę  $L$  można znormalizować, dzieląc jej wartość przez ogólną liczbę atrybutów warunkowych  $nL(r) = L(r)/|A|$ . Założymy, że miary  $L$  i  $nL$  traktowane są jako miary kosztu.

Zauważmy, że na ich podstawie można otrzymać miary korzyści  $|A| - L$  oraz  $1 - nL$ . Jeśli wszystkie oceniane reguły są wyznaczone, możliwy jest również inny sposób normalizacji, polegający na podzieleniu długości każdej reguły przez długość reguły najdłuższej.

Miary  $L$  i  $nL$  przeznaczone są do oceny reguł złożonych z prostych warunków elementarnych. Miary te nie są odpowiednie do oceny reguł zawierających warunki złożone. Złożony warunek elementarny może być zbudowany na podstawie wielu atrybutów [265], w szczególnych przypadkach atrybuty te (dokładniej ich wartości) podlegają potęgowaniu (np. jeśli reguła zawiera elipsoidalne warunki elementarne [64]). Miara (3.26) bierze pod uwagę liczbę atrybutów i ich wystąpienie we wszystkich warunkach elementarnych, z jakich zbudowana jest reguła:

$$L^*(r) = \sum_{a \in A} fa(r). \quad (3.26)$$

Liczba  $fa(r)$  oznacza sumaryczną liczbę wystąpień atrybutu  $a$  w przesłance  $r$ . Zliczane jest każde wystąpienie atrybutu  $a$  w jakimkolwiek warunku elementarnym. Zakłada się również, że złożony warunek elementarny może zawierać wyrażenia  $a^{s/k}$  ( $s, k \in \mathbf{Z}$ ), w takiej sytuacji zliczanych jest  $s+1$  wystąpień atrybutu  $a$ .

Sposób normalizacji miary  $L^*$  nie jest oczywisty, wymaga on nałożenia ograniczeń na maksymalną liczbę warunków elementarnych ( $\max Ec$ ), z jakich może być zbudowana przesłanka, oraz maksymalnej liczby atrybutów tworzących warunek elementarny ( $\max Attr$ ). Jeśli liczby te są ustalone, to  $nL^*(r) = L^*(r)/(\max Ec \cdot \max Attr)$ . W zadanym zbiorze reguł  $RUL$  normalizację  $L^*$  można przeprowadzić, dzieląc  $L^*(r)$  przez  $\max\{L^*(r) : r \in RUL\}$ .

### 3.2.2. Unikalnie pokrywane przykłady pozytywne

W najprostszym przypadku miara informująca o liczbie przykładów pozytywnych, pokrywanych jedynie przez ocenianą regułę, definiowana jest jako miara bezwzględna. Wartość miary równa jest liczbie unikalnie pokrywanych przez nią przykładów pozytywnych  $p^u$ . Innym rozwiązaniem jest odniesienie liczby  $p^u$  do rozmiaru klasy decyzyjnej (3.27) lub do ogólnej liczby przykładów pozytywnych, pokrywanych przez regułę (3.28):

$$U_P(r) = \frac{p^u}{P}, \quad (3.27)$$

$$U_p(r) = \frac{p^u}{p}. \quad (3.28)$$

Obie miary  $U_P$ ,  $U_p$  przyjmują wartości w przedziale  $[0,1]$ . Miary te nie są równoważne ze względu na porządek reguł. Miara (3.27) informuje o bezwzględnym unikalnym pokryciu reguły  $r$ . Miara (3.28) informuje o względnym unikalnym pokryciu reguły  $r$ .

### 3.2.3. Podobieństwo reguł

Rozważania na temat liczby przykładów unikalnie pokrywanych przez reguły mają ścisły związek z tzw. semantycznym podobieństwem reguł [314]. Traktując problem bardziej ogólnie, pomiędzy dwiema regułami możemy rozważyć podobieństwa semantyczne i syntaktyczne [91,314]. Ocena podobieństwa reguł  $r_i$ ,  $r_j$  odbywa się na podstawie informacji zawartych w tablicy kontyngencji (tabela 3.7).

Tabela 3.7

Tablica kontyngencji przeznaczona do badania podobieństwa reguł

		$r_j$		$\Sigma$
		tak	nie	
$r_i$	tak	$a$	$b$	$a + b$
	nie	$c$	$d$	$c + d$
	$\Sigma$	$a + c$	$b + d$	$a + b + c + d$

Tablicę kontyngencji (tabela 3.7) można wykorzystać do obliczenia zarówno syntaktycznego, jak i semantycznego podobieństwa reguł. Badanie podobieństwa syntaktycznego polega na analizie budowy przesłanek. Rozważając podobieństwo syntaktyczne, przez  $a$  oznaczamy liczbę warunków elementarnych, występujących w obu porównywanych regułach,  $b$  oznacza liczbę warunków występujących w regule  $r_i$  i niewystępujących w regule  $r_j$ ,  $c$  to liczba warunków występujących w regule  $r_j$  i niewystępujących w regule  $r_i$ , wreszcie  $d$  oznacza liczbę warunków niewystępujących w żadnej z porównywanych reguł. Podczas badania podobieństwa semantycznego brane są pod uwagę przykłady treningowe, pokrywane przez porównywane reguły. Znaczenie komórek w tabeli 3.7 jest wtedy następujące:  $a$  to liczba przykładów pokrywanych przez obie reguły,  $b$  to liczba przykładów pokrywanych przez regułę  $r_i$  i niepokrywanych przez regułę  $r_j$  itd.

Do obliczania podobieństwa pomiędzy regułami  $r_i$  i  $r_j$  najczęściej używa się symetrycznych miar podobieństwa. W pracy [314] Tsumoto sugeruje zastosowanie do tego celu jednej z siedmiu miar. Miary te to m.in.: współczynnik Jaccarda (3.29), współczynnik korelacji (3.30) oraz miara Kulczyńskiego (3.31):

$$\text{sim}_{\text{Jaccard}}(r_i, r_j) = \frac{a}{a + b + c}, \quad (3.29)$$

$$\text{sim}_{\text{Corr}}(r_i, r_j) = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}, \quad (3.30)$$

$$\text{sim}_{\text{Kulcz}}(r_i, r_j) = \frac{1}{2} \left( \frac{a}{a+b} + \frac{a}{a+c} \right). \quad (3.31)$$

Miary te spełniają warunek  $\text{sim}(r_i, r_j) = \text{sim}(r_j, r_i)$ . Tsumoto przedstawia propozycję wizualizacji podobieństwa i niepodobieństwa reguł za pomocą metody skalowania wielowymiarowego [314]. Na podstawie macierzy podobieństwa (niepodobieństwa) pomiędzy regułami dokonuje się wizualizacji ich położenia na płaszczyźnie. Dla podobieństw syntaktycznego i semantycznego przeprowadza się oddzielne wizualizacje. Jako przykład zastosowania proponowanej metodyki Tsumoto przedstawia wizualizację i analizę zbioru reguł będących wynikiem analizy danych medycznych.

Inną miarę umożliwiającą badanie podobieństwa pomiędzy regułami proponują Gago i Beneto [91]. Bierze ona pod uwagę podobieństwo warunków elementarnych, tworzących porównywane reguły. Jeśli w obu regułach występują warunki zbudowane na podstawie tego samego atrybutu, to sprawdzane jest, w jakim stopniu pokrywają się zakresy tych warunków. Gago i Beneto proponują, aby warunki elementarne, których zakresy pokrywają się w stopniu większym niż 66%, uznać za identyczne. Efektywność proponowanej miary omawiana jest na przykładzie analizy zbioru danych, opisującego klientów banku. Miara proponowana przez Gago i Beneto stosowana jest w algorytmie filtracji, którego celem jest znalezienie reguł najbardziej do siebie niepodobnych.

W pracy [257] autor wraz z Grucią przedstawili symetryczną miarę syntaktycznego podobieństwa reguł (3.32), opisujących geny za pomocą terminów ontologicznych (ang. *Gene Ontology terms – GO terms*):

$$\text{sim}_{\text{GO}}(r_i, r_j) = 1 - \frac{b+c}{a+b+a+c} = \frac{2a}{2a+b+c}. \quad (3.32)$$

Budowa tych reguł jest specyficzna, gdyż w przesłance zawierają one jedynie warunki elementarne o postaci  $t=1$ , co oznacza, że dany termin ontologiczny opisuje geny pokrywające regułę. Miara (3.32) jest równoważna mierze Jaccarda. W definicji miary (3.32) najistotniejszy jest sposób zliczania unikalnych warunków elementarnych, zawartych w porównywanych regułach (więcej na ten temat napisano w podrozdziale 6.3).

Na podstawie miary *Jaccarda*, propozycji Gego i Beneto oraz miary (3.32) można zdefiniować symetryczną miarę syntaktycznego podobieństwa reguł (3.33), biorącą pod uwagę stopień pokrywania się zakresów warunków elementarnych porównywanych reguł:

$$\text{sim}_{\text{Flex}}^{\text{syn}}(r_i, r_j) = \frac{\sum_{w \in r_i} \text{sim}_{r_i}^w + \sum_{w \in r_j} \text{sim}_{r_j}^w}{|Ec(r_i)| + |Ec(r_j)|}. \quad (3.33)$$

We wzorze (3.33)  $Ec(r)$  jest zbiorem warunków elementarnych reguły  $r$ , natomiast postać  $\text{sim}_r^w$  uzależniona jest od typu atrybutu, na podstawie którego utworzono warunek  $w$ . Jeśli atrybut ten jest typu ciągłego, to  $\text{sim}_r^w$  definiowane jest zgodnie ze wzorem (3.34):

$$\text{sim}_r^w = \frac{v_{\max}^{\cap w} - v_{\min}^{\cap w}}{v_{\max}^w - v_{\min}^w}. \quad (3.34)$$

We wzorze (3.34)  $v_{\max}^w$  i  $v_{\min}^w$  to odpowiednio wartości maksymalna i minimalna, definiujące zakres warunku  $w$ . Gdy zakres warunku nie jest ograniczony, pod uwagę brana jest maksymalna lub minimalna wartość atrybutu, na podstawie którego warunek ten jest zbudowany. Przez  $v_{\max}^{\cap w}$ ,  $v_{\min}^{\cap w}$  oznaczono odpowiednio wartości maksymalną i minimalną, definiujące największy przedział zawierający się jednocześnie w zakresie warunku  $w$  oraz w zakresie odpowiadającego mu warunku w porównywanej regule. Podczas obliczania  $\text{sim}_r^w$  zakresy warunków traktowane są jako przedziały domknięte.

Dla atrybutów symbolicznych oraz porządkowych  $\text{sim}_r^w$  obliczane jest zgodnie ze wzorem (3.35):

$$\text{sim}_r^w = \frac{|Z^{\cap w}|}{|Z^w|}. \quad (3.35)$$

Przez  $Z^w$  oznaczono zbiór wartości warunku  $w$ , przez  $Z_w^{\cap w}$  – zbiór będący iloczynem zbiorów wartości występujących w warunku  $w$  oraz w odpowiadającym mu warunku w porównywanej regule. W szczególności jeśli w regułach dopuszczalne są jedynie warunki postaci  $w \equiv a = v$ , to  $\text{sim}_r^w \in \{1,0\}$ .

Jeśli porównywane reguły będą składały się jedynie z warunków identycznych lub różnych (o rozłącznych zakresach), miara (3.33) przyjmie postać (3.32). Miara ta jest symetryczna i przyjmuje wartości w przedziale  $[0,1]$ . Wartość 0 oznacza, że reguły zbudowane są z różnych warunków elementarnych, wartość 1 oznacza, że reguły są identyczne.

**Przykład 3.5.** Jeżeli warunek elementarny  $w$  w regule  $r_i$  jest postaci  $a \in [1,2]$  oraz odpowiadający mu warunek w regule  $r_j$  jest postaci  $a \in (1.5,2.1]$ , to  $\text{sim}_{r_i}^w = \frac{2-1.5}{2-1} = 0.5$  oraz  $\text{sim}_{r_j}^w = \frac{2-1.5}{2.1-1.5} = 0.83$ .

Miarę  $\text{sim}_{\text{Flex}}^{\text{syn}}$  można przekształcić do miary niesymetrycznej, obliczającej podobieństwo reguły  $r_i$  do reguły  $r_j$  (3.36):

$${}^a\text{sim}_{\text{Flex}}^{\text{syn}}(r_i, r_j) = \frac{\sum_{w \in r_i} \text{sim}_{r_i}^w}{|\text{attr}(r_i, r_j)|}. \quad (3.36)$$

Przez  $\text{attr}(r_i, r_j)$  oznaczono zbiór atrybutów, na podstawie których zbudowane są warunki elementarne w regułach  $r_i, r_j$ . Obliczając wartość miar  $\text{sim}_{\text{Flex}}^{\text{syn}}(r_i, r_j)$ ,  ${}^a\text{sim}_{\text{Flex}}^{\text{syn}}(r_i, r_j)$ , zakładamy, że jeśli w regule  $r_i$  jest warunek  $w$ , zbudowany na podstawie atrybutu  $a$ , oraz w regule  $r_j$  nie ma warunku zbudowanego na podstawie atrybutu  $a$ , to  $\text{sim}_{r_i}^w = 0$ .

**Przykład 3.6.** Założmy, że dane są dwie reguły:  $r_i \equiv a \in [1,2] \wedge b = 3 \rightarrow d = 3$  oraz  $r_j \equiv a \in [1.5, 2] \wedge b = 3 \wedge c = 4 \rightarrow d = 3$ ; wówczas  ${}^a\text{sim}_{\text{Flex}}^{\text{syn}}(r_i, r_j) = (0.5 + 1)/3 = 0.5$  oraz  ${}^a\text{sim}_{\text{Flex}}^{\text{syn}}(r_j, r_i) = (1 + 1 + 0)/3 = 0.66$ .

Zauważmy, że jeśli reguła  $r_j$  jest specjalizacją reguły  $r_i$ , to  ${}^a\text{sim}_{\text{Flex}}^{\text{syn}}(r_i, r_j) < {}^a\text{sim}_{\text{Flex}}^{\text{syn}}(r_j, r_i)$ . Dodatkowo, jeśli  $r_j$  jest specjalizacją  $r_i$  oraz warunki elementarne  $r_j$  zbudowane są na podstawie takiego samego zbioru atrybutów jak warunki  $r_i$ , to  ${}^a\text{sim}_{\text{Flex}}^{\text{syn}}(r_j, r_i) = 1$ . Dowód tych własności wynika wprost z definicji miary  ${}^a\text{sim}_{\text{Flex}}^{\text{syn}}$ .

Miary  $\text{sim}_{\text{Flex}}^{\text{syn}}$  oraz  ${}^a\text{sim}_{\text{Flex}}^{\text{syn}}$  można w prosty sposób zamienić w miary semantyczne  $\text{sim}_{\text{Flex}}^{\text{sem}}$ ,  ${}^a\text{sim}_{\text{Flex}}^{\text{sem}}$ . Założmy, że warunek  $w$  zbudowany jest na podstawie atrybutu  $a$ . W mierze semantycznej, obliczając  $\text{sim}_r^w$  w mianowniku będziemy zliczać przykłady pokrywające  $w$ , natomiast w liczniku zliczane będą przykłady, których wartość atrybutu  $a$  mieści się przedziale od  $v_{\min}^{\cap w}$  do  $v_{\max}^{\cap w}$  lub należy do zbioru  $Z^{\cap w}$ , przy czym tym razem w obliczeniach pod uwagę brane jest to, czy zakres warunku elementarnego jest przedziałem domkniętym czy otwartym. W przypadku nierównomiernego rozłożenia przykładów w przedziałach pokrywanych przez ciągłe warunki elementarne, miara semantyczna będzie odzwierciedlać podobieństwo reguł lepiej niż miara syntaktyczna, która takiego rozkładu nie uwzględnia.

Obliczając wartości miar podobieństwa  $sim$ , pomiędzy parami wszystkich reguł należących do zbioru  $RUL$  możemy dla każdej reguły zdefiniować miary (3.37) i (3.38), których wartości odzwierciedlają stopień, w jakim dana reguła podobna jest do innych reguł należących do zbioru  $RUL$ :

$$q_{Sim}^{max}(r, RUL) = \max\{sim(r_j, r) : r_j \neq r, r_j \in RUL\}, \quad (3.37)$$

$$q_{Sim}^{avg}(r, RUL) = \frac{1}{|RUL|} \sum_{r_j \in RUL, r_j \neq r} sim(r_j, r). \quad (3.38)$$

We wzorach (3.37) i (3.38)  $sim(r_j, r)$  oznacza dowolną miarę umożliwiającą pomiar podobieństwa pomiędzy regułami. Miara  $q_{Sim}^{max}(r, RUL)$  odzwierciedla maksymalne podobieństwo ocenianej reguły do innych reguł należących do  $RUL$ . Można ją potraktować jako miarę kosztu, gdyż z punktu widzenia selekcji reguł najbardziej unikalnych, a więc niepodobnych do innych reguł, im wartość  $q_{Sim}^{max}$  jest niższa, tym ocena reguły jest wyższa.

Na podstawie (3.37) bądź (3.38) można zdefiniować porządek reguł, który następnie może być użyty w algorytmie filtracji.

Przedstawione miary podobieństwa przeznaczone są do porównywania reguł zbudowanych z prostych warunków elementarnych. Do reguł zawierających złożone warunki elementarne można wprost zastosować przedstawione miary, pod warunkiem, że badane będzie semantyczne podobieństwo reguł. Proponowane miary nie nadają się do badania syntaktycznego podobieństwa. Po pierwsze, większość złożonych warunków elementarnych (np. skośnych) będzie uniklana w obu porównywanych regułach. Po drugie, niemożliwe jest obliczenie  $sim_r^w$  dla złożonych warunków elementarnych, zbudowanych na podstawie atrybutów ciągłych.

### 3.2.4. Równomierność rozkładu pokrywanych przykładów pozytywnych

Obecnie przedstawimy propozycję miary, której zadaniem jest pomiar równomierności rozłożenia przykładów pozytywnych w przedziałach pokrywanych przez ciągłe warunki elementarne. Miara ta w obecnej postaci przeznaczona jest do oceny reguł zbudowanych z prostych warunków elementarnych. Założymy, że dany jest parametr  $k$ , informujący o tym, na ile różnych i parami rozłącznych przedziałów należy podzielić zakres wartości każdego warunku elementarnego, zbudowanego na podstawie dowolnego atrybutu ciągłego. Jeśli reguła zawiera warunek  $a \in Za$  oraz  $a$  jest atrybutem symbolicznym lub porządkowym, to zakres warunku dzielony jest na tyle podzbiorów, ile wartości zawartych jest w zbiorze  $Za$  (dla warunków  $a = v$  liczba ta będzie równa 1). Zakładamy również, że jeśli liczba przykładów pozytywnych  $p$ , pokrywanych przez regułę, jest mniejsza od  $k$ , to  $k := p$ . Dla dowolnego warunku  $w$ , znajdującego się w regule  $r$ , przez  $p_w^j$  oznaczymy liczbę

przykładów pozytywnych, pokrywanych przez  $r$ , które należą do  $j$ -tego elementu podziału. Przez  $\bar{p} = [p/k]$  oznaczmy oczekiwany liczbę przykładów pozytywnych, jaka powinna się znaleźć w każdym elemencie podziału warunku  $w$  przy założeniu równomiernego rozkładu przykładów w tym warunku. Średnie odchylenie bezwzględne (3.39), obliczone dla  $w$  na podstawie informacji o  $p_w^j$  i  $\bar{p}$ , zawiera informację o równomierności rozlożenia przykładów w zakresie tego warunku. Średnia arytmetyczna z odchyleń bezwzględnych (3.40), obliczonych dla wszystkich warunków elementarnych ocenianej reguły, jest miarą równomierności rozlożenia przykładów w pokrywanym przez nią „obszarze” przestrzeni atrybutów:

$$D(w, k) = \sum_{j=1}^k \frac{|p_w^j - \bar{p}|}{k}, \quad (3.39)$$

$$Distr(r) = \sum_{w \in Ec(r)} \frac{|D(w, k)|}{L(r)}. \quad (3.40)$$

Miara  $Distr(r)$  jest miarą kosztu; im mniejsza wartość miary, tym przykłady pozytywne, pokrywające regułę, są równomiernej rozłożone. Przy rozkładzie równomiernym  $Distr(r)=0$  maksymalna wartość miary wynosi  $(k-1)\bar{p} + p - \bar{p}$ . Wartość tę otrzymamy, jeśli wszystkie przykłady skupią się w jednym elemencie każdego z podziałów warunków elementarnych. Jest to oczywiście wartość teoretyczna, gdyż sposób indukcji reguły (w szczególności określania zakresów warunków elementarnych) gwarantuje, że sytuacja taka nigdy nie nastąpi. Wartości miary minimalna i maksymalna są dobrymi punktami odniesienia dla interpretacji wyników oceny konkretnej reguły. Wartość miary (3.40) zależy od wartości parametru  $k$ , która ustalana jest przez użytkownika.

Poza wykorzystaniem miary (3.40) jako kolejnego wskaźnika opisującego jakość reguły można wyobrazić sobie użycie informacji o wartości (3.39) w algorytmie klasyfikacji. Jeśli przykład testowy, pokrywany przez regułę, znajdzie się w obszarze, w którym zagęszczenie przykładów pozytywnych pokrywanych przez tę regułę jest mniejsze od przeciętnego, to siła głosu reguły powinna być zmniejszana. Jeśli przykład testowy znajdzie się w obszarze o większym zagęszczeniu przykładów pozytywnych, siła głosu reguły powinna być zwiększana.

Przedstawiona idea modyfikacji sposobu głosowania jest zgodna z sugestiami Webba [327], mówiącymi o tym, że w klasyfikacji przykładów oraz indukcji reguł należy brać pod uwagę sąsiedztwo przykładów. Sąsiedztwo stanowi podstawę metody klasyfikacji, zwanej metodą  $k$ -najbliższych sąsiadów [101]. Informacja o sąsiedztwie przykładów wykorzystywana jest również przez algorytm indukcji reguł. W jednej z wielu modyfikacji

algorytmu AQ [336] dla atrybutów ciągłych punkt odcięcia, określający „początek” i „koniec” zakresu warunku elementarnego, zależy od sąsiedztwa „pierwszego” i „ostatniego” przykładu pozytywnego, pokrywanego przez tworzony warunek. Idea uwzględniania sąsiedztwa przykładów realizowana jest także w algorytmie RIONA [101] oraz w przedstawionej przez autora [263] modyfikacji algorytmu indukcji reguł regresyjnych M5 [237].

### 3.2.5. Miary złożone

Każda miara jakości porządkuje (w sensie matematycznym jest to słaby porządek częściowy) zbiór ocenianych reguł. W praktyce porządek utożsamiany jest z rankingiem. Ranking uzyskujemy przez posortowanie reguł zgodnie z nierośącymi (miary korzyści) lub niemalejącymi (miary kosztu) wartościami miary.

Ocena reguł ze względu na kilka kryteriów jednocześnie wykonywana jest najczęściej przez użytkownika, który na podstawie rankingów definiowanych przez różne, także subiektywne, miary jakości dokonuje wyboru reguł najbardziej go interesujących.

Założmy, że w zbiorze miar  $\{q_1, q_2, \dots, q_l\}$  zdefiniowano porządek  $\langle q_{i1}, q_{i2}, \dots, q_{il} \rangle$ , odzwierciedlający ważności miar dla całkowitej oceny reguł. Na podstawie  $\langle q_{i1}, q_{i2}, \dots, q_{il} \rangle$  możliwe jest ułożenie reguł w porządku leksykograficznym. W praktyce porządek ten tworzony jest jedynie na podstawie reguł spełniających wymogi minimalnej jakości ze względu na każdą z miar należących do zbioru  $\{q_1, q_2, \dots, q_l\}$ .

Stosowanie kilku miar jakości w algorytmach indukcji lub filtracji zbiorów reguł polega na zdefiniowaniu miary złożonej, biorącej pod uwagę jednocześnie wszystkie rozważane kryteria oceny. Miary złożone definiowane są zazwyczaj jako suma lub iloczyn wartości miar składowych (3.41), (3.42):

$$q_{Compl}^{add}(r) = \sum_{i=1}^l \alpha_i q_i(r), \quad (3.41)$$

$$q_{Compl}^{prod}(r) = \prod_{i=1}^l q_i(r). \quad (3.42)$$

We wzorze (3.41)  $\alpha_i \in [0,1]$  dla każdego  $i \in \{1, 2, \dots, l\}$ . Każde  $\alpha_i$  jest wagą odzwierciedlającą istotność miary składowej  $q_i$ . Na wartości wag można nałożyć dodatkowe ograniczenie, mówiące o tym, że ich suma powinna być równa 1.

Definiując miarę złożoną, należy znormalizować tworzące ją miary składowe. Bez normalizacji wartości jednych miar mogą dominować wartości innych miar składowych. Wynikowa ocena będzie wtedy zdominowana przez ocenę miary przyjmującej największe wartości. Dla ustalonej klasy decyzyjnej miara monotoniczna osiąga wartość maksymalną,

gdy spełnione zostaną warunki  $p = P$  oraz  $n = 0$ . Wartość minimalną otrzymamy, gdy  $p = 0$ ,  $n = N$ . W mierze złożonej wszystkie miary składowe powinny być albo miarami korzyści albo miarami kosztu. Miara złożona  $q_{Compl}$  powinna także spełniać warunek monotoniczności. Zakładając, że wszystkie miary składowe  $\{q_1, q_2, \dots, q_l\}$  są miarami korzyści, warunek ten można zapisać jako:

$$\forall r_1, r_2 \forall i \in \{1, 2, \dots, l\} - \{j\} ((q_i(r_1) = q_i(r_2) \wedge q_j(r_1) \leq q_j(r_2)) \Rightarrow q_{Compl}(r_1) \leq q_{Compl}(r_2)).$$

Inny sposób złożonej oceny reguł polega na wykorzystaniu informacji o miejscu danej reguły w rankingach definiowanych przez miary składowe. Ocena reguły nie jest wtedy bezpośrednio uzależniona od bezwzględnych wartości jej ocen cząstkowych, ale od zajmowanego przez nią miejsca w rankingach. Zakładając, że najlepsza reguła zajmuje miejsce o numerze 1, najgorsza reguła zajmuje miejsce o najwyższym numerze oraz reguły nieroróżnicialne zajmują to samo miejsce w rankingu, możliwe jest podanie różnych konkretyzacji złożonej miary jakości. Poniżej przedstawiono trzy takie konkretyzacje:

- średnie miejsce w rozpatrywanych rankingach (miara kosztu),
- najlepsza pozycja w rozpatrywanych rankingach (miara kosztu),
- najgorsza pozycja w rozpatrywanych rankingach (miara kosztu).

Jeśli ocena odbywa się po zakończeniu etapu indukcji, do selekcji reguł najbardziej interesujących można zastosować metody wielokryterialne. Bayardo i Agrawal [15] wykazali, że dla zbioru reguł wskazujących na identyczną klasę decyzyjną reguły optymalne względem wybranych miar jakości (np. *Laplace*, *RI*, *Lift*) znajdują się w zbiorze reguł Pareto-optymalnych ze względu na wsparcie i dokładność. Autorzy ci dowiedli również, że w zbiorze tym znajdują się reguły optymalne dla dowolnej miary monotonicznej względem wsparcia (przy stałej dokładności) i dokładności (przy stałym wsparciu). Optymalność rozumiana jest jako maksymalna wartość miary przy zadany wsparciu i dokładności.

W pracach Szczęch (Brzezińskiej), Greco i Ślowińskiego [43, 300] do wielokryterialnej oceny reguł użyto wsparcia oraz miar konfirmacji (m.in.  $f$  i  $s$ ), jako semantycznie bardziej użytecznych od dokładności. Autorzy ci rozszerzyli prace Bayardo i Agrawala, wykazując, że dla ustalonych  $P$  i  $N$  oraz przy niezmieniającym się wsparciu występuje monotoniczna zależność pomiędzy dokładnością reguły a wartością dowolnej miary definiowanej na podstawie tablicy kontyngencji i charakteryzującej się własnościami M1–M4. Ponadto, podali warunki, jakie musi spełniać miara charakteryzująca się własnościami M1–M4, aby dla ustalonych  $P$  i  $N$  oraz przy niezmieniającej się dokładności mogła być miarą monotoniczną względem wsparcia. W przywoływanych pracach wykazano również, że zbiory reguł Pareto-optymalnych w przestrzeni wsparcie–dokładność i wsparcie– $f$  są identyczne oraz że są nadzbiorem zbioru Pareto-optymalnego w przestrzeni wsparcie– $s$

(miara  $s$ ). Uzyskane wyniki były podstawą do zdefiniowania schematu eliminacji reguł nieinteresujących.

Wyniki przedstawione przez Bayrodo i Agrawala oraz Szczęch, Greco i Słowińskiego nie oznaczają, że w zbiorze reguł, które nie są Pareto-optymalne, nie można znaleźć reguł interesujących. Nie oznacza to również, że posługując się jedynie regułami Pareto-optimalnymi, osiągniemy dobre wyniki klasyfikacji. W przywoływanych pracach ocena reguł wykonywana jest jedynie z punktu widzenia miar jakości, definiowanych na podstawie tablicy kontyngencji.

Do złożonej oceny reguł stosowane są także metody maszynowego uczenia się [1,3] i wielokryterialnego podejmowania decyzji [178]. Metody te można nazwać nadzorowanymi, gdyż wykorzystują one informacje pozyskane od eksperta. Lenca [178] wraz ze współpracownikami proponuje zastosowanie metody PROMTHEE [36] do wyboru miary atrakcyjności, która w sposób najbardziej zbliżony do użytkownika porządkuje zbiór wybranych przez niego reguł. Wybrana miara stosowana jest do uporządkowania całego, nierzadko bardzo licznego, zbioru reguł.

Metodę, dla której podstawą jest algorytm maszynowego uczenia się, przedstawia Abe wraz ze współpracownikami [1, 3]. W propozycji tej ekspert klasyfikuje prezentowane mu reguły do jednej z trzech kategorii: interesujące, nieinteresujące, niezrozumiałe. Następnie dla reguł tych obliczane są wartości szerokiego spectrum miar atrakcyjności. W ten sposób otrzymujemy zbiór metadanych, w których każdy przykład (reguła) opisany jest wektorem wartości miar atrakcyjności, a atrybut decyzyjny wskazuje, do jakiej kategorii należy reguła. W zbiorze metadanych uruchamiany jest algorytm uczenia. Klasyfikator będący efektem uczenia jest następnie stosowany do określania, czy nowe reguły, wyznaczone na podstawie innych zbiorów danych, są interesujące czy nie. Autorzy stosowali metodę do oceny atrakcyjności reguł opisujących problemy diagnostyki medycznej. Proponowana metoda pozwalała na wskazywanie reguł interesujących z 72% dokładnością. Nie pozwala ona jednak na utworzenie rankingu reguł.

W dalszej części rozdziału przedstawiono propozycję wielokryterialnej, nadzorowanej przez użytkownika oceny reguł. Metoda pozwala na uporządkowanie zbioru reguł zgodnie z wartościami funkcji użyteczności, która estymowana jest na podstawie rankingu reguł, utworzonego przez użytkownika. W wielokryterialnym podejmowaniu decyzji metoda estymacji funkcji użyteczności znana jest pod nazwą UTA [141, 269]. Jest ona jedną z pierwszych metod stosowanych w wielokryterialnym wspomaganiu decyzji. W metodzie tej funkcja użyteczności dokonuje agregacji wartości wszystkich cząstkowych kryteriów oceny do jednej globalnej wartości. Wartość ta pozwala ocenić użyteczność każdego z rozważanych wariantów decyzyjnych, tym samym możliwe jest uporządkowanie zbioru wariantów od najlepszego(-ych) do najgorszego(-ych). Podstawą do estymacji funkcji użyteczności jest

uporządkowany podzbiór zbioru wariantów decyzyjnych. Porządek ten definiowany jest przez decydenta (eksperta) na podstawie jego subiektywnych odczuć. Ocena wykonywana przez eksperta musi mieć związek z wartościami ocen cząstkowych, jakie uzyskują warianty decyzyjne. Zakłada się, że ocena eksperta spełnia warunki monotoniczności, tzn. nie występuje sytuacja, w której wariant  $d1$ , o wartościach ocen cząstkowych lepszych lub identycznych z wariantem  $d2$ , jest przez eksperta oceniany gorzej niż  $d2$ .

Przenieśmy metodę UTA do interesującej nas dziedziny zastosowania. Założymy, że zbiór kryteriów będziemy utożsamiać ze zbiorem dowolnych miar jakości  $\{q_1, q_2, \dots, q_l\}$ . Zbiór wariantów decyzyjnych będziemy utożsamiać ze zbiorem reguł  $RUL$ . W szczególnym przypadku w zbiorze  $RUL$  mogą znajdować się jedynie reguły wskazujące na identyczną klasę decyzyjną.

Miary jakości spełniają warunek (3.43):

$$\begin{cases} q_i(r_1) > q_i(r_2) \Leftrightarrow r_1 \succ r_2 \\ q_i(r_1) = q_i(r_2) \Leftrightarrow r_1 \equiv r_2 \end{cases} \quad (3.43)$$

Zapis  $r_1 \succ r_2$  oznacza, że reguła  $r_1$  jest lepsza niż reguła  $r_2$  ( $r_1$  jest preferowana nad  $r_2$ ). Zapis  $r_1 \equiv r_2$  oznacza, że reguła  $r_1$  nie jest ani lepsza, ani gorsza od  $r_2$  (reguły są neutralne względem siebie). Zakres wartości każdej mocy  $q_i$  w zbiorze  $RUL$  oznaczamy jako  $[vq_{i*}, vq_i^*]$ . Ze zbioru reguł wybrano zbiór referencyjny  $RUL_R \subset RUL$ , który uporządkowany zostanie przez eksperta. W zbiorze tym mogą znajdować się reguły, pomiędzy którymi zachodzi relacja  $\equiv$ . Uporządkowanie zbioru referencyjnego nazwiemy porządkiem referencyjnym.

Z każdą regułą  $r \in RUL$  można związać wektor wartości miar  $\mathbf{q}^r = (vq_1^r, vq_2^r, \dots, vq_l^r)$ , gdzie  $vq_i^r = q_i(r)$ ,  $i \in \{1, 2, \dots, l\}$ .

Metoda UTA poszukuje addytywnej funkcji użyteczności o postaci (3.44):

$$u(\mathbf{q}^r) = \sum_{i=1}^l \alpha_i u_i(vq_i^r), \quad (3.44)$$

spełniającej ograniczenia (3.45):

$$\begin{cases} \sum_{i=1}^l \alpha_i = 1 \\ u_i(vq_{i*}) = 0 \quad u_i(vq_i^*) = 1 \quad \forall i \in \{1, 2, \dots, l\} \end{cases} \quad (3.45)$$

Funkcje  $u_i$  nazywane są cząstkowymi funkcjami użyteczności i są one znormalizowane. Funkcję (3.44) i ograniczenia (3.45) można zapisać równoważnie jako (3.46) i (3.47) [269]:

$$u(\mathbf{q}^r) = \sum_{i=1}^l u_i(vq_i^r), \quad (3.46)$$

$$\begin{cases} \sum_{i=1}^l u_i(vq_i^*) = 1 \\ u_i(vq_{i*}) = 0 \quad \forall i \in \{1, 2, \dots, l\}. \end{cases} \quad (3.47)$$

Globalna funkcja użyteczności oraz funkcje cząstkowe spełniają warunek monotoniczności (3.48):

$$\begin{cases} u(\mathbf{q}^{r1}) > u(\mathbf{q}^{r2}) \Leftrightarrow r_1 \succ r_2 \\ u(\mathbf{q}^{r1}) = u(\mathbf{q}^{r2}) \Leftrightarrow r_1 \equiv r_2. \end{cases} \quad (3.48)$$

We wzorze (3.48)  $\mathbf{q}^{ri}$  jest wektorem wartości miar obliczonych dla reguły  $r_i$ . Warunek (3.48) oznacza, że relacje  $\succ$  oraz  $\equiv$  są przechodnie. Dla relacji  $\equiv$  oraz dowolnych reguł  $r_1, r_2, r_3$  mamy zatem  $r_1 \equiv r_2 \wedge r_2 \equiv r_3 \Rightarrow r_1 \equiv r_3$  oraz  $r \equiv r$ , a także  $r_1 \equiv r_2 \Rightarrow r_2 \equiv r_1$ , co oznacza, że  $\equiv$  jest relacją równoważności.

Biorąc pod uwagę (3.46), (3.47) oraz warunki monotoniczności (3.48), globalną ocenę każdej reguły zawartej w  $RUL_R$  można wyrazić jako (3.49):

$$u'(\mathbf{q}^r) = \sum_{i=1}^l u_i(vq_i^r) + \sigma(r), \quad (3.49)$$

gdzie  $\sigma(r)$  jest potencjalnie popełnionym błędem ze względu na rzeczywistą wartość  $u(\mathbf{q}^r)$ .

W metodzie UTA wymaga się, aby cząstkowe funkcje użyteczności były przedziałami liniowymi. Miary jakości  $q_i$  nie muszą być liniowe, aby cząstkowe funkcje użyteczności  $u_i$  były odcinkami liniowymi; stosuje się interpolację liniową. Zakres wartości każdej miary  $q_i$  dzielony jest na  $t_i - 1$  równych przedziałów (wartości  $t_i$  ustalane są przez użytkownika). Jeśli wartości miary  $q_i$  są liczbami całkowitymi i w zbiorze  $RUL$  wartości tych jest niewiele, to  $t_i = |V_{q_i}|$ , gdzie  $V_{q_i}$  jest zbiorem wartości miary  $q_i$  w  $RUL$ .

Wartość cząstkowej funkcji użyteczności na końcu każdego przedziału można wyznaczyć na podstawie wzoru (3.50):

$$vq_i^j = vq_{i*} + \frac{j-1}{t_i-1}(vq_i^* - vq_{i*}), \quad \forall j \in \{1, 2, \dots, t_i\}. \quad (3.50)$$

Wartość cząstkowej funkcji użyteczności  $u_i$  jest dla każdego  $q_i(r) = vq_i^r \in [vq_i^j, vq_i^{j+1}]$  obliczana zgodnie z (3.51):

$$u_i(q_i(r)) = u_i(vq_i^r) = u_i(vq_i^j) + \frac{vq_i^r - vq_i^j}{vq_i^{j+1} - vq_i^j} [u_i(vq_i^{j+1}) - u_i(vq_i^j)]. \quad (3.51)$$

Zakładamy, że w zbiorze  $RUL_R$  reguły ponumerowane są w taki sposób, aby  $r_1$  była regułą najlepszą, a  $r_m$  ( $m = |RUL_R|$ ) była regułą najgorszą. Oznacza to, że dla każdej pary uporządkowanej  $\langle r_k, r_{k+1} \rangle$  spełnione jest albo  $r_k \succ r_{k+1}$ , albo  $r_k \equiv r_{k+1}$ . Wobec przyjętego

założenia dla każdej pary reguł  $\langle r_k, r_{k+1} \rangle$ , wartość  $\Delta(r_k, r_{k+1}) = u'(\mathbf{q}^{rk}) - u'(\mathbf{q}^{r(k+1)})$  będzie spełniała jeden z następujących warunków (3.52):

$$\begin{cases} \Delta(r_k, r_{k+1}) \geq \delta & \Leftrightarrow r_k \succ r_{k+1} \\ \Delta(r_k, r_{k+1}) = 0 & \Leftrightarrow r_k \equiv r_{k+1}. \end{cases} \quad (3.52)$$

Każda z cząstkowych funkcji użyteczności także powinna spełniać warunek  $u_i(vq_i^{j+1}) - u_i(vq_i^j) \geq s_i$ ,  $\forall j \in \{1, 2, \dots, t_i\}$ ,  $\forall i \in \{1, 2, \dots, l\}$ , gdzie  $s_i \geq 0$  jest tzw. progiem neutralności.

Na podstawie warunków (3.46), (3.47), (3.52) oraz wymagań  $u_i(vq_i^{j+1}) - u_i(vq_i^j) \geq s_i$  cząstkowe funkcje użyteczności estymowane są przez rozwiązywanie następującego problemu programowania liniowego z ograniczeniami (3.53):

$$\left\{ \begin{array}{l} [\min] F = \sum_{r \in RUL_R} \sigma(r) \\ \text{z ograniczeniami} \\ \left\{ \begin{array}{l} \Delta(r_k, r_{k+1}) \geq \delta \Leftrightarrow r_k \succ r_{k+1} \\ \Delta(r_k, r_{k+1}) = 0 \Leftrightarrow r_k \equiv r_{k+1} \end{array} \right\} \\ u_i(vq_i^{j+1}) - u_i(vq_i^j) \geq 0 \\ \sum_{i=1}^l u_i(vq_i^*) = 1 \\ u_i(vq_{i*}) = 0 \end{array} \right. \quad \begin{array}{l} \forall k \\ \forall i, j \\ \forall r \in RUL_R, \forall i, j. \end{array} \quad (3.53)$$

W konstruowaniu macierzy dla programu optymalizacyjnego pomocne będzie użycie następujących zależności [269]:  $w_{ij} = u_i(vq_i^{j+1}) - u_i(vq_i^j)$  oraz  $u_i(vq_i^1) = 0$ ,  $u_i(vq_i^j) = \sum_{t=1}^{j-1} w_{it}$  dla każdego  $i = 1, 2, \dots, l$  oraz  $j = 2, 3, \dots, t_i - 1$ .

W pracy [141] Jacquet-Lagreze i Siskos sugerują, że wartość  $\delta$  nie może być większa niż  $Q^{-1}$ , gdzie  $Q$  jest liczbą klas abstrakcji relacji  $\equiv$  w zbiorze  $RUL_R$ .

Jakość otrzymanego rozwiązania sprawdza się, porównując porządek reguł, zdefiniowany w  $RUL_R$  przez eksperta, z porządkiem otrzymanym po estymacji funkcji użyteczności. Jeśli porządkie nie są identyczne, to oblicza się ich podobieństwo. Do tego celu stosuje się współczynnik korelacji  $\tau$  Kendalla. W pracy Siskosa [269] można znaleźć opis uogólnień metody UTA. Przedstawiono tam również sposoby postępowania w sytuacji, gdy rozwiązania problemu (3.53) jest wiele lub gdy istnieje jedynie rozwiązanie quasi-optymalne.

Wartości parametrów  $t_i$  oraz  $\delta$  definiowane są przez użytkownika. Zadanie (3.53) można rozwiązać za pomocą odpowiedniego oprogramowania (np. pakietu obliczeń inżynierskich Matlab). Jeśli otrzymane rozwiązanie jest satysfakcyjne, wykorzystujemy funkcję użyteczności do uporządkowania całego dostępnego zbioru reguł. Jeśli rozwiązanie

nie jest satysfakcjonujące, można próbować: zmienić referencyjny zbiór reguł, zwiększyć wartości parametrów  $t_i$ , zmienić wartość parametru  $\delta$ . Proces estymacji wystarczająco dobrej funkcji użyteczności może wymagać przeprowadzenia kilku prób.

Zastosowanie metody UTA do oceny reguł definiowanych w celu opisu grup genów jest przedmiotem badań prowadzonych aktualnie przez autora wspólnie z Grucą [117]. Podczas prowadzonych prac dla jakości uzyskanych wyników istotna okazała się metodyka wyboru reguł tworzących zbiór referencyjny  $RUL_R$ . Przed zdefiniowaniem problemu (3.53) wartości cząstkowych funkcji użyteczności dzielone są na przedziały. Zbiór referencyjny powinien zawierać reguły zapewniające jak najszerzą reprezentację utworzonych przedziałów. Przykładowo, jeśli zakres wartości funkcji  $q_i$  podzielono na trzy przedziały:  $t_i^1, t_i^2, t_i^3$ , to najlepiej, aby w zbiorze referencyjnym znalazły się reguły  $r_1, r_2, r_3$ , takie że  $q_i(r_1) \in t_i^1$ ,  $q_i(r_2) \in t_i^2$ ,  $q_i(r_3) \in t_i^3$ .

### 3.3. Ocena zbioru reguł

Ocena zdolności klasyfikacyjnych zbioru reguł odbywa się za pomocą miar omówionych w rozdziale 2.6.1. W pracy [342] Yao i Zhou wymieniają miary stosowane do tzw. makrooceny reguł. Ocena w skali makro dotyczy zbiorów reguł. Miary przeznaczone do tego celu zostały przez Yao i Zhou podzielone na dwie kategorie: miary informujące o zdolnościach klasyfikacyjnych oraz miary złożoności. Poza wprowadzonymi w rozdziale 2.6.1 miarami informującymi o zdolnościach klasyfikacyjnych zbioru reguł, dla ustalonej tablicy decyzyjnej  $\mathbf{DT}=(U, A \cup \{d\})$  zbioru reguł  $RUL$  oraz utworzonego na jego podstawie klasyfikatora  $f$  definiowane są precyzja (3.54) oraz ogólność (ang. *generality*) zbioru reguł (3.55):

$$precision(f) = \frac{|\{x \in U : f(x) = d(x)\}|}{|\{x \in U : \exists_{r \in RUL} x \in [\varphi r]\}|}, \quad (3.54)$$

$$generality(RUL) = \frac{|\{x \in U : \exists_{r \in RUL} x \in [\varphi r]\}|}{|U|}. \quad (3.55)$$

We wzorach (3.54) i (3.55)  $[\varphi r]$  oznacza zbiór przykładów pokrywanych przez przesłankę  $r$ . Miara (3.54) jest modyfikacją całkowitej dokładności klasyfikacji (2.13). Informuje ona, ile spośród rozpoznanych przykładów udało się poprawnie przyporządkować do odpowiadających im klas decyzyjnych. Miara (3.54) informuje o tym, jaka jest zdolność rozpoznawania przykładów przez zbiór  $RUL$ , przy czym dokładność klasyfikacji jest dla tej miary nieistotna. Wartości miar (3.54), (3.55) mogą być wyznaczone na podstawie dowolnego zbioru przykładów (testowego, walidacyjnego lub treningowego).

Najprostszymi parametrami informującymi o złożoności zbioru reguł są: liczba reguł (3.56) oraz liczba warunków elementarnych, z których zbudowane są reguły (3.57):

$$nRules(RUL) = |RUL|, \quad (3.56)$$

$$nEc(RUL) = \sum_{r \in RUL} |Ec(r)|. \quad (3.57)$$

Zamiast sumarycznej liczby warunków elementarnych (3.57) częściej podawana jest średnia liczba warunków w regule. Na podstawie średniej arytmetycznej można zdefiniować wiele miar oceniających jakość zbioru reguł. Przykładowo, można obliczyć średnią jakość reguł należących do  $RUL$  (3.58):

$$Q^q(RUL) = \frac{1}{|RUL|} \sum_{r \in RUL} q(r). \quad (3.58)$$

We wzorze (3.58)  $q$  jest jedną z miar jakości, przeznaczonych do oceny reguł. W szczególności możemy oceniać zbiór reguł ze względu na: średnią dokładność, średnie pokrycie, średni poziom statystycznej istotności, średnią liczbę unikalnie pokrywanych przykładów pozytywnych, średnie maksymalne podobieństwo itd. Jeśli oceniany jest zbiór reguł regresyjnych, to dodatkowo można obliczyć średnie odchylenie standardowe oraz średni rozstęp. Przez rozstęp rozumiemy różnicę pomiędzy wartościami maksymalną a minimalną zmiennej decyzyjnej, jaką przewiduje reguła na podstawie treningowego zbioru przykładów.

Jeżeli  $q$  we wzorze (3.58) jest miarą korzyści, wówczas  $Q^q$  również jest miarą korzyści, jeśli  $q$  jest miarą kosztu,  $Q^q$  również jest miarą kosztu.

Do scharakteryzowania zbioru reguł można również użyć miary informującej o liczbie przykładów poprawnie i unikalnie pokrywanych przez wyznaczone reguły:

$$Unique(RUL) = \frac{|\{x \in U : \exists!_{r \in RUL} x \in [r]\}|}{|U|}. \quad (3.59)$$

Miara (3.59) informuje o bezwzględnym unikalnym pokryciu zbioru  $RUL$ . Może ona być traktowana jako odpowiednik miary (3.27), definiowanej dla oceny bezwzględnego unikalnego pokrycia reguły. Wstawiając  $|\{x \in U : \exists_{r \in RUL} x \in [\varphi r]\}|$  do mianownika (3.59), otrzymamy miarę informującą o względnym unikalnym pokryciu zbioru  $RUL$ . Miara ta może być traktowana jako odpowiednik miary (3.28), definiowanej dla oceny względnego unikalnego pokrycia reguły.

Ostatnia z omawianych miar, (3.60), mierzy poziom konfliktów w zbiorze przykładów pokrywanych przez reguły. Wyższe wartości miary oznaczają większą liczbę konfliktów, z tego też powodu należy ją traktować jako miarę kosztu. To, czy konflikty rozstrzygane są poprawnie czy niepoprawnie, nie ma znaczenia dla wartości miary:

$$\text{Conflict}(RUL) = \frac{|\{x \in U : \exists_{r_1, r_2 \in RUL} (x \in [pr1] \wedge x \in [pr2]) \wedge (\psi r_1 \neq \psi r_2)\}|}{|\{x \in U : \exists_{r \in RUL} x \in [\varphi r]\}|}. \quad (3.60)$$

Ocena zbioru reguł, podobnie jak ocena pojedynczej reguły, może być wykonywana równocześnie ze względu na kilka kryteriów jakości. Najczęściej pod uwagę brane są zdolności klasyfikacyjne (*Classification*) i opisowe (*Description*), a do oceny stosuje się jedną z dwóch miar – (3.61) lub (3.62):

$$Q(RUL) = \text{Classification}(RUL)\alpha_1 + \text{Description}(RUL)\alpha_2, \quad (3.61)$$

$$Q(RUL) = \text{Classification}(RUL) \cdot \text{Description}(RUL). \quad (3.62)$$

Wartości parametrów  $\alpha_1, \alpha_2 \in [0,1]$  odzwierciedlają preferencje użytkownika, dotyczące istotności poszczególnych kryteriów. Przede wszystkim można wymagać, aby  $\alpha_1 + \alpha_2 = 1$ .

Każde kryterium, *Classification* i *Description*, można wyrazić za pomocą prostej lub złożonej miary oceniającej [5, 222, 244, 297, 298, 273, 275]. Zdolności klasyfikacyjne mierzone są za pomocą miary (3.54) lub miar wymienianych w rozdziale 2.6.1. Zdolności opisowe wiążą się często ze złożonością i liczbą reguł. Im mniejsze są liczba i złożoność reguł, tym lepsze są ich zdolności opisowe.

W tym miejscu warto wspomnieć, że ocenę zdolności klasyfikacyjnych i złożoności reguł można wykonać, wykorzystując zasadę minimalnej długości kodu [221, 222, 238, 297, 298]. Ocena zbioru reguł *RUL* utożsamiana jest z długością kodu, niezbędną do zakodowania informacji o przynależności każdego przykładu ze zbioru *U* do odpowiadającej mu klasy decyzyjnej. Przy założeniu, że przyporządkowanie to odbywa się za pomocą zbioru reguł *RUL*, miara informująca o jakości tego zbioru określona jest następująco:

$$Q^{MDL}(RUL) = \text{Length}_K(RUL) + \text{Length}_K(U | RUL), \quad (3.63)$$

gdzie:

- $\text{Length}_K(RUL)$  jest długością kodu, niezbędną do zakodowania każdej reguły należącej do *RUL*. Aby zakodować regułę, należy zakodować informację o tym, które atrybuty tworzą występujące w regule warunki elementarne, a następnie zakodować informację o każdym warunku. Przykładowo, zakodowanie warunku  $a \in Za$ , utworzonego na podstawie atrybutu symbolicznego, wymaga kodu o długości  $|Za| \cdot \log(|Va|)$ ;
- $\text{Length}_K(U | RUL)$  jest długością kodu, niezbędną do zakodowania informacji o przynależności każdego przykładu ze zbioru *U* do odpowiedniej klasy decyzyjnej, przy założeniu, że zbiór reguł *RUL* jest znany. W kodowaniu należy wziąć pod uwagę, że:
  - informacja o przykładach poprawnie klasyfikowanych przez zbiór reguł *RUL* nie wymaga kodowania, gdyż przynależność tych przykładów wynika z działania klasyfikatora,

- zakodowania wymaga informacja o poprawnej przynależności przykładów, które zostały sklasyfikowane niepoprawnie; należy zakodować informację potrzebną do wskazania takich przykładów oraz przyporządkowania ich do właściwych klas decyzyjnych,
- zakodowania wymaga informacja o przykładach, dla których klasyfikator nie podejmuje żadnych decyzji (z sytuacją taką możemy mieć do czynienia, gdy klasyfikacja odbywa się jedynie przez reguły pasujące i nie jest stosowana reguła domyślna); należy zakodować informację potrzebną do wskazania takich przykładów oraz przyporządkowania ich do właściwych klas decyzyjnych.

Miara  $Q^{MDL}$  jest miarą kosztu. Efektywność oceny proponowanej przez miarę silnie zależy od przyjętego sposobu kodowania reguł [57, 222, 238, 297, 298, 333]. Zasada minimalnej długości kodu została użyta do nadzorowania procesu indukcji drzew decyzyjnych [238] i reguł decyzyjnych [57, 222, 297, 298]. W szczególności w algorytmie RIPPER [57] wartość miary  $Q^{MDL}$  decyduje o kontynuacji lub zaprzestaniu indukcji kolejnych reguł.

Ocena zbioru reguł, wykonywana przez wszystkie wymienione w niniejszym rozdziale miary, dokonuje się w kontekście ustalonego zbioru przykładów. W szczególności podczas oceny składnika odpowiedzialnego za badanie zdolności klasyfikacyjnych znaczenie ma zastosowana technika oceny eksperymentalnej oraz schemat klasyfikacji.

Porównanie jakości kilku zbiorów reguł może odbywać się przez obliczenie dla każdego z nich wartości jednej z miar wymienionych w niniejszym rozdziale. Należy tutaj zadbać co najmniej o to, aby podczas oceny zdolności klasyfikacyjnych stosować tę samą technikę oceny eksperymentalnej.

Jeśli w ocenie pod uwagę branych jest kilka kryteriów równocześnie i trudne lub niemożliwe jest zawarcie ich w jednej mierze jakości, to proces wyboru najbardziej odpowiedniego regułowego modelu danych nadzorowany jest przez użytkownika. Użytkownik, podobnie jak podczas oceny pojedynczych reguł, na podstawie porządków zbiorów reguł, otrzymanych przez różne (także subiektywne) miary jakości, dokonuje wyboru najbardziej odpowiadającego mu modelu. Przykładowo, jeśli nadziednym celem indukcji jest klasyfikacja oraz użytkownik jest w stanie określić najmniejszą akceptowaną przez niego dokładność klasyfikacji, to spośród wszystkich zbiorów reguł o akceptowanej dokładności klasyfikacji jako najlepszy wybierany jest często najmniej liczny zbiór reguł. W szczególności można posłużyć się testem statystycznym, aby zidentyfikować zbiory reguł, których dokładność nie różni się istotnie od zbioru osiągającego najwyższą dokładność. Następnie do tak zidentyfikowanych zbiorów można stosować kolejne kryteria jakości.

## **4. EFEKTYWNOŚĆ I WŁASNOŚCI WYBRANYCH MIAR OCENY REGUŁ**

W niniejszym rozdziale przedstawiono wyniki badania efektywności miar jakości, definiowanych na podstawie tablicy kontyngencji. W analizie pod uwagę wzięto zdolności klasyfikacyjne i opisowe zbiorów reguł otrzymanych w efekcie użycia miary w pokryciowym algorytmie indukcji reguł. Wyniki badań empirycznych uzupełniono analizą własności i równoważności miar. Na podstawie badań empirycznych określono zbiór miar najbardziej odpowiednich do nadzorowania procesu indukcji w algorytmie pokryciowym. Na podstawie analizy własności miar zidentyfikowano miary najbardziej odpowiednie do oceny zdolności opisowych reguł.

Badania empiryczne nad efektywnością miar definiowanych na podstawie tablicy kontyngencji i przeznaczonych do nadzorowania procesu indukcji reguł prowadzone były m.in. przez Agotnesa [5], Ana i Cercone [8], Bruhę [40, 42], Janssena i Fürnkranza [143]. W większości przywoływanych prac autorzy koncentrują się na analizie całkowitej dokładności klasyfikacji uzyskanych zbiorów reguł. Rzadziej pod uwagę brana jest liczba reguł. W pracach prowadzonych przez autora [254, 261, 264] przeprowadzono analizę najszerzego zestawu miar i zbiorów danych. Miary stosowano zarówno w algorytmie q-ModLEM, jak i w RMatrix. Efektywność miar badano pod kątem całkowitej dokładności klasyfikacji, średniej dokładności klas decyzyjnych oraz złożoności klasyfikatora. Badając złożoność klasyfikatora, pod uwagę brano jego zdolności klasyfikacyjne oraz dokładność i pokrycie tworzących go reguł.

W ostatnich latach podejmowano prace polegające na stosowaniu metauczenia do nadzorowania procesu wyboru miary nadzorującej indukcję reguł. Główną ideą tych badań była chęć zdefiniowania zestawu cech charakteryzujących zbiór danych treningowych i powiązanie wartości tych cech z efektywnością (najczęściej rozumianą jako całkowita dokładność klasyfikacji) miar jakości. Pierwsze prace tego typu przeprowadzili An i Cercone [8], opisując zbiór danych za pomocą następujących cech charakterystycznych, mianowicie liczby: klas decyzyjnych, przykładów, atrybutów symbolicznych, atrybutów numerycznych. Zbiór danych treningowych opisywała również cecha symboliczna, informująca o tym, czy rozkład liczby przykładów reprezentujących klasy decyzyjne jest zrównoważony czy nie. Za pomocą wymienionych wskaźników scharakteryzowano 27 zbiorów danych z bazy UCI [82].

Następnie w zbiorach tych przeprowadzono indukcję reguł, używając do tego celu 12 miar jakości. Efektywność miar badano pod kątem całkowitej dokładności klasyfikacji. Zakres otrzymanych dokładności klasyfikacji podzielono arbitralnie na cztery przedziały: bardzo dobra, dobra, średnia, niedobra. Działania te doprowadziły do zdefiniowania 12 zbiorów metadanych, z których każdy zawierał 27 przykładów. Każdy zbiór związany był z jedną miarą jakości. W zbiorze metadanych każdy przykład reprezentował zbiór z bazy UCI i opisany był przez wartości cech charakterystycznych. Atrybut decyzyjny wskazywał, jaką dokładność otrzymano dla tego przykładu po zastosowaniu ustalonej miary jakości. W każdym ze zbiorów metadanych dokonano indukcji metareguł. Następnie spośród metaregułów wybrano reguły pozwalające określić, kiedy zastosowanie danej miary prowadzi do uzyskania dokładności bardzo dobrej i dobrej, a kiedy średniej i złej. Wyznaczony zbiór metaregułów zastosowano następnie do selekcji najbardziej odpowiedniej miary dla danego zbioru danych treningowych. Zasadniczym problemem w weryfikacji otrzymanych wyników było to, że autorzy weryfikowali efektywność rekomendacji metaregułów na podstawie tych samych 27 zbiorów danych, na podstawie których reguły te utworzono. Wyniki nie były zatem wiarygodne. Eksperymenty przeprowadzone przez Ana i Cercone powtórzył autor, weryfikując użyteczność metaregułów na podstawie niezależnych zbiorów danych. Otrzymane wyniki nie były zadowalające. Duża grupa miar nierekomendowanych przez metareguły pozwalała na utworzenie klasyfikatorów o dokładności wyższej od klasyfikatorów utworzonych na podstawie rekomendacji metaregułów. W związku z tym autor rozszerzył zbiór cech charakterystycznych m.in. o: cechy informujące o korelacji pomiędzy atrybutami warunkowymi a atrybutem decyzyjnym oraz cechę odzwierciedlającą w sposób numeryczny, jak bardzo niezrównoważony jest rozkład klas [253]. Zamiast opisywać każdy zbiór danych, opisywano oddziennie każdą klasę decyzyjną. Uzyskano w ten sposób większe zbiorów metadanych, a rekomendacje metaregułów dotyczyły zastosowania miary do nadzorowania indukcji w konkretnej klasie decyzyjnej. Działania te pozwoliły na poprawienie jakości wyników, w dalszym ciągu jednak rekomendacje metaregułów nie zawsze wskazywały na miarę najbardziej efektywną. Dokładny opis prezentowanego tutaj w skrócie postępowania można znaleźć w [253].

Badania nad opracowaniem schematu metauczenia dla celów doboru miary w pokryciowym algorytmie indukcji reguł prowadzone były także przez Janssena i Fürnkranza [143]. Autorzy z każdą z reguł związali 9 cech (cechy te to m.in.:  $p$ ,  $n$ ,  $P, N, p/(p+n)$ ,  $P/(P+N)$ ) i na tej podstawie próbowali przewidywać rzeczywistą (niezależną od zbioru treningowego) liczbę przykładów pozytywnych  $p_{real}$  i negatywnych  $n_{real}$ , pokrywanych przez regułę. Oszacowanie  $p_{real}$  i  $n_{real}$  pozwoliłoby na ustalenie rzeczywistej dokładności reguły. Próby oszacowania rzeczywistej dokładności reguł generowanych przez algorytm pokryciowy podejmowane były przez Fürnkranza już kilka lat

wcześniej [88]. Janssen i Fürnkranz na podstawie analizy reguł wyznaczonych dla ponad 30 zbiorów danych z bazy UCI próbowali, m.in. metodą regresji liniowej, powiązać wartości  $p_{real}$  i  $n_{real}$  z wartościami wymienionych wcześniej 9 cech. Przewidywane w ten sposób wartości  $p_{real}$  i  $n_{real}$  użyto następnie w algorytmie indukcji reguł. Otrzymana dokładność klasyfikacji była jednak gorsza od dokładności, jaką zapewniało zastosowanie algorytmu RIPPER. Interesujące natomiast było określenie wpływu  $p$ ,  $n$ ,  $P$ ,  $N$ ,  $p/(p+n)$ ,  $P/(P+N)$  itd. na  $p_{real}$  i  $n_{real}$ .

W pracy [143] Janssen i Fürnkranz przedstawili także metodę doboru optymalnych wartości parametrów miar (m.in. dla miar:  $m$ ,  $F$ ,  $Klösgen$ ). Wartości te ustalono za pomocą strategii podobnej do wspinaczki. Globalne, odpowiednie zdaniem autorów dla dowolnego zbioru danych wartości parametrów ustalono na podstawie analizy ponad 30 zbiorów danych z bazy UCI.

Badania teoretyczne, obejmujące analizę monotoniczności i własności miar definiowanych na podstawie tablicy kontyngencji, przeprowadzono jedynie dla pewnej grupy miar zawartych w tabeli 3.5. Badaniem monotoniczności i równoważności miar ze względu na porządek reguł zajmowali się Fürnkranz i Flach [90] oraz Bruha [40]. Badania własności opisowych prowadzone były przeważnie w szerszym kontekście miar oceny atrakcyjności reguł, z tego też powodu wielu miar zwartych w tabeli 3.5 nie brano w tej analizie pod uwagę. Teoretyczna analiza miar obejmuje analizę ich pochodzenia (określenie, co właściwie jest mierzone) oraz analizę ich własności [27, 128, 132, 176, 341]. Na uwagę zasługują tutaj wyniki prac Greco, Szczęch i Słowińskiego, dotyczące analizy własności miar konfirmacji [43, 112, 113, 300].

Analizę podobieństwa miar ze względu na porządek reguł przeprowadzano głównie dla miar atrakcyjności [2, 4, 129, 307, 314, 319]. Aby zidentyfikować grupy miar podobnych, generowano reguły asocjacyjne i porównywano rankingi reguł, otrzymane po zastosowaniu różnych miar. Drogą grupowania danych [319] i/lub analizy wzajemnych korelacji pomiędzy rankingami identyfikowano grupy miar podobnych [2, 4, 136]. Aby uwolnić się od obciążenia związanego z analizą konkretnych zbiorów danych, badania przeprowadzano również na podstawie danych wygenerowanych w sposób syntetyczny, a indukcję reguł powtarzano, zmieniając wartość minimalnego wsparcia, jakim musiały charakteryzować się reguły. Oshaki i Tsumoto wraz ze współpracownikami [215] przeprowadzili analizę podobieństwa miar, porównując rankingi reguł, otrzymane przez miary atrakcyjności, z rankingiem zdefiniowanym przez eksperta. Podobne, nieco szersze badania opisano w [47], gdzie porównywano rankingi reguł klasyfikacyjnych, uzyskane na podstawie analizy 8 zbiorów danych. Reguły oceniano za pomocą 11 miar atrakcyjności. Z każdego zbioru reguł, dla każdej miary wybrano po 9 reguł (3 reguły najwyższej oceniane, 3 oceniane najniżej

oraz 3 znajdujące się pośrodku rankingu). Tak uzyskane rankingi konfrontowane były z rankingami definiowanymi przez ekspertów dziedzinowych.

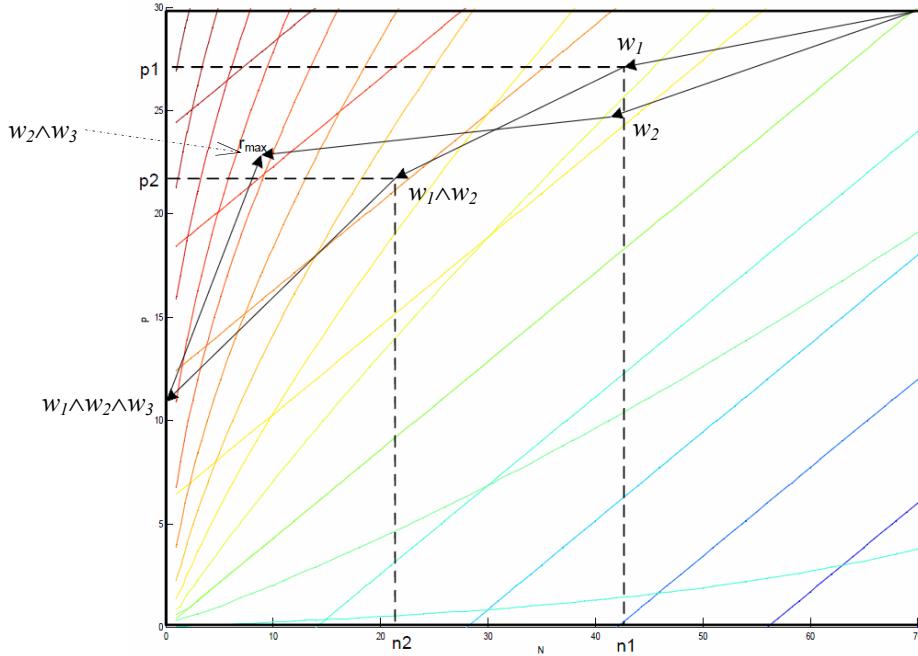
#### 4.1. Rola miary w pokryciowym algorytmie indukcji

Rolę miary jakości w pokryciowym algorytmie indukcji reguł przedstawiono już w rozdziale 2, przy okazji omawiania najpopularniejszych algorytmów oraz podczas opisywania algorytmów q-ModLEM i RMatrix. Powróćmy na chwilę do tej analizy. Na rysunku 4.1 przedstawiono kolejne etapy indukcji reguły. W fazie wzrostu do przesłanki dodawane są kolejno warunki  $w_1$ ,  $w_2$ ,  $w_3$ . W fazie przycinania warunek  $w_1$  okazuje się zbędny. Na rysunku 4.1 tłem dla procesu indukcji są izolinie miar  $RSS$  i  $C2$ . Indukcja reguły  $r_{max}$ , złożonej z warunków  $w_2$ ,  $w_3$ , nadzorowana jest przez miarę  $RSS$ . Z rysunku można odczytać, że do pewnego etapu wzrostu (koniunkcja  $w_1 \wedge w_2$ ) wartość  $RSS$  rośnie, a następnie spada (koniunkcja  $w_1 \wedge w_2 \wedge w_3$ ). W fazie przycinania warunki usuwane są w taki sposób, aby jakość reguły była co najmniej taka jak jakość reguły będącej rezultatem fazy wzrostu. Dla miary  $RSS$  regułą o najwyższej jakości jest reguła, której przesłanka ma postać  $w_2 \wedge w_3$ . Zauważmy, że gdyby proces indukcji nadzorowany był przez miarę  $C2$ , regułą o najwyższej jakości byłaby reguła złożona z warunków  $w_1$ ,  $w_2$ ,  $w_3$ . Reguła ta byłaby również regułą najlepszą, gdyby faza wzrostu nadzorowana była przez  $RSS$ , a faza przycinania – przez  $C2$ .

Do chwili obecnej zakładano, że w fazie wzrostu poszukiwanie kolejnego warunku elementarnego realizowane jest na podstawie oceny jakości reguły rozszerzonej o ten warunek. Jakość ta obliczana jest na całym zbiorze przykładów treningowych. Faza wzrostu może być realizowana również w inny sposób. Pewna grupa algorytmów po dodaniu warunku elementarnego nie ocenia jakości rozszerzonej reguły, ale ocenia, na ile rozszerzona reguła jest lepsza od reguły nierozszerzonej. Miary stosowane w takim podejściu oceniają korzyść rozszerzenia reguły o dany warunek elementarny. W tabeli 3.5 znajdują się miary  $cFoil$  i  $Corr$ , które są bardzo bliskie tej idei. Założymy, że w zbiorze  $P$  przykładów pozytywnych i  $N$  przykładów negatywnych dana jest reguła  $\varphi \wedge w \rightarrow \psi$ , pokrywana przez  $p$  przykładów pozytywnych i  $n$  przykładów negatywnych. Zastępując w  $cFoil$  i  $Corr$  wartości  $P$  i  $N$  przez  $p'$  i  $n'$ , oznaczające odpowiednio liczbę przykładów pozytywnych i negatywnych pokrywanych przez  $\varphi \rightarrow \psi$ , otrzymamy miary realizujące ideę oceny korzyści rozszerzenia  $\varphi \rightarrow \psi$  o warunek  $w$  [90, 143].

W części algorytmów etap oceny warunku elementarnego oddzielony jest od oceny reguły. Ocena warunku dodawanego do  $r$  może być wykonana globalnie w kontekście

jakości reguły rozszerzonej o ten warunek (np. q-ModLEM) lub lokalnie w kontekście jakości tego warunku w zbiorze  $[r]$  przykładów pokrywanych przez  $r$  (np. MODLEM).



Rys. 4.1. Ilustracja faz wzrostu i przycinania reguły ( $P=30, N=70$ ).

Jako tło zastosowano poziomice miar RSS i C2

Fig. 4.1. Illustration of the rule growing and pruning phases ( $P=30, N=70$ ).

Contour lines of the measures RSS and C2 were applied as a background

W pokryciowym algorytmie indukcji reguł miara jakości stosowana jest w fazach wzrostu i przycinania. Na każdym z tych etapów może to być inna miara. Miarę jakości wykorzystuje się także w algorytmie decyzyjnym, aby odzwierciedlić zaufanie do reguł. Dla  $n$  miar jakości możemy zatem przedstawić  $n^3$  różnych sposobów użycia tych miar do utworzenia klasyfikatora regułowego.

W opisanych w niniejszym rozdziale badaniach efektywność miary badano, stosując ją na każdym etapie indukcji reguł i podczas klasyfikacji. W celu indukcji reguł stosowano algorytm q-ModLEM. Algorytm ten, używający miary w fazach wzrostu i przycinania oraz podczas klasyfikacji, oznaczany będzie jako q-ModLEM(MMM). Jako uzupełnienie badań przedstawiono także wyniki stosowania miary jedynie na etapach przycinania i klasyfikacji. W badaniach tych do indukcji reguł użyto algorytmu q-ModLEM, który w fazie wzrostu do oceny warunków elementarnych stosuje entropię warunkową (w taki sam sposób jak w algorytmie MODLEM [283, 284]). Algorytm q-ModLEM, stosujący miarę jedynie w fazie przycinania i podczas klasyfikacji, oznaczany będzie jako q-ModLEM(EMM).

Zastosowanie algorytmu q-ModLEM(MMM) pozwoli na zbadanie, jaki jest rzeczywisty wpływ danej miary jakości na zdolności klasyfikacyjne i złożoność wyznaczanych reguł.

## 4.2. Badanie efektywności miar w pokryciowym algorytmie indukcji reguł

Badania eksperymentalne przeprowadzono na grupie 34 zbiorów danych, pochodzących głównie z repozytorium UCI [82]. Zbiory te będziemy nazywali treningowymi. W tabeli 4.1 przedstawiono charakterystykę wszystkich zbiorów treningowych. Wyniki badań eksperymentalnych zostały wykorzystane m.in. do określenia zbioru miar najbardziej obiecujących oraz zdefiniowania miary złożonej. Efektywność zbioru miar najbardziej obiecujących oraz miary złożonej zostanie w dalszej części rozdziału zweryfikowana na oddzielnej grupie 14 zbiorów danych. Zbiory te będą nazywali testowymi. Ich charakterystykę przytoczono na końcu tabeli 4.1.

Zbiory treningowe reprezentują szeroką gamę problemów klasyfikacyjnych. Badania tak szerokiego spektrum danych powinny zapewnić, że informacje na temat efektywności miar będą mogły zostać przeniesione również do innych problemów związanych z budową klasyfikatorów regułowych.

Opisane poniżej badania stanowią, według najlepszej wiedzy autora, najbardziej wszechstronną empiryczną analizę wpływu miary jakości, użytej przez pokryciowy algorytm indukcji, na jakość wyznaczonego zbioru reguł.

W prezentowanych w dalszej części rozdziału tabelach zastosowano następujące oznaczenia i wskaźniki: *Acc* – całkowita dokładność klasyfikacji, wyrażona w procentach; *BAcc* – średnia dokładność klas decyzyjnych (*balanced accuracy*), wyrażona w procentach; *Rozpoznano* – oznacza, ile procent klasyfikowanych przykładów zostało rozpoznanych (pokrytych) przez reguły biorące udział w klasyfikacji; *Konflikty* – to wyrażona w procentach liczba konfliktów klasyfikacji; *Błędnie* – oznacza, jaki procent spośród wszystkich konfliktów został rozstrzygnięty nieprawidłowo.

Do opisu złożoności zbioru reguł użyto następujących wskaźników: *Liczba* – liczba reguł, *Warunki* – średnia liczba warunków w regule; *U*. – stosunek liczby przykładów pozytywnych, unikniętych pokrywanych przez regułę, do ogólnej liczby pokrywanych przez nią przykładów pozytywnych (w tabelach podawana jest wartość wyrażona w procentach); *prec.* – średnia dokładność reguł; *cov.* – średnie pokrycie reguł. Jako uzupełnienie podano również informację o średnim poziomie statystycznej istotności wyznaczonych reguł (w tabelach oznaczane jako *p<sub>val</sub>*).

Eksperymenty wykonano, stosując 10-krotną warstwową walidację krzyżową. Dla każdej miary eksperymenty wykonano na podstawie identycznych podziałów zbiorów danych. Prezentowane w tabeli 4.2 oraz w pozostałych tabelach wyniki są wartościami średnimi z wartości średnich (średnia z 34 wyników 10-krotnej walidacji krzyżowej).

Charakterystyka zbiorów danych użytych do badań

Tabela 4.1

Nazwa zbioru	Liczba klas	Liczba przykładów	Najliczniejsza klasa [%]	Liczba atrybutów	W tym atrybutów nominalnych
Zbiory treningowe					
anneal	5	898	76.17	38	32
audiology	24	226	25.22	69	69
auto-mpg	3	398	62.56	7	2
autos	6	205	32.68	25	10
balance-scale	3	625	33.33	4	0
breast-cancer	2	286	70.28	9	9
breast-w	2	699	65.52	9	0
car	4	1728	70.02	6	6
cleveland	5	303	54.13	13	7
contact-lenses	3	24	62.50	4	4
credit-g	2	1000	70.00	20	13
cylinder-bands	2	540	57.78	35	18
diabetes	2	768	65.1	8	0
echocardiogram	2	131	67.18	11	2
ecoli	8	336	42.56	7	0
flag	4	194	46.91	28	18
heart-statlog	2	270	55.56	13	0
hepatitis	2	155	79.35	19	13
horse-colic	2	368	63.04	22	15
hungarian-heart-disease	2	294	63.95	13	7
iris	3	150	33.33	4	0
mammographic-masses	2	961	53.69	5	2
mushroom	2	8124	51.8	22	22
prnn-synth	2	250	50.00	2	0
segment	7	2310	14.29	19	0
sick-euthyroid	2	3772	93.88	29	22
sonar	2	208	53.37	60	0
soybean	19	683	13.47	35	35
titanic	2	2201	67.7	3	3
vehicle	4	846	25.77	18	0
vote	2	435	61.38	16	16
wine	3	178	39.89	13	0
yeast	10	1484	31.2	8	0
zoo	7	101	40.59	16	15
Zbiory testowe					
bupa-liver-disorders	2	345	57.97	6	0
credit-a	2	690	55.51	15	9
hayes-roth	3	132	33.33	4	4
heart-c	2	303	54.46	13	7
hypothyroid	4	3772	92.29	29	22
glass	6	214	35.51	9	0
ionosphere	2	351	64.1	34	0
kdd-synthetic-control	6	600	16.66	60	0
kr-vs-kp	2	3196	52.22	36	36
labor	2	57	64.91	16	8
lymph	4	148	54.73	18	15
primary-tumor	21	339	24.78	17	17
splice	3	3190	51.88	60	60
tic-tac-toe	2	958	65.34	9	9

W tabeli 4.2 zamieszczono wyniki algorytmu q-ModLEM(MMM). Miary posortowane są malejąco ze względu na dokładność klasyfikacji otrzymanego na ich podstawie klasyfikatora.

Tabela 4.2  
Charakterystyka klasyfikatorów regułowych, wyznaczonych  
przez algorytm q-ModLEM(MMM)

Miara	Klasyfikacja			Konflikty		Reguły					p <sub>val</sub>
	Acc	BAcc	Rozpoznano	Suma	Błędnie	Liczba	Warunków	U.	prec.	cov.	
<i>C2</i>	<b>82.29</b>	<b>76.32</b>	98.3	33	9	127	2.8	4.7	0.96	0.36	0.039
<i>C1</i>	82.00	<b>76.99</b>	97.2	29	8	151	2.5	4.0	0.98	0.28	0.051
<i>g</i>	<b>81.41</b>	74.19	98.2	32	10	145	2.9	4.6	0.96	0.32	0.008
<i>m</i>	81.28	73.99	99.1	39	13	110	3.4	6.8	0.91	0.44	0.008
<i>Klösgen</i>	81.14	75.26	99.0	44	14	78	3.5	5.8	0.89	0.53	0.010
<i>wLap</i>	81.11	<b>77.88</b>	97.2	29	9	157	2.6	3.8	0.98	0.26	0.016
<i>Laplace</i>	81.08	74.34	97.1	29	9	157	2.6	3.8	0.98	0.26	0.016
<i>J</i>	80.97	75.22	99.2	47	14	71	3.6	6.9	0.87	0.57	0.010
<i>CN2</i>	80.97	75.22	99.2	47	14	71	3.6	6.9	0.87	0.57	0.010
<i>LS</i>	80.75	75.76	96.2	29	8	172	2.4	3.8	0.98	0.19	0.057
<i>E<sup>φ</sup></i>	80.50	74.74	98.6	45	14	104	2.7	4.5	0.95	0.35	0.038
<i>s</i>	80.33	74.94	98.8	51	15	86	2.6	5.2	0.96	0.34	0.041
<i>F</i>	80.02	73.05	99.1	49	16	61	3.6	5.7	0.83	0.60	0.016
<i>Corr</i>	79.86	74.26	99.2	53	16	44	3.7	8.3	0.81	0.68	0.012
<i>YAILS</i>	79.67	71.91	99.1	56	16	50	2.7	5.9	0.92	0.49	0.035
<i>Cohen</i>	79.43	73.8	99.3	52	17	44	3.7	7.8	0.78	0.68	0.011
<i>Gain</i>	79.24	72.32	99.3	57	17	38	3.6	10.3	0.81	0.70	0.012
<i>RIPPER</i>	78.87	74.26	94.3	26	7	165	2.3	4.3	0.98	0.16	0.083
<i>Precision</i>	78.63	73.76	94.3	26	8	165	2.3	4.3	0.98	0.16	0.083
<i>TWS</i>	78.60	71.31	99.3	60	18	40	3.7	10.6	0.76	0.70	0.093
<i>cFoil</i>	78.60	71.31	99.3	60	18	40	3.7	10.6	0.76	0.70	0.093
<i>f</i>	78.48	73.88	94.3	26	8	165	2.3	4.3	0.98	0.16	0.083
<i>WRA</i>	78.31	70.12	99.2	58	18	34	3.6	12.0	0.77	0.74	0.012
<i>Novelty</i>	78.31	70.12	99.3	58	18	34	3.6	12.0	0.77	0.74	0.012
<i>RI</i>	78.31	70.12	99.3	58	18	34	3.6	12.0	0.77	0.74	0.012
<i>RSS</i>	78.18	72.33	99.3	58	18	34	3.6	12.0	0.77	0.74	0.012
<i>Q2</i>	77.86	71.20	99.4	62	19	48	3.4	9.6	0.71	0.58	0.180
<i>MS</i>	77.79	70.48	99.1	59	19	33	3.8	8.2	0.73	0.77	0.032
<i>Lift</i>	77.70	74.89	94.3	26	9	165	2.3	4.3	0.98	0.16	0.083
<i>Accuracy</i>	77.66	65.55	99.0	46	17	59	3.0	6.6	0.86	0.5	0.031
<i>OWS</i>	77.54	74.39	94.3	26	9	165	2.3	4.3	0.98	0.16	0.083
<i>DF</i>	77.51	74.00	94.3	26	9	165	2.3	4.3	0.98	0.16	0.083
<i>Odds</i>	76.93	72.57	99.3	61	19	36	3.0	13.1	0.91	0.50	0.015
<i>RR</i>	67.38	67.19	98.9	65	28	17	3.5	12.0	0.73	0.77	0.023

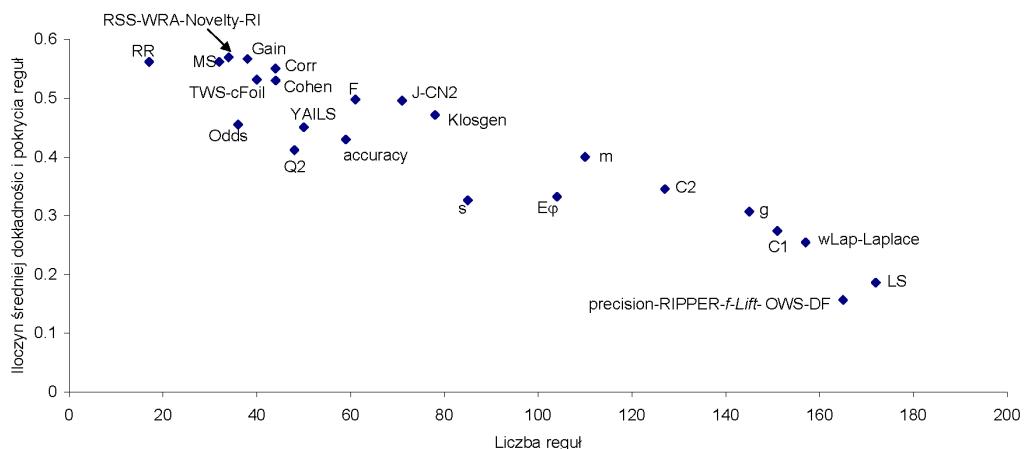
W tabeli 4.2 wyróżniono 3 miary najlepsze ze względu na wartości *Acc* i *BAcc*.

Na podstawie wyników zamieszczonych w tabeli 4.2 można wyciągnąć następujące wnioski:

- część miar prowadzi do uzyskania identycznych wyników, identyczność ta dotyczy rezultatów klasyfikacji oraz charakterystyki reguł (liczby, warunków itd.) lub jedynie charakterystyki reguł (w dalszej części rozdziału przeprowadzona zostanie analiza równoważności tych miar);
- najwyższą dokładność klasyfikacji i średnią dokładność klas decyzyjnych uzyskują (poza kilkoma wyjątkami) miary prowadzące do indukcji większej liczby reguł; średnia dokładność reguły otrzymanych przez takie miary jest większa niż 0.95, średnie pokrycie nie jest większe niż 0.36; jeśli indukcja reguł odbywa się dla celów klasyfikacji, celowe jest stosowanie miar przywiązujących największą wagę do dokładności reguł;

- w klasyfikatorach otrzymanych na podstawie miar prowadzących do indukcji mniejszej liczby bardziej ogólnych reguł (średnia dokładność nie wyższa niż 0.96, średnie pokrycie nie mniejsze niż 0.34) występuje duża liczba konfliktów klasyfikacji; duża liczba tych konfliktów rozstrzygana jest niepoprawnie, co tłumaczy niższą jakość klasyfikatorów;
- im większe jest pokrycie reguł, tym większa jest liczba przykładów pozytywnych, pokrywanych przez nie unikalnie;
- średnia dokładność i średnie pokrycie reguł są skorelowane z liczbą wyznaczanych reguł;
- średnia liczba warunków elementarnych skorelowana jest z liczbą wyznaczanych reguł oraz z ich średnim pokryciem.

Dwa ostatnie spostrzeżenia omówimy nieco dokładniej. Na rysunku 4.2 zamieszczono wykres obrazujący zależność pomiędzy średnią liczbą reguł, generowaną przez każdą z miar, a iloczynem średniej dokładności i średniego pokrycia tych reguł. Widoczna jest wyraźna liniowa zależność. Otrzymano ją na podstawie wyników uśrednionych. Aby przeprowadzić dokładniejszą analizę tej zależności, dla każdej z miar jakości oraz każdego z 340 zbiorów reguł (34 zbiorów danych treningowych, 10-krotna walidacja krzyżowa) obliczono współczynnik korelacji liniowej pomiędzy liczbą generowanych reguł a iloczynem ich średniej dokładności i średniego pokrycia. Uzyskane wyniki zaprezentowano w tabeli 4.3.



Rys. 4.2. Ilustracja zależności pomiędzy liczbą wyznaczanych reguł a iloczynem ich średniej dokładności i średniego pokrycia

Fig. 4.2. Illustration of the dependences between the number of induced rules and the product of their average precision and coverage

Tabela 4.3

Korelacja pomiędzy liczbą reguł, a iloczynem ich średniej dokładności i średniego pokrycia

Max.	Q3	Median	Średnia	Q1	Min.
-0.9879	-0.9446	-0.9012	-0.8682	-0.8436	-0.3860

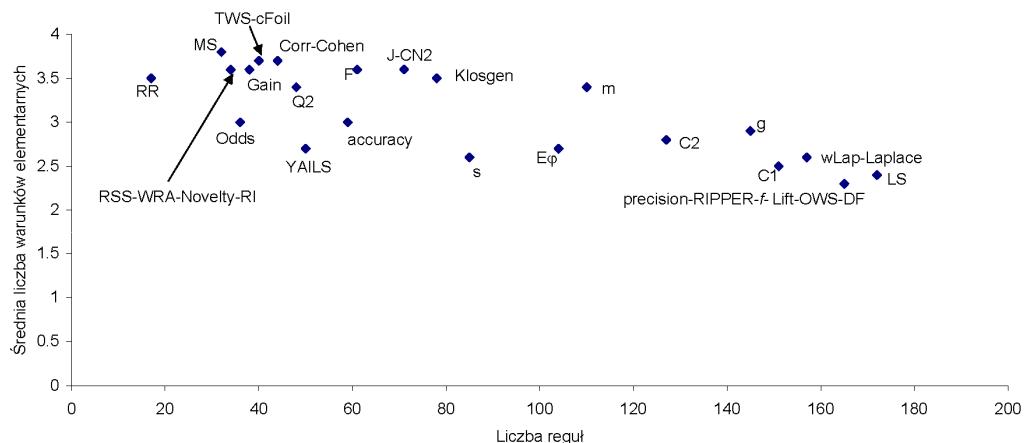
Najmniejszą korelację (-0.3860) odnotowano dla zbioru *titanic*. W większości pozostałych przypadków korelacja była bardzo wysoka. Zauważmy, że dokładność klasyfikacji, średnia dokładność klas decyzyjnych, średnia dokładność i średnie pokrycie reguł wyrażane są za pomocą liczb z przedziału [0,1]. Założymy, że dane są: zbiór reguł *RUL*, zbudowany na jego podstawie klasyfikator  $f$  oraz testowa tablica decyzyjna  $T$ . Przez  $\text{Avgprec}(RUL)$  i  $\text{Avgcov}(RUL)$  oznaczmy odpowiednio średnią dokładność i średnie pokrycie reguł zawartych w zbiorze *RUL*. W myśl zależności ilustrowanej na rysunku 4.2 i w tabeli 4.3 im większa jest wartość  $\text{Avgprec}(RUL) \cdot \text{Avgcov}(RUL)$ , tym mniej reguł znajduje się w *RUL*.

Z punktu widzenia czytelności modelu pożądane jest, aby klasyfikator był złożony z jak najmniejszej liczby reguł. Ponadto, zdolności klasyfikacyjne zbioru reguł będą tym większe, im większa jest wartość  $\text{Acc}(f, T)$ . Chcemy również, żeby klasyfikator nie promował większościowych klas decyzyjnych, w związku z czym zależy nam także na jak największej wartości  $\text{AvgAcc}(f, T)$ . Kryterium oceniające, które bierze pod uwagę wszystkie te wymagania, spełnia miara *AvC* (ang. *Accuracy vs. Complexity*), wyrażona za pomocą wzoru (4.1):

$$\text{AvC}(RUL, f, T) = \text{Avgprec}(RUL) \cdot \text{Avgcov}(RUL) \cdot \text{Acc}(f, T) \cdot \text{AvgAcc}(f, T). \quad (4.1)$$

Miara (4.1) jest prosta i intuicyjnie zrozumiała. Preferuje klasyfikatory o wysokich zdolnościach klasyfikacji, składające się z reguł jak najdokładniejszych i jak najogólniejszych. W przeciwieństwie do innych miar umożliwiających ocenę zbioru reguł miara (4.1) nie wymaga ani definiowania wartości żadnych parametrów, ani kodowania reguł, tak jak jest to np. w przypadku stosowania miar definiowanych na podstawie zasady minimalnej długości kodu [57, 221, 222, 238, 297, 298]. W dalszych badaniach informacja o wartościach miary *AvC* stanowić będzie informację uzupełniającą dla miar *Acc* (2.13) i *AvgAcc* (*BAcc*) (2.26). Miara *AvC* nadaje się do porównania zbiorów reguł otrzymanych za pomocą różnych miar jakości, ale w obrębie jednego ustalonego algorytmu indukcji. Ze względu na kryterium *AvC* najlepszymi miarami okazały się *RSS*, *Corr* i *MS*.

Ostatni z wniosków wynikających z przeprowadzonych eksperymentów dotyczy związku pomiędzy średnią liczbą warunków elementarnych a liczbą reguł. Zależność tę zaprezentowano na rysunku 4.3.



Rys. 4.3. Ilustracja zależności pomiędzy liczbą wyznaczanych reguł a ich średnią liczbą warunków elementarnych

Fig. 4.3. Illustration of the dependences between the number of induced rules and their average number of elementary conditions

Reguły wyznaczone za pomocą miar prowadzących do indukcji dużej liczby dokładnych reguł ( $precision > 0.95$ ) zawierają mniej warunków elementarnych niż reguły otrzymane za pomocą miar promujących reguły ogólniejsze. Oznacza to, że zakresy warunków elementarnych, ustalonych za pomocą takich miar, jak *Cohen*, *Gain*, *J*, *MS*, *RSS*, są szerokie. To z kolei powoduje, że każdy warunek rozważany oddzielnie pokrywa dużą liczbę przykładów, w tym również przykładów negatywnych. Udokładnienie reguł zawierających „szerokie” warunki elementarne wymaga umieszczenia w przesłance większej liczby warunków. Miary ukierunkowane na indukcję reguł dokładnych tworzą warunki z węższymi zakresami, przez co liczba warunków w przesłance jest mniejsza.

#### 4.2.1. Identyfikacja miar najbardziej efektywnych

Wybierając grupę miar najbardziej obiecujących ze względu na całkowitą dokładność klasyfikacji (*Acc*), średnią dokładność klas decyzyjnych (*BAcc*) oraz kryterium *AvC*, wykorzystano wyniki prezentowane w tabeli 4.2 oraz wyniki testów statystycznych, wykonanych dla każdego z 34 zbiorów danych i każdej miary jakości. W tabeli 4.4 przedstawiono liczbę zwycięstw (w) i porażek (l) każdej z miar ze względu na każde z trzech (*Acc*, *BAcc*, *AvC*) kryteriów jakości zbioru reguł. W porównaniach wykorzystano dwustronny test t dla zmiennych zależnych, poziom istotności ustalono na 0.05. Uznawano, że miara przegrywała na danym zbiorze, jeśli istniała co najmniej jedna inna miara, która uzyskała na tym zbiorze statystycznie lepszy wynik. W przeciwnym razie uznawano, że miara zwyciężała na danym zbiorze. Kolejność miar w tabeli 4.4 jest taka sama jak kolejność miar w tabeli 3.5. W tabeli tej wyróżniono po trzy najlepsze miary ze względu na każde z kryteriów *Acc*, *BAcc* i *AvC*.

Tabela 4.4  
Liczba zwycięstw i porażek na 34 treningowych zbiorach danych

Miara	Acc	BAcc	AvC	Miara	Acc	BAcc	AvC
	(w/l)	(w/l)	(w/l)		(w/l)	(w/l)	(w/l)
<i>accuracy</i>	5/29	2/32	3/31	<i>precision</i>	7/27	8/26	0/34
<i>f</i>	5/29	8/26	0/34	<i>Laplace</i>	18/16	11/23	0/34
<i>wLap</i>	16/18	23/11	0/34	<i>RIPPER</i>	6/28	9/25	0/34
<i>Lift</i>	4/30	9/25	0/34	<i>DF</i>	3/31	8/26	0/34
<i>g</i>	16/18	14/20	2/32	<i>OWS</i>	3/31	6/28	0/34
<i>LS</i>	14/20	15/19	0/34	<i>RSS</i>	12/22	11/23	18/16
<i>WRA-Novelty-RI</i>	10/24	8/26	16/18	<i>m</i>	17/17	13/21	2/32
<i>Cohen</i>	11/23	11/23	15/19	<i>MS</i>	10/24	7/27	18/16
<i>F</i>	15/19	10/24	6/28	<i>C1</i>	22/12	22/12	0/34
<i>C2</i>	19/15	16/18	1/33	<i>Klösgen</i>	17/17	15/19	2/32
<i>Corr</i>	9/25	11/23	18/16	<i>s</i>	14/20	14/20	1/33
<i>YAILS</i>	17/17	14/20	3/31	<i>Odds</i>	11/23	11/23	4/30
<i>RR</i>	5/29	5/29	7/27	<i>Q2</i>	10/24	8/26	3/31
<i>TWS-cFoil</i>	10/23	9/25	12/22	<i>CN2-J</i>	14/20	13/21	6/28
<i>Gain</i>	8/26	7/27	17/17	<i>E<sup>ρ</sup></i>	14/20	12/22	1/33

W pierwszym etapie selekcji najbardziej interesujących miar z rozważanego zbioru usunięto miary: *Klösgen*, *m*, *F*. Użyto w nich wartość parametru, którą dobrano na podstawie analizy prawie identycznego zbioru danych (patrz: wspominany na początku rozdziału artykuł Janssena i Fürnkranza [143]). Wyniki otrzymane przez te miary są zatem obciążone sposobem dobrania wartości parametru. Zauważmy, że nawet przy tak dobranych wartościach parametrów miary te nie są miarami najlepszymi ze względu na żadne z rozważanych kryteriów jakości. Dotyczy to zarówno bezwzględnych wartości *Acc*, *BAcc* oraz *AvC*, jak i liczby zwycięstw i porażek. Ponadto wykresy miar zamieszczone w dodatku A sugerują, że w zbiorze pozostałych miar znajdują się miary podobne do *Klösgen*, *m* i *F*. Potwierdzają to również wyniki badań, dotyczące podobieństwa miar (patrz: dalsza część rozdziału).

Drugi etap selekcji polegał na usunięciu miar, które możemy określić jako zdominowane.

**Definicja 4.1.** Miara  $q_2$  dominuje nad miarą  $q_1$  w rodzinie zbiorów danych  $\mathfrak{R}$  oraz ze względu na kryterium jakości zbioru reguł i metodę porównania wyników  $c$  (ozn.  $q_2 \geq_c^{\mathfrak{R}} q_1$ ) wtedy i tylko wtedy, gdy  $\forall A \in \mathfrak{R} \quad (q_2 \geq_c^A q_1)$ .

W definicji 4.1 zapis  $q_2 \geq_c^A q_1$  oznacza, że zbiór reguł, uzyskany przez algorytm *ALG*, używający miary  $q_2$ , osiąga w zbiorze danych  $A$  wynik nie gorszy ze względu na kryterium

jakości i metodę porównania wyników  $c$  niż zbiór reguł otrzymanych przez  $ALG$ , używający miary  $q_1$ .

Przez  $\mathfrak{I}$  oznaczmy grupę 34 zbiorów danych. Przeprowadzone badania porównawcze pozwoliły w rozważanej grupie miar zidentyfikować kilka miar zdominowanych:

- $wLap \geq_{(Acc,test-t,0.05)}^{\mathfrak{I}} Lift; wLap \geq_{(BAcc,test-t,0.05)}^{\mathfrak{I}} Lift; wLap \geq_{(AvC,test-t,0.05)}^{\mathfrak{I}} Lift;$
- $wLap \geq_{(Acc,test-t,0.05)}^{\mathfrak{I}} OWS; wLap \geq_{(BAcc,test-t,0.05)}^{\mathfrak{I}} OWS; wLap \geq_{(AvC,test-t,0.05)}^{\mathfrak{I}} OWS;$
- $Laplace \geq_{(Acc,test-t,0.05)}^{\mathfrak{I}} RIPPER; wLap \geq_{(BAcc,test-t,0.05)}^{\mathfrak{I}} RIPPER; Laplace \geq_{(AvC,test-t,0.05)}^{\mathfrak{I}} RIPPER;$
- $Laplace \geq_{(Acc,test-t,0.05)}^{\mathfrak{I}} precision; wLap \geq_{(BAcc,test-t,0.05)}^{\mathfrak{I}} precision; Laplace \geq_{(AvC,test-t,0.05)}^{\mathfrak{I}} precision.$

Po usunięciu miar zdominowanych wybrano dwie grupy miar najbardziej obiecujących. W pierwszej grupie zawarto miary pozwalające na uzyskanie najwyższych wartości miar  $Acc$ ,  $BAcc$ ,  $AvC$ . Drugą grupę tworzy quasi-minimalny zbiór miar zwyciężających na każdym z 34 zbiorów danych treningowych. Quasi-minimalny zbiór miar został zdefiniowany oddziennie dla każdego kryterium jakości  $Acc$ ,  $BAcc$ ,  $AvC$ .

Pierwsza grupa miar oznaczana będzie jako  $Max$ . Do grupy tej zaliczono miary:  $C1$ ,  $C2$ ,  $Corr$ ,  $g$ ,  $wLap$ ,  $LS$ ,  $MS$ ,  $RSS$ , s. Tworząc grupę  $Max$ , najpierw wybrano trzy miary najlepsze ze względu na całkowitą dokładność klasyfikacji ( $Acc$ ), następnie trzy miary najlepsze ze względu na średnią dokładność klas decyzyjnych ( $BAcc$ ), do tego zestawu dodano także trzy miary najlepsze ze względu na kryterium  $AvC$ . W ten sposób otrzymano zbiór miar  $Max1 = \{C1, C2, Corr, g, wLap, MS, RSS\}$ , nie zapewniał on jednak zwycięstw dla każdego z 34 zbiorów danych treningowych i każdego z 3 kryteriów jakości. Zbiór  $Max1$  rozszerzono o najmniejszy w sensie inkluzji zbiór miar, który zwycięstwa te zapewniał. W ten sposób otrzymano zbiór  $Max2 = Max1 \cup \{LS, RR, s\}$ . Z  $Max2$  usunięto jednak miarę  $RR$ , otrzymując ostatecznie zbiór  $Max = Max2 - \{RR\}$ . Miarę  $RR$  usunięto z tego względu, że klasyfikatory uzyskane na jej podstawie charakteryzuje bardzo niska dokładność klasyfikacji. Usunięcie  $RR$  spowodowało, że grupa  $Max$  nie zwyciężała jedynie na 5 zbiorach danych ze względu na kryterium  $AvC$ .

Quasi-minimalne zbiory miar wybrano, stosując prostą strategię. Dla ustalonego kryterium jakości jako pierwszą wybierano miarę o najmniejszej liczbie porażek w rodzinie 34 zbiorów treningowych. Następnie usuwano z niej wszystkie zbiory, dla których wybrana miara nie przegrywała (zwyciężała lub remisowała z najlepszą miarą). Na zredukowanym zbiorze danych ponownie wybierano miarę zapewniającą najmniejszą liczbę porażek. Cały proces powtarzano tak długo, dopóki nie pokryto całej rodzinie zbiorów treningowych. W ten sposób otrzymano trzy zbiory miar:  $MinAcc$ ,  $MinBAcc$  oraz  $MinAvC$ , które zaprezentowano w tabeli 4.5.

Tabela 4.5  
Grupy miar najbardziej obiecujących

Nazwa zbioru	Miary
<i>Max</i>	<i>C1, C2, Corr, g, LS, MS, RSS, s, wLap</i>
<i>MinAcc</i>	<i>C1, CN2, g, LS, RSS</i>
<i>MinBAcc</i>	<i>CN2, LS, s, wLap</i>
<i>MinAvC</i>	<i>Corr, Gain, MS, Odds, RR, RSS</i>

Nie wszystkie z wybranych w ten sposób quasi-minimalnych zbiorów miar są zbiorami minimalnymi w sensie inkluzyji. Można zidentyfikować kilkanaście minimalnych zbiorów miar, nieprzegrywających na żadnym ze zbiorów treningowych. Zbiory minimalne złożone są z 4 (kryteria jakości *Acc*, *BAcc*) i 6 (kryterium *AvC*) miar jakości. Dla kryterium *Acc* w każdym ze zbiorów minimalnych znajdują się miary *LS* i *RSS*, a dla kryterium *AvC* – miary *MS*, *RR*, *RSS*. Oznacza to, że miary te są jedynymi nieprzegrywającymi miarami na co najmniej jednym z rozważanych zbiorów treningowych.

Ze względu na to, iż zidentyfikowano kilkanaście zbiorów minimalnych, przedstawiony sposób wyboru quasi-minimalnych zbiorów miar wydaje się rozsądnym rozwiązaniem, pozwalającym na wskazanie jednego, zredukowanego zbioru miar dla każdego z rozważanych kryteriów jakości. Wybrane zbiory *MinBAcc*, *MinAvC* są zbiorami minimalnymi w sensie inkluzyji, natomiast zbiór miar *MinAcc* zawiera o jedną miarę więcej niż zbiór minimalny.

Poniżej przedstawiono wyniki analiz statystycznych, mających na celu określenie, czy któryś z miar można uznać za najefektywniejszą. W tym celu, za pomocą testu kolejności par Wilcoxona, porównano parami wszystkie miary. Poziom istotności ustalono na 0.05. Podstawą do porównań były wyniki otrzymane w grupie 34 zbiorów treningowych. Nie przytaczamy tutaj porównań wykonanych za pomocą testu Friedmana, gdyż w połączeniu z testem Nemenyi uzyskujemy bardzo restrykcyjną procedurę porównawczą [95]. Porównanie wszystkich miar wykazywało statystyczne różnice pomiędzy nimi, jednakże test Nemenyi nie wykazywał różnic pomiędzy najlepszymi miarami. Dla kryteriów *Acc* i *BAcc* na CD diagramach tworzyła się grupa 5 najlepszych miar, dla kryterium *AvC* grupa ta zawierała 8 miar.

W kolejnych wierszach tabel 4.6, 4.7, 4.8 umieszczone zostały najlepsze miary ze względu na określone kryterium jakości *Acc*, *BAcc*, *AvC*. Dla ustalonej miary  $q$  w kolumnach tabel 4.6, 4.7, 4.8 umieszczone zostały nazwy wszystkich miar, których wyniki nie różnią się statystycznie od  $q$ . Jeśli komórka  $c_{ij}$  tabeli jest pusta, to pomiędzy miarami z  $i$ -tego wiersza oraz  $j$ -tej kolumny odnotowano statystyczną różnicę (miara  $i$  jest lepsza od  $j$ ). W tabelach nie zamieszczono miar, których wyniki są statystycznie gorsze od każdej z miar, których nazwy pojawiają się w wierszach tabel 4.6, 4.7, 4.8.

Tabela 4.6  
p-wartość testu Wilcoxona pomiędzy miarami najlepszymi  
ze względu na całkowitą dokładność klasyfikacji (*Acc*)

	<i>C1</i>	<i>C2</i>	<i>g</i>	<i>m</i>	<i>Laplace</i>	<i>wLap</i>
<i>C1</i>	-	0.7645	0.4100	0.1510		
<i>C2</i>	0.2388	-	0.1638	0.0541		
<i>g</i>	0.5940	0.8387	-	0.1575	0.2849	0.0561

Tabela 4.7  
p-wartość testu Wilcoxona pomiędzy miarami najlepszymi  
ze względu na średnią dokładność klas decyzyjnych (*BAcc*)

	<i>C1</i>	<i>C2</i>	<i>wLap</i>	<i>LS</i>	<i>s</i>
<i>C1</i>	-	0.0992	0.9703		
<i>C2</i>	0.90025	-	0.9886	0.1367	0.1413
<i>wLap</i>			-		

Tabela 4.8  
p-wartość testu Wilcoxona pomiędzy miarami najlepszymi  
ze względu na *AvC*

	<i>Corr</i>	<i>Cohen</i>	<i>Gain</i>	<i>MS</i>	<i>RSS</i>	<i>WRA</i>
<i>Corr</i>	-	0.1093	0.0783	0.0724	0.7539	0.2858
<i>MS</i>	0.9289	0.8163	0.3977	-	0.9488	0.5585
<i>RSS</i>	0.2493	0.0820	0.0739	0.0522	-	

Przytoczone wyniki pokazują, że miary *C1*, *C2* pozwalają na definiowanie dobrych klasyfikatorów, zarówno jeśli chodzi o całkowitą dokładność klasyfikacji, jak i ze względu na średnią dokładność klas decyzyjnych. Dla tego ostatniego kryterium bezkonkurencyjna jest miara *wLap*, która z punktu widzenia testu Wilcoxona jest lepsza od każdej innej miary. Ze względu na kryterium *AvC* miara *C2* jest statystycznie lepsza od miar *C1*, *g*, *wLap* (p-wartość co najwyżej 0.0003).

Wyniki porównań uzyskane dla kryterium *AvC* pokazały, że pomiędzy miarami *Corr* i *MS* oraz *RSS* i *MS* wystąpiłaby statystycznie istotna różnica, jeśli poziom istotności ustalono by na 0.1. Ponadto miara *Corr* prowadzi do statystycznie lepszych wyników od miar *RSS* i *MS* ze względu na całkowitą dokładność klasyfikacji (*Acc*) oraz średnią dokładność klas decyzyjnych (*BAcc*) (p-wartość co najwyżej 0.006).

#### 4.2.2. Miara złożona i adaptacyjna metoda doboru miary

Na podstawie analizy wyników efektywności miar jakości podjęto próbę poprawienia jakości zbiorów reguł generowanych przez algorytm q-ModLEM(MMM). Badania prowadzono w dwóch kierunkach. Po pierwsze, podjęto próbę zdefiniowania miary złożonej, której wartości byłyby uzależnione od wartości miar zawartych w tabeli 3.5. Po drugie,

sprawdzono, czy metoda adaptacyjnego dobru miary poprawi wyniki algorytmu q-ModLEM(MMM). W metodzie adaptacyjnej wykorzystano zidentyfikowane w poprzednim rozdziale zbiory miar najbardziej obiecujących.

Definiując miarę złożoną, badano różne sposoby łącznia wyników oceny, wykonywanych przez miary zawarte w tabeli 3.5. Najlepsze wyniki w grupie danych treningowych udało się uzyskać dla miary złożonej, będącej średnią arytmetyczną miar składowych. Miarę taką oznaczać będziemy jako *MM* (ang. *Multi-Measure*). Miarę tę zdefiniowano oddzielnie dla każdego z rozważanych kryteriów jakości zbioru reguł. Miary składowe wchodzące w skład *MM* dobrano, stosując strategię wspinaczki.

Początkowo *MM* składała się z jednej miary, która w grupie danych treningowych była najlepsza według zadanego kryterium jakości zbioru reguł (np. ze względu na całkowitą dokładność klasyfikacji była to *C2*). Następnie zbiór miar tworzących *MM* rozszerzano (zgodnie z rankingiem wyników, patrz tabela 3.2), tak aby maksymalizować wartość rozważanego kryterium jakości zbioru reguł. Jeśli dodanie kolejnej miary nie poprawiało wyniku, proces definiowania miary *MM* kończono.

Miarę *MM* definiowano dla podstawowych i normalizowanych postaci miar składowych. Zauważmy, że wartości niektórych miar (np. *LS* lub *Odds*) mogą być znacznie wyższe od wartości innych miar (np. *Corr*, *g*). Aby zrównoważyć wpływ wartości każdej miary na wartości *MM*, przeprowadzono proces normalizacji miar składowych. Normalizacja polegała na podzieleniu oryginalnej wartości miary przez maksymalną wartość, jaką może uzyskać miara w analizowanym zbiorze danych. Monotoniczne miary korzyści (mające własności  $M_p$ ,  $M_n$ ) osiągną wartość maksymalną *max* dla reguły pokrywającej wszystkie przykłady pozytywne i żadnego przykładu negatywnego, a wartość minimalną *min* – dla reguły niepokrywającej żadnego przykładu pozytywnego i pokrywającej wszystkie przykłady negatywne.

Wartość miary po normalizacji wyznaczano na podstawie wzoru  $(vq - \min)(\max - \min)$ , gdzie *vq* jest wartością miary przed normalizacją. Znormalizowaną wersję miary *MM* oznaczamy przez *nMM*.

Składniki miary złożonej oraz zbiory miar najbardziej obiecujących zidentyfikowano na podstawie analizy wyników otrzymanych w grupie 34 zbiorów treningowych. Badając efektywność miar *MM*, *nMM* oraz grup miar najbardziej obiecujących, dodatkowo użyto 14 zbiorów testowych.

Wykonując eksperymenty, sprawdzono, w ilu przypadkach zbiory miar najbardziej obiecujących oraz miary złożone *MM* i *nMM* zapewniały uzyskanie nieprzegrywającego zbioru reguł. Uzyskane wyniki zamieszczono w tabelach 4.9, 4.10, 4.11. Liczba porażek podawana dla grup *Max*, *MinAcc*, *MinBAcc*, *MinAvC* oznacza, na ilu zbiorach danych ani jednej z miar tworzących grupy *Max*, *MinAcc*, *MinBAcc*, *MinAvC* nie udało się wygrać (lub

co najmniej nie przegrać). Wyniki uzyskane na grupie zbiorów treningowych są oczywiście zbyt optymistyczne.

W prezentowanych dalej tabelach obok miary złożonej podano także nazwy tworzących ją miar składowych. W niektórych przypadkach jest to po prostu jedna, najlepsza miara według danego kryterium jakości. W tabelach, dla porównania, przytoczono także wynik uzyskany przez najlepszą miarę.

Tabela 4.9

Liczba porażek dla: najlepszych miar, grup miar najbardziej obiecujących oraz miary złożonej (*Acc*)

Miary	34 zbiory treningowe	14 zbiorów testowych
<i>C2</i>	14	6
<i>g</i>	16	6
<i>MM(C2)</i>	14	6
<i>nMM(C1,Odds,C2,g)</i>	11	5
<i>Max</i>	0	0
<i>MinAcc</i>	0	1

Tabela 4.10

Liczba porażek dla: najlepszych miar, grup miar najbardziej obiecujących oraz miary złożonej (*BAcc*)

Miary	34 zbiory treningowe	14 zbiorów testowych
<i>wLap</i>	11	6
<i>MM(wLap,s,C1,C2)</i>	10	7
<i>nMM(wLap)</i>	11	6
<i>Max</i>	0	1
<i>MinBAcc</i>	0	1

Tabela 4.11

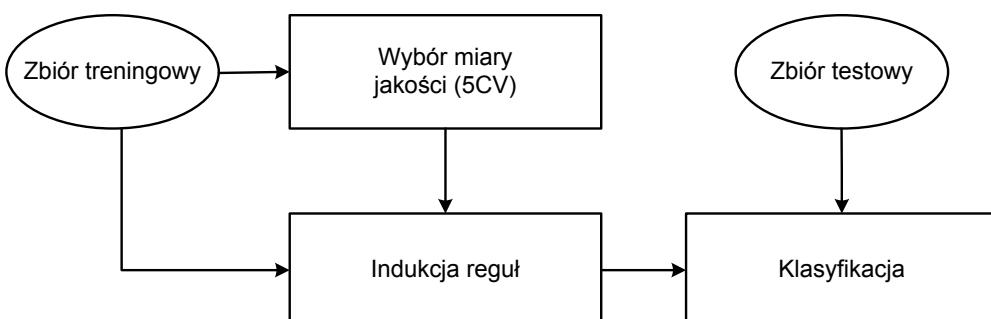
Liczba porażek dla: najlepszych miar, grup miar najbardziej obiecujących oraz miary złożonej (*AvC*)

Miary	34 zbiory treningowe	14 zbiorów testowych
<i>RSS</i>	16	6
<i>MM(RSS, YALIS, TWS)</i>	11	7
<i>nMM(RSS, YALIS, TWS)</i>	12	7
<i>Max</i>	5	1
<i>MinAvC</i>	0	0

Eksperyment, którego wyniki przytoczono w tabelach 4.9–4.11, odpowiada zadaniu analizy danych, w którym analityk testuje różne rozwiązania i do praktycznego zastosowania wybiera najefektywniejsze z nich. Zaprezentowane wyniki pokazują, że zbiór *Max* oraz zbiory quasi-minimalne zawierają miary, których zastosowanie prawie za każdym razem

skutkuje wyznaczeniem najlepszych zbiorów reguł. Miara złożona  $nMM(C1, Odds, C2, g)$ , optymalizowana pod kątem maksymalizacji całkowitej dokładności klasyfikacji, pozwala na uzyskanie nieco mniejszej liczby porażek niż miara  $C2$ . Dotyczy to zarówno zbiorów treningowych, jak i testowych. Ze względu na miarę składową  $Odds$  miara  $nMM(C1, Odds, C2, g)$  wykorzystuje znormalizowane postaci miar składowych.

Kolejny eksperyment polegał na ponownym przeanalizowaniu wszystkich zbiorów danych za pomocą algorytmu, w którym miara jakości dobierana była w każdym kroku walidacji krzyżowej w sposób całkowicie automatyczny. Dobór ten odbywał się drogą wewnętrznej walidacji krzyżowej, przeprowadzonej na aktualnym zbiorze przykładów treningowych. Taki dobór parametrów stosowany jest w wielu systemach analitycznych, np. w programach GhostMiner i Weka [333]. Metoda automatyczna wykorzystuje ustalony zbiór miar oraz jedno z kryteriów jakości ( $Acc$ ,  $BAcc$ ,  $AvC$ ). Dla każdej z miar w trybie  $k$ -krotnej walidacji krzyżowej uruchamiany jest algorytm indukcji reguł, a otrzymane wyniki są uśredniane. Miara uzyskująca najlepszy wynik jest następnie użyta do wyznaczenia reguł w całym zbiorze treningowym. Reguły te stosowane są do klasyfikacji przykładów testowych (rys. 4.4).



Rys. 4.4. Schemat adaptacyjnego doboru miary w pokryciowym algorytmie indukcji reguł

Fig. 4.4. Scheme of adaptive measure selection in the sequential covering rule induction algorithm

Przeprowadzając eksperymenty, do wyboru miary jakości wykorzystano 5-krotną walidację krzyżową. Przyjęta metodyka oznacza, że dla każdego z  $34 \cdot 10$  oraz  $14 \cdot 10$  zbiorów danych miara dobierana była na podstawie wyników średnich, uzyskanych w trybie wewnętrznej 5-krotnej walidacji krzyżowej. Metodę adaptacyjnego doboru miary nazwano *AMS* (ang. *Adaptive Measure Selection*).

Wyniki eksperymentu przedstawiono w tabelach 4.12, 4.13, 4.14. Zapis *AMS-X* wskazuje na wyniki adaptacyjnej wersji algorytmu q-ModLEM(MMM), stosującego miary należące do zbioru  $X$ .

Tabela 4.12

Wyniki szczegółowe dla: najlepszych miar, miary złożonej oraz metody *AMS* – optymalizacja i ocena klasyfikatora ze względu na *Acc*

Miarы	34 zbiorów treningowych				14 zbiorów testowych			
	Acc	BAcc	Porażki	Reguły	Acc	BAcc	Porażki	Reguły
<i>C2</i>	82.29	76.32	14	127	82.10	76.91	6	88
<i>g</i>	81.41	74.19	16	145	82.45	76.91	6	120
<i>MM(C2)</i>	82.29	76.32	14	127	82.10	76.91	6	88
<i>nMM(C1,Odds,C2,g)</i>	82.81	76.47	11	140	82.28	77.53	5	104
<i>AMS-Max</i>	82.90	77.56	6	122	83.75	78.86	3	96
<i>AMS-MinAcc</i>	82.63	77.06	7	121	82.77	77.88	3	106

Tabela 4.13

Wyniki szczegółowe dla: najlepszych miar, miary złożonej oraz metody *AMS* – optymalizacja i ocena klasyfikatora ze względu na *BAcc*

Miarы	34 zbiorów treningowych				14 zbiorów testowych			
	Acc	BAcc	Porażki	Reguły	Acc	BAcc	Porażki	Reguły
<i>wLap</i>	81.11	77.88	11	157	82.02	79.87	6	135
<i>MM(wLap,s,C1,C2)</i>	81.89	77.93	10	145	82.32	79.97	7	115
<i>nMM(wLap)</i>	81.11	77.88	11	157	82.02	79.87	6	135
<i>AMS-Max</i>	82.57	78.25	6	125	83.35	79.90	3	114
<i>AMS-MinBAcc</i>	82.29	78.41	7	127	83.57	80.84	3	114

Tabela 4.14

Wyniki szczegółowe dla: najlepszych miar, miary złożonej oraz metody *AMS* – optymalizacja i ocena klasyfikatora ze względu na *AvC*

Miarы	34 zbiorów treningowych				14 zbiorów testowych			
	Acc	BAcc	Porażki	Reguły	Acc	BAcc	Porażki	Reguły
<i>RSS</i>	78.18	72.33	16	34	77.55	71.37	6	26
<i>MM(RSS,YALIS,TWS)</i>	79.48	72.28	11	38	78.70	71.67	7	31
<i>nMM(RSS,YALIS,TWS)</i>	79.76	72.91	12	40	79.00	72.18	7	33
<i>AMS-Max</i>	79.49	73.47	4	40	79.44	74.90	2	25
<i>AMS-MinAvC</i>	79.62	74.03	3	38	78.16	74.57	3	23

Wszystkie miary złożone i metoda *AMS* osiągają wyższe wartości *Acc*, *BAcc* i *AvC* niż zbioru reguł, otrzymane na podstawie miar zawartych w tabeli 3.5. Metoda *AMS* zdecydowanie obniża również liczbę porażek klasyfikatora. Dotyczy to zarówno 34 zbiorów treningowych, jak i 14 zbiorów testowych.

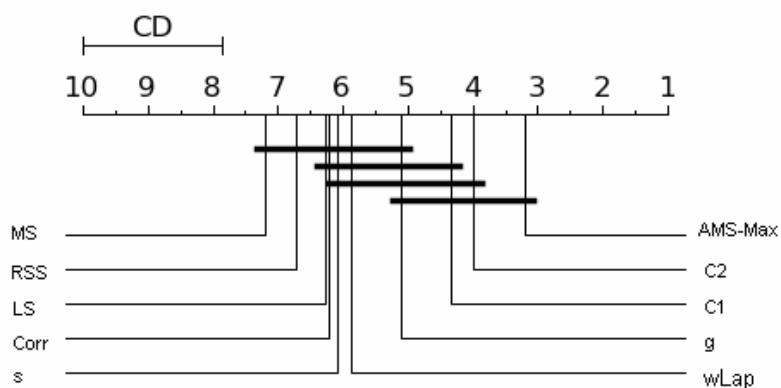
Statystyczną analizę wyników przeprowadzono, porównując algorytmy parami (test kolejności par Wilcoxona) oraz porównując grupę algorytmów (testy Friedmana i Nemeniego). W obu testach poziom istotności ustalono na 0.05. Porównanie miar złożonych z najlepszymi miarami jakości nie wykazało statystycznie istotnych różnic.

Analiza wyników otrzymanych przez adaptacyjną metodę doboru miary jakości wykazuje, że najbardziej efektywna jest grupa miar *Max*. Quasi-minimalne zbiory miar również sprawdzały się bardzo dobrze. Nie odnotowano jednak statystycznych różnic pomiędzy adaptacyjną wersją algorytmu q-ModLEM(MMM), wykorzystującą zbiory quasi-minimalne, a algorytmem q-ModLEM(MMM), stosującym jedną, najlepszą miarę. Wyniki porównań dla metody *AMS* i zbioru miar *Max* zaprezentowano w tabeli 4.15.

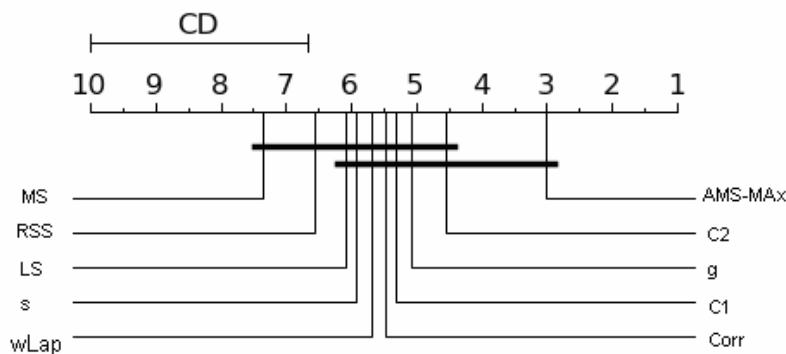
Tabela 4.15  
p-wartość testu Wilcoxona pomiędzy metodą *AMS-Max*  
a najlepszymi miarami jakości

	Acc		BAcc		AvC	
	34 zbiory treningowe	14 zbiorów testowych	34 zbiory treningowe	14 zbiorów testowych	34 zbiory treningowe	14 zbiorów testowych
	<i>C2</i>	<i>g</i>	<i>wLap</i>	<i>wLap</i>	<i>RSS</i>	<i>RSS</i>
<i>AMS-Max</i>	0.107	0.065	0.241	0.501	0.024	0.168

Do porównań statystycznych za pomocą testu Friedmana wybrano klasyfikatory regułowe, uzyskane na podstawie miar zawartych w zbiorze *Max*. W zbiorze tym znajdują się miary najlepsze ze względu na każde z rozważanych przez nas kryteriów jakości (*Acc*, *BAcc*, *AvC*). Test Friedmana wykazał statystyczne różnice pomiędzy klasyfikatorami ze względu na każde kryterium jakości *Acc*, *BAcc*, *AvC*. Kolejnym krokiem było zatem wykonanie testu Nemenyiego. Za każdym razem klasyfikator otrzymany na podstawie metody *AMS-Max* znajdował się na początku rankingu, jednakże w żadnym przypadku test Nemenyiego nie wykazał, że klasyfikator ten jest statystycznie lepszy od klasyfikatora będącego wynikiem zastosowania najlepszej, według danego kryterium jakości, miary. Na rysunkach 4.5, 4.6 zaprezentowano CD diagramy, przedstawiające w postaci graficznej porównanie całkowitej dokładności klasyfikacji rozważanych klasyfikatorów.



Rys. 4.5. Porównanie klasyfikatorów w grupie 34 zbiorów treningowych  
Fig. 4.5. Classifiers comparison on 34 training data sets



Rys. 4.6. Porównanie klasyfikatorów w grupie 14 zbiorów testowych  
Fig. 4.6. Classifiers comparison on 14 testing data sets

#### 4.2.3. Porównanie z innymi algorytmami pokryciowymi

Przedstawione do tej pory wyniki eksperymentów dotyczyły algorytmu q-ModLEM(MMM). W praktycznych zastosowaniach w etapach wzrostu i przycinania często stosowane są różne miary jakości [57, 76, 90, 283, 284]. Poniżej przedstawiono wyniki badania efektywności miar jakości stosowanych jedynie na etapie eliminacji warunków elementarnych oraz w czasie klasyfikacji. W trakcie badań wykorzystano algorytm q-ModLEM(EMM), jako bardzo podobny do opracowanego przez Stefanowskiego algorytmu MODLEM. Szczegółowe wyniki zaprezentowano w tabelach 4.16 i 4.17. W tabeli 4.16 zamieszczono wyniki klasyfikatorów, w których konflikty rozstrzygano drogą głosowania (wyniki te można porównywać z wynikami algorytmu q-ModLEM(MMM), tabela 3.2). Tabela 4.17 zawiera wyniki otrzymane po zastosowaniu schematu największego zaufania.

W tabelach 4.16 i 4.17 komórki z szarym tłem wskazują na 3 miary najlepsze ze względu na algorytm q-ModLEM(MMM) i kryteria *Acc* oraz *BAcc*. Wyniki zapisane pogrubioną czcionką wskazują na 3 miary najlepsze ze względu na algorytm q-ModLEM(EMM) i kryteria *Acc* oraz *BAcc*.

Tabela 4.16

Charakterystyka klasyfikatorów regułowych, wyznaczonych  
przez algorytm q-ModLEM(EMM) – 34 zbiory treningowe – głosowanie

Miara	Klasyfikacja			Konflikty		Reguły				p <sub>val</sub>	
	Acc	BAcc	Rozpoznano	Suma	Błędnie	Liczba	Warunków	U.	prec.		
<i>g</i>	<b>81.80</b>	74.19	99.2	36	11	51	2.9	0.07	0.91	0.36	0.037
<i>C2</i>	<b>81.62</b>	<b>76.21</b>	98.1	24	7	67	3.1	0.08	0.97	0.30	0.092
<i>Laplace</i>	<b>81.32</b>	74.87	98.0	23	7	65	3.1	0.07	0.96	0.28	0.057
<i>Odds</i>	80.75	<b>76.17</b>	98.9	36	11	57	3.0	0.09	0.94	0.34	0.068
<i>Klösgen</i>	80.73	74.51	99.4	44	14	36	2.7	0.08	0.89	0.41	0.069
<i>C1</i>	<b>80.50</b>	<b>75.41</b>	97.5	22	7	73	3.2	0.07	0.98	0.27	0.088
<i>YAILS</i>	80.43	72.94	98.0	27	9	64	3.1	0.08	0.97	0.31	0.089
<i>m</i>	80.32	71.74	99.6	46	15	36	2.7	0.09	0.85	0.42	0.053
<i>E<sup>ρ</sup></i>	80.38	74.80	98.2	27	7	61	2.7	0.08	0.94	0.34	0.067
<i>wLap</i>	80.30	<b>76.49</b>	98.0	23	8	65	3.1	0.07	0.96	0.28	0.057
<i>J</i>	80.19	74.33	99.6	51	16	29	2.6	0.09	0.85	0.46	0.057
<i>CN2</i>	80.19	74.33	99.6	51	16	29	2.6	0.09	0.85	0.46	0.057
<i>LS</i>	80.07	74.66	97.2	20	6	78	3.3	0.07	0.98	0.22	0.085
<i>Corr</i>	80.05	74.45	99.6	54	16	26	2.5	0.10	0.85	0.47	0.072
<i>Cohen</i>	79.80	74.30	99.7	59	18	21	2.2	0.10	0.77	0.54	0.065
<i>s</i>	79.66	74.53	98.0	27	9	66	3.1	0.08	0.96	0.29	0.094
<i>F</i>	79.40	70.61	99.7	63	18	21	2.1	0.09	0.77	0.59	0.070
<i>Gain</i>	78.82	72.04	99.6	57	18	26	2.5	0.11	0.85	0.48	0.075
<i>precision</i>	78.72	74.28	97.2	21	8	79	3.3	0.08	0.98	0.21	0.101
<i>f</i>	78.68	74.41	97.2	21	8	79	3.3	0.08	0.98	0.21	0.101
<i>RIPPER</i>	78.58	73.94	97.2	21	8	79	3.3	0.08	0.98	0.21	0.101
<i>Lift</i>	78.21	74.74	97.2	21	8	79	3.3	0.08	0.98	0.21	0.101
<i>DF</i>	78.03	74.55	97.2	21	8	79	3.3	0.08	0.98	0.21	0.101
<i>OWS</i>	78.03	74.64	97.2	21	8	79	3.3	0.08	0.98	0.21	0.101
<i>RSS</i>	77.52	72.43	99.8	65	20	18	2.0	0.13	0.75	0.58	0.069
<i>MS</i>	77.27	68.07	99.8	74	21	15	1.8	0.12	0.70	0.70	0.107
<i>accuracy</i>	77.16	64.08	99.6	45	18	32	2.4	0.10	0.84	0.45	0.072
<i>WRA</i>	77.11	68.58	99.8	65	21	18	2.0	0.13	0.75	0.58	0.069
<i>Novelty</i>	77.11	68.58	99.8	65	21	18	2.0	0.13	0.75	0.58	0.069
<i>RI</i>	77.11	68.58	99.8	65	21	18	2.0	0.13	0.75	0.58	0.069
<i>TWS</i>	77.11	70.81	99.8	75	21	17	2.0	0.13	0.71	0.58	0.158
<i>cFoil</i>	77.11	70.81	99.8	75	21	17	2.0	0.13	0.71	0.58	0.158
<i>Q2</i>	77.08	72.34	99.8	78	21	19	2.2	0.13	0.64	0.51	0.239
<i>RR</i>	74.25	71.11	99.1	49	19	52	2.8	0.15	0.91	0.39	0.093

Tabela 4.17  
 Dokładność klasyfikacji klasyfikatorów regułowych, wyznaczonych  
 przez q-ModLEM(EMM) – 34 zbiory treningowe –  
 klasyfikacja według największego zaufania

Miara	Klasyfikacja		Miara	Klasyfikacja	
	Acc	BAcc		Acc	BAcc
<i>C2</i>	<b>81.74</b>	<b>76.66</b>	<i>s</i>	78.49	74.91
<i>g</i>	<b>81.18</b>	72.52	<i>wLap</i>	78.33	<b>75.52</b>
<i>Laplace</i>	<b>81.13</b>	74.43	<i>Cohen</i>	78.26	73.38
<i>YAILS</i>	80.75	73.03	<i>Gain</i>	77.05	70.40
<i>C1</i>	80.44	<b>76.17</b>	<i>OWS</i>	77.03	74.18
<i>precision</i>	80.33	75.35	<i>Lift</i>	76.86	74.09
<i>RIPPER</i>	80.33	75.35	<i>DF</i>	76.86	74.09
<i>f</i>	80.28	<b>75.53</b>	<i>MS</i>	76.50	66.25
<i>LS</i>	80.14	74.82	<i>Q2</i>	76.45	73.72
<i>E<sup>ρ</sup></i>	80.08	74.21	<i>TWS</i>	75.49	69.67
<i>F</i>	79.73	70.1	<i>cFoil</i>	75.49	69.67
<i>Odds</i>	79.68	75.57	<i>WRA</i>	75.44	66.77
<i>J</i>	79.06	73.46	<i>Novelty</i>	75.44	66.77
<i>CN2</i>	79.06	73.46	<i>RI</i>	75.44	66.77
<i>Corr</i>	78.93	74.00	<i>RSS</i>	74.87	71.78
<i>m</i>	78.91	67.98	<i>accuracy</i>	74.85	61.05
<i>Klösgen</i>	78.59	73.40	<i>RR</i>	69.95	67.58

Analizy dokładności i pokrycia otrzymanych reguł, a także ich ogólnej liczby oraz liczby unikalnie przez nie pokrywanych przykładów pozytywnych wykazują, że w algorytmie q-ModLEM(EMM) pomiędzy tymi wartościami występują podobne zależności jak w algorytmie q-ModLEM(MMM). Algorytm q-ModLEM(EMM) generuje reguły złożone z mniejszej liczby warunków elementarnych. Wyznaczane reguły są nieco mniej dokładne, ale bardziej ogólne. Łączna liczba reguł generowanych przez q-ModLEM(EMM) jest o ponad połowę mniejsza niż liczba reguł generowanych przez q-ModLEM(MMM). Miary prowadzące do indukcji większej liczby reguł pozwalają podobnie jak w q-ModLEM(MMM), na utworzenie klasyfikatorów o większej dokładności klasyfikacji.

Interesujące może być spostrzeżenie, że w regułach indukowanych za pomocą q-ModLEM(EMM) zachodzi przeciwna niż w regułach generowanych przez q-ModLEM(MMM) zależność pomiędzy liczbą reguł a liczbą warunków elementarnych. Miary przywiązuje większą wagę do pokrycia reguł tym razem prowadzą do indukcji krótszych reguł niż miary kładące większy nacisk na ich dokładność. Jest tak, ponieważ warunki elementarne budowane są w każdym przypadku na podstawie identycznej miary (entropii warunkowej), a uzyskanie reguł o wyższej dokładności wymaga wtedy większej liczby warunków. W algorytmie q-ModLEM(MMM) ocena warunków za pomocą miar kładących większy nacisk na dokładność powoduje szybsze udokładnianie reguły, tym samym liczba warunków jest mniejsza.

Porównanie zdolności klasyfikacyjnych zbiorów reguł wyznaczonych przez q-ModLEM(MMM) i q-ModLEM(EMM) jest niejednoznaczne. W zależności od zastosowanej miary jakości możemy otrzymać statystycznie istotne różnice (test kolejności par Wilcoxona, poziom istotności 0.05). Przykładowo dla miar  $g$  i  $C2$  oraz całkowitej dokładności klasyfikacji nie odnotowano statystycznych różnic, natomiast algorytm q-ModLEM(MMM), stosujący miarę  $C1$ , tworzy statystycznie dokładniejsze klasyfikatory niż stosujący tę miarę q-ModLEM(EMM). Reasumując, gdy popatrzymy na całkowitą dokładność klasyfikacji oraz średnią dokładność klas decyzyjnych, to zauważymy, że w większości przypadków algorytm q-ModLEM(MMM) pozwala na utworzenie lepszych klasyfikatorów niż q-ModLEM(EMM). Nie zawsze jednak różnice te są statystycznie istotne.

Wykorzystanie w q-ModLEM(EMM) miar należących do zbiorów  $Max$ ,  $MinAcc$ ,  $MinBAcc$ ,  $MinAvC$  pozwala na uzyskanie klasyfikatorów zwycięskich ze względu na każde z trzech rozważanych kryteriów jakości. Dotyczy to zarówno grupy zbiorów treningowych, jak i testowych. Jedyny wyjątek stanowi zbiór  $MinBAcc$ . Dla jednego zbioru treningowego nie udało się za pomocą miar należących  $MinBAcc$  uzyskać zwycięskiego klasyfikatora. Dokładniejsza analiza pokazała, że w algorytmie q-ModLEM(EMM) do zapewnienia zwycięstwa na każdym ze zbiorów danych treningowych i testowych wystarczające były następujące zbiory miar:

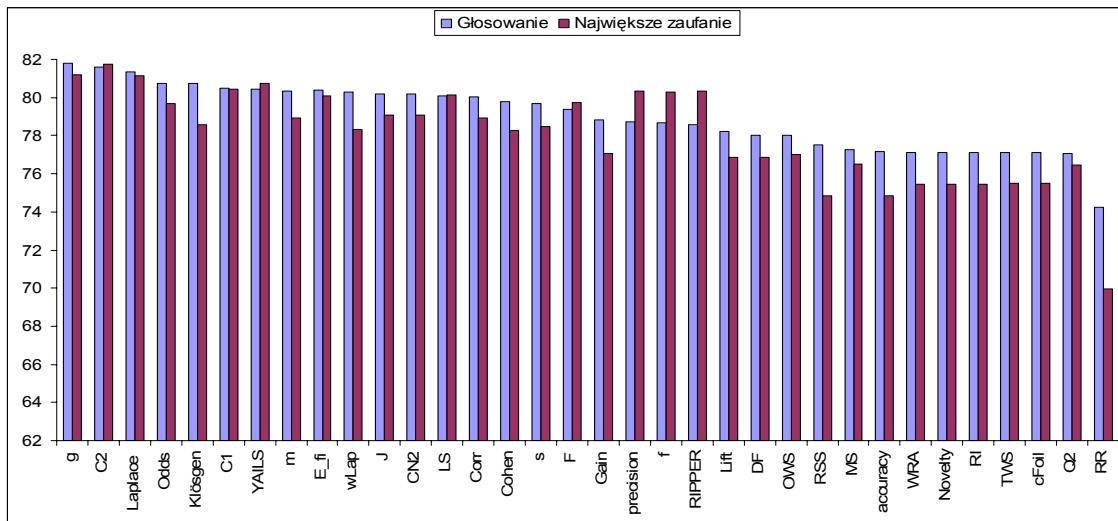
- $\{g, LS, RSS\}$  dla całkowitej dokładności klasyfikacji,
- $\{CN2, wLap, LS\} \cup \{C2\}$  dla średniej dokładności klas decyzyjnych,
- $\{Corr, MS, RSS\}$  dla kryterium  $AvC$ .

Otrzymane wyniki są zgodne z oczekiwaniemi, ponieważ zbiory reguł indukowane przez q-ModLEM(EMM) są mniej zróżnicowane niż zbiory generowane przez q-ModLEM(MMM).

W realizacji praktycznych zadań przydatne mogą być również wyniki porównania różnych schematów klasyfikacji. Tabele 4.16 i 4.17 zawierają szczegółowe informacje o wynikach klasyfikatorów regułowych, utworzonych przez q-ModLEM(EMM) dla dwóch schematów klasyfikacji. Na rysunku 4.7 zaprezentowano różnice całkowitej dokładności klasyfikacji pomiędzy schematami największego zaufania a głosowaniem.

Dla większości miar głosowanie pozwala na uzyskanie większej dokładności klasyfikacji. Dla wielu miar różnice pomiędzy głosowaniem a schematem największego zaufania są statystycznie istotne (test kolejności par Wilcoxona, poziom istotności 0.05). Największe różnice odnotowano dla miar wysoko oceniających reguły o dużym pokryciu (np.  $Gain$ ,  $RR$ ,  $RSS$ ,  $WRA$ ). Jedynie dla miar  $f$ ,  $precision$  i  $RIPPER$  schemat największego zaufania wyraźnie zwycięża ze schematem głosowania. Analiza wpływu mechanizmu rozstrzygania konfliktów klasyfikacji na średnią dokładność klas decyzyjnych nie pozwala na wyciągnięcie jednoznacznego wniosku. W obrębie szczególnie interesującego nas zbioru miar  $Max$

głosowanie zapewnia otrzymanie wyższej średniej dokładności klas decyzyjnych (jedyną wyjątek stanowi miara  $MS$ ).



Rys. 4.7. Porównanie dwóch schematów klasyfikacji: głosowania i największego zaufania  
Fig. 4.7. Comparison of two classification schemes: weighted voting and maximal confidence

Ostatni eksperyment z wykorzystaniem algorytmu q-ModLEM(EMM) polegał na zastosowaniu w nim metody adaptacyjnego doboru miary. Szczegółowe wyniki zaprezentowano w tabeli 4.18.

Tabela 4.18  
Algorytm q-ModLEM(EMM) – wyniki szczegółowe dla najlepszych miar jakości oraz metody adaptacyjnej

Kryterium jakości zbioru reguł	Miara	34 zbiory treningowe				14 zbiorów testowych			
		Acc	BAcc	Porażki	Reguły	Acc	BAcc	Porażki	Reguły
Acc	<i>C2</i>	81.6	76.2	13	66	80.9	76.7	3	43
	<i>g</i>	81.8	74.2	8	51	81.4	74.9	5	38
	<i>AMS-Max</i>	82.5	76.7	4	53	81.9	76.9	0	36
BAcc	<i>wLap</i>	80.3	76.5	10	65	80.0	76.7	4	46
	<i>C2</i>	81.6	76.2	12	66	80.9	76.7	4	43
	<i>AMS-Max</i>	82.0	77.5	5	54	81.5	77.3	0	40
AvC	<i>MS</i>	77.3	68.1	6	15	73.6	64.8	6	15
	<i>RSS</i>	77.5	72.4	19	18	76.8	70.7	1	15
	<i>AMS-Max</i>	78.1	72.0	4	17	76.2	70.5	2	18

Uzyskane wyniki potwierdzają efektywność grupy miar *Max*. Zbiory reguł, budowane na podstawie miar należących do grupy *Max*, charakteryzują się najwyższą jakością. Rezultaty algorytmu q-ModLEM(EMM), w którym zastosowano metodę *AMS*, są – ze względu na całkowitą dokładność klasyfikacji oraz średnią dokładność klas decyzyjnych – statystycznie lepsze (test kolejności par Wilcooxona, poziom istotności 0.05) od wyników uzyskanych przez q-ModLEM(EMM), wykorzystujący miarę *g* (kryterium *Acc*) oraz miarę *C2* (kryterium *BAcc*). Dotyczy to obu grup danych: treningowych i testowych. Natomiast nie

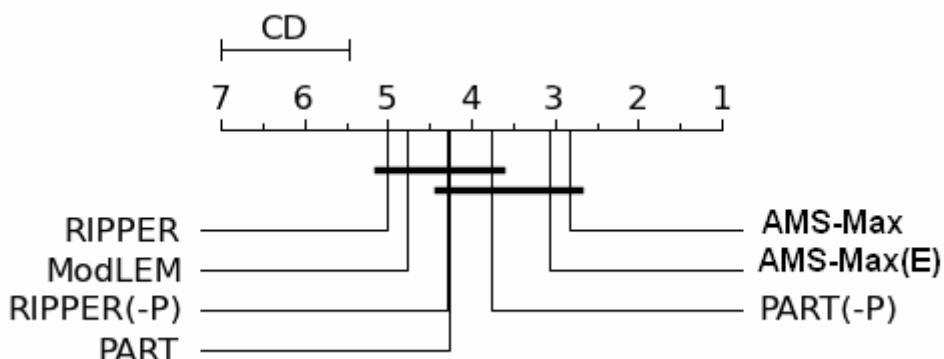
odnotowano statystycznych różnic pomiędzy metodą adaptacyjną a miarą *RSS* ze względu na *AvC*.

Aby podsumować eksperymenty, porównano wyniki klasyfikatorów utworzonych za pomocą adaptacyjnych wersji algorytmów q-ModLEM(MMM) i q-ModLEM(EMM) z innymi pokryciowymi algorytmami indukcji reguł. Do porównań wybrano standardowy algorytm MODLEM (implementacja z pakietu Weka) oraz algorytmy RIPPER i PART. Przebadano dwie konkretyzacje algorytmów RIPPER i PART. Symbol *P* oznacza, że algorytm uruchamiany był z włączoną opcją przycinania. Do eksperymentów wykorzystano pakiet Weka [333]. Wykonano je, stosując 10-krotną warstwową walidację krzyżową. Wyniki zamieszczono w tabeli 4.19. Zapis *AMS-Max* wskazuje na adaptacyjną wersję algorytmu q-ModLEM(MMM), a zapis *AMS-Max(E)* – na adaptacyjną wersję algorytmu q-ModLEM(EMM). W metodzie adaptacyjnej optymalizacja doboru miary przebiegała w kierunku maksymalizacji całkowitej dokładności klasyfikacji.

Tabela 4.19  
Porównanie dokładności klasyfikacji adaptacyjnych wersji algorytmu  
q-ModLEM z innymi algorytmami pokryciowymi

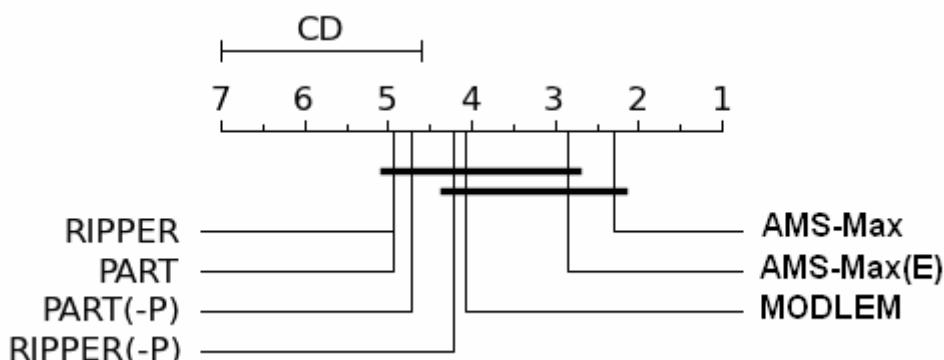
Algorytm	34 zbioru treningowego			14 zbiorów testowych		
	Acc	BAcc	Reguły	Acc	BAcc	Reguły
AMS-Max	82.9	77.6	122	83.8	78.9	96
AMS-Max(E)	82.5	76.7	53	82.6	78.0	48
MODLEM	78.7 <sup>(+)(-)</sup>	73.0	50	76.9 <sup>(-)(0)</sup>	71.3	40
PART (P)	81.6 <sup>(-)(0)</sup>	76.1	22	79.6 <sup>(-)(-)</sup>	75.2	23
PART	80.4 <sup>(+)(-)</sup>	75.6	42	80.9 <sup>(-)(0)</sup>	76.2	43
RIPPER (P)	80.6 <sup>(-)(-)</sup>	74.4	8	81.2 <sup>(-)(0)</sup>	76.2	8
RIPPER	79.7 <sup>(-)(-)</sup>	73.0	19	80.7 <sup>(-)(0)</sup>	75.0	17

W tabeli 4.19 obok wyników każdego z porównywanych algorytmów umieszczono ciąg znaków  $(*)(*)$ . W pierwszym nawiasie zamieszczono informację o tym, czy dany wynik jest statystycznie gorszy (-), lepszy (+) czy też nie ma statystycznej różnicy (0) pomiędzy danym algorytmem a adaptacyjną wersją q-ModLEM(MMM). W drugim nawiasie zamieszczono podobne informacje, dotyczące porównania algorytmu z adaptacyjną wersją q-ModLEM(EMM). Do porównań użyto testu kolejności par Wilcooxona oraz poziomu istotności 0.05. Na rysunkach 4.8 i 4.9 zamieszczono CD diagramy, obrazujące różnice pomiędzy porównywanyymi klasyfikatorami.



Rys. 4.8. Porównanie całkowitej dokładności klasyfikacji pokryciowych algorytmów indukcji reguł – 34 zbiory treningowe

Fig. 4.8. Comparison of the overall classification accuracy of sequential covering rule induction algorithms – 34 training data sets



Rys. 4.9. Porównanie całkowitej dokładności klasyfikacji pokryciowych algorytmów indukcji reguł – 14 zbiorów testowych

Fig. 4.9. Comparison of the overall classification accuracy of sequential covering rule induction algorithms – 14 testing data sets

Adaptacyjna metoda doboru miary pozwala na utworzenie klasyfikatorów o najwyższej całkowitej dokładności klasyfikacji i najwyższej średniej dokładności klas decyzyjnych. Spostrzeżenie to dotyczy zarówno 34 zbiorów treningowych, jak i 14 zbiorów testowych. W większości przypadków klasyfikatory utworzone na podstawie metody adaptacyjnej charakteryzują się statystycznie wyższą całkowitą dokładnością klasyfikacji od innych algorytmów pokryciowych, wymienionych w tabeli 4.19.

Ze względu na całkowitą dokładność klasyfikacji pomiędzy adaptacyjnymi wersjami algorytmów q-ModLEM(MMM) i q-ModLEM(EMM) nie odnotowano statystycznych różnic. Różnice występują ze względu na średnią dokładność klas decyzyjnych, niezależnie od tego, czy optymalizacja przebiega w kierunku maksymalizacji *Acc* czy *BAcc* (q-ModLEM(MMM) jest lepszy od q-ModLEM(EMM)). Niezależnie od celu optymalizacji również liczba reguł generowana przez q-ModLEM(EMM) jest około dwukrotnie mniejsza

niz liczba reguł generowana przez q-ModLEM(MMM). Różnica ta jest każdorazowo statystycznie istotna.

### 4.3. Badanie podobieństwa miar ze względu na porządek reguł

Ostatni z etapów eksperymentalnej analizy zachowania miar jakości polegał na identyfikacji grup miar tworzących podobne porządki reguł. Analizę przeprowadzono w dwóch etapach. W etapie pierwszym dla 48 zbiorów danych (połączono 34 zbiorów treningowe i 14 zbiorów testowych) wyznaczono reguły za pomocą algorytmu q-ModLEM(MMM). Proces indukcji nadzorowany był przez miarę *precision*, co oznacza, że wszędzie, gdzie było to możliwe, generowano reguły dokładne. Następnie reguły te uporządkowano według malejącej wartości miary *C2*. Reguły oceniano oddzielnie w ramach każdej klasy decyzyjnej. Zgodnie z metodyką opisaną w rozdziale 3.1.4, każdy z 48 zbiorów reguł uporządkowano za pomocą miar zawartych w tabeli 3.5. Dla każdego z 48 zbiorów otrzymano 34 porządki reguł (analizie poddano 33 miary jakości z tabeli 3.5 oraz miarę  $E^{\phi}$ ) wraz z odpowiadającymi im reprezentacjami. Aby możliwe było wyznaczenie wartości każdej z rozważanych miar, podczas tworzenia porządków reguł użyto zmodyfikowanych (patrz: rozdział 3.1.6) postaci miar *Corr*, *CN2*, *J*, *LS*, *Odds*, s.

Pomiędzy reprezentacjami porządków obliczano wartość współczynnika korelacji  $\tau$  Kendalla. Podstawą utworzenia grup miar podobnych była średnia wartość uzyskanej w tej sposób korelacji. Korelację tę nazwiemy *korelacją wzajemną* pomiędzy miarami.

W drugim etapie analizy zidentyfikowano grupy miar podobnych. Proces grupowania przebiegał następująco. Najpierw zidentyfikowano miary o wzajemnej korelacji większej od 0.99 (ze względu na możliwe błędy zaokrągleń, wynikające z implementacji, nie użyto wartości 1.00). Miary o tak dużej korelacji wzajemnej są miarami kandydującymi do zbadania, czy są one równoważne ze względu na porządek reguł. W tabelach zamieszczonych w dalszej części rozdziału miary o tak dużej korelacji wzajemnej oznaczono w ramach każdej grupy symbolem \*. Następnie do zidentyfikowanych grup miar dodawano kolejne miary, tak aby wzajemna korelacja pomiędzy miarami tworzącymi daną grupę była większa od 0.87. W ten sposób otrzymano 8 grup miar. Charakterystykę grup przedstawiono w tabeli 4.20, gdzie w kolumnie *Korelacja pomiędzy grupami* podano największą z najmniejszych oraz największą wartość współczynnika korelacji wzajemnej pomiędzy miarami tworzącymi różne grupy. W nawiasie podano numer grupy, której to porównanie dotyczy.

Tabela 4.20

## Grupy miar podobnych

Numer grupy	Miary	Minimalna korelacja wzajemna w ramach grupy	Korelacja pomiędzy grupami
1	<i>precision</i> <sup>*</sup> , <i>f</i> <sup>*</sup> , <i>RIPPER</i> <sup>*</sup> , <i>Lift</i> <sup>*</sup> , <i>DF</i> <sup>*</sup> , <i>OWS</i> <sup>*</sup> , <i>LS</i> <sup>*</sup>	1.00	0.28 (8) 0.83 (3)
2	<i>CI</i>	1.00	0.60 (8) 0.91 (3)
3	<i>wLap</i> <sup>*</sup> , <i>Laplace</i> <sup>*</sup> , <i>g</i> , <i>Odds</i>	0.92	0.67 (8) 0.90 (2)
4	<i>s</i> , <i>YAILS</i> , <i>C2</i> , <i>E<sup>ρ</sup></i>	0.92	0.61 (1) 0.90 (6)
5	<i>J</i> <sup>*</sup> , <i>CN2</i> <sup>*</sup> , <i>Klösgen</i> , <i>m</i>	0.88	0.45 (1) 0.94 (6)
6	<i>Corr</i> , <i>Gain</i> <sup>*</sup> , <i>RR</i>	0.90	0.43 (1) 0.96 (8)
7	<i>cFoli</i> <sup>*</sup> , <i>TWS</i> <sup>*</sup>	1.00	0.54 (1) 0.98 (8)
8	<i>RSS</i> <sup>*</sup> , <i>WRA</i> <sup>*</sup> , <i>Novelty</i> <sup>*</sup> , <i>RI</i> <sup>*</sup> , <i>Cohen</i> , <i>Q2</i> , <i>MS</i> , <i>accuracy</i> , <i>F</i>	0.95	0.38 (1) 0.97 (6)

Aby dokładniej sprawdzić podobieństwo miar, w ramach każdej grupy obliczono minimalną korelację pomiędzy reprezentacjami porządków. Wyniki porównań zamieszczono w tabelach 4.21–4.25, gdzie jako uzupełnienie przedstawiono także wartości korelacji dla dziesiątego percentyla i pierwszego kwartyla. W grupach o numerach 1, 2, 7 minimalna wartość współczynnika korelacji pomiędzy miarami wynosiła zawsze 1.00.

Tabela 4.21

## Minimalna korelacja pomiędzy reprezentacjami porządków reguł utworzonych na podstawie miar należących do grupy 3

	Min.	Per10	Q1		Min.	Per10	Q1
<i>g, wLap</i>	0.41	0.86	0.96	<i>wLap, Laplace</i>	0.99	1.00	1.00
<i>g, Laplace</i>	0.41	0.86	0.96	<i>wLap, Odds</i>	0.68	0.79	0.91
<i>g, Odds</i>	0.71	0.81	0.91	<i>Laplace, Odds</i>	0.68	0.79	0.91

Tabela 4.22

## Minimalna korelacja pomiędzy reprezentacjami porządków reguł utworzonych na podstawie miar należących do grupy 4

	Min.	Per10	Q1		Min.	Per10	Q1
<i>C2, s</i>	0.64	0.83	0.91	<i>C2, E<sup>ρ</sup></i>	0.69	0.79	0.85
<i>C2, YAILS</i>	0.80	0.89	0.95	<i>YAILS, E<sup>ρ</sup></i>	0.78	0.80	0.91
<i>s, YAILS</i>	0.81	0.81	0.93	<i>s, E<sup>ρ</sup></i>	0.85	0.94	0.97

Tabela 4.23

Minimalna korelacja pomiędzy reprezentacjami porządków reguł utworzonych na podstawie miar należących do grupy 5

	Min.	Per10	Q1
<i>J–CN2, m</i>	0.39	0.83	0.92
<i>J–CN2, Klösgen</i>	0.92	0.97	0.99
<i>m, Klösgen</i>	0.16	0.70	0.90

Tabela 4.24

Minimalna korelacja pomiędzy reprezentacjami porządków reguł utworzonych na podstawie miar należących do grupy 6

	Min.	Per10	Q1
<i>Corr, Gain</i>	0.95	0.97	0.98
<i>Corr, RR</i>	0.42	0.67	0.95
<i>Gain, RR</i>	0.47	0.67	0.93

Tabela 4.25

Minimalna korelacja pomiędzy reprezentacjami porządków reguł utworzonych na podstawie miar należących do grupy 8

	Min.	Per10	Q1		Min.	Per10	Q1
<i>RSS–WRA–Novelty–RI, Cohen</i>	0.92	0.97	1.00	<i>Cohen, F</i>	0.71	0.86	0.96
<i>RSS–WRA–Novelty–RI, Q2</i>	0.69	0.95	0.99	<i>Q2, MS</i>	0.60	0.84	0.97
<i>RSS–WRA–Novelty–RI, MS</i>	0.66	0.95	0.98	<i>Q2, accuracy</i>	0.81	0.91	0.96
<i>RSS–WRA–Novelty–RI, accuracy</i>	0.82	0.94	0.97	<i>Q2, F</i>	0.71	0.93	0.96
<i>RSS–WRA–Novelty–RI, F</i>	0.71	0.84	0.94	<i>MS, accuracy</i>	0.65	0.96	0.99
<i>Cohen, Q2</i>	0.83	0.92	0.98	<i>MS, F</i>	0.71	0.81	0.94
<i>Cohen, MS</i>	0.69	0.98	0.99	<i>accuracy, F</i>	0.71	0.96	0.96
<i>Cohen, accuracy</i>	0.86	0.98	0.99				

Dopełnieniem wyników jest tabela 4.26, w której dla każdej z miar przedstawiono najmniejszą z minimalnych oraz największą z minimalnych wartości współczynnika korelacji pomiędzy daną miarą a miarami należącymi do innych grup.

Tabela 4.26  
Minimalne i maksymalne korelacje pomiędzy  
reprezentacjami porządków reguł

Miara	Grupa	Minimum z minimalnej korelacji poza grupą	Maksimum z minimalnej korelacji poza grupą
<i>precision*</i>	1	-0.8311 ( <i>MS</i> ; 8)	0.1988 ( <i>CI</i> ; 2)
<i>CI</i>	2	-0.4008 ( <i>MS</i> , 8)	0.6710 ( <i>Odds</i> ; 3)
<i>g</i>	3	-0.3446 ( <i>RR</i> ; 6)	0.6299 ( <i>m</i> ; 5)
<i>wLap*</i>		-0.4116 ( <i>MS</i> , 8)	0.5023 ( <i>CI</i> ; 2)
<i>Odds</i>		-0.1226 ( <i>precision*</i> ; 1)	0.6710 ( <i>CI</i> ; 2)
<i>C2</i>	4	-0.3543 ( <i>precision*</i> ; 1)	0.6179 ( <i>Odds</i> ; 3)
<i>s</i>		-0.5317 ( <i>precision*</i> ; 1)	0.4898 ( <i>Klösgen</i> ; 5)
<i>YAILS</i>		-0.2570 ( <i>precision*</i> ; 1)	0.5836 ( <i>RR</i> ; 6)
<i>E<sup>ρ</sup></i>		-0.5426 ( <i>precision*</i> ; 1)	0.5145 ( <i>Klösgen</i> ; 5)
<i>CN2*</i>	5	-0.5853 ( <i>precision*</i> ; 1)	0.7083 ( <i>Gain</i> ; 6)
<i>Klösgen</i>		-0.6150 ( <i>precision*</i> ; 1)	0.8158 ( <i>Gain</i> ; 6)
<i>m</i>		-0.3873 ( <i>precision*</i> ; 1)	0.6299 ( <i>g</i> ; 3)
<i>Corr</i>	6	-0.7373 ( <i>precision*</i> ; 1)	0.7614 ( <i>Klösgen</i> ; 5)
<i>Gain</i>		-0.6535 ( <i>precision*</i> ; 1)	0.8158 ( <i>Klösgen</i> ; 5)
<i>RR</i>		-0.7863 ( <i>precision*</i> ; 1)	0.5836 ( <i>YAILS</i> ; 4)
<i>TWS*</i>	7	-0.7235 ( <i>precision*</i> ; 6)	0.9558 ( <i>RSS*</i> ; 2)
<i>accuracy</i>	8	-0.7943 ( <i>precision*</i> ; 1)	0.6927 ( <i>TWS*</i> ; 7)
<i>Cohen</i>		-0.7818 ( <i>precision*</i> ; 1)	0.9097 ( <i>TWS*</i> ; 7)
<i>F</i>		-0.6513 ( <i>precision*</i> ; 1)	0.7083 ( <i>TWS*</i> ; 7)
<i>MS</i>		-0.8311 ( <i>precision*</i> ; 1)	0.7076 ( <i>TWS*</i> ; 7)
<i>Q2</i>		-0.6544 ( <i>precision*</i> ; 1)	0.9356 ( <i>TWS*</i> ; 7)
<i>RSS*</i>		-0.7942 ( <i>precision*</i> ; 1)	0.9558 ( <i>TWS*</i> ; 7)

Na podstawie zamieszczonych w tabelach wyników można przedstawić następujące wnioski:

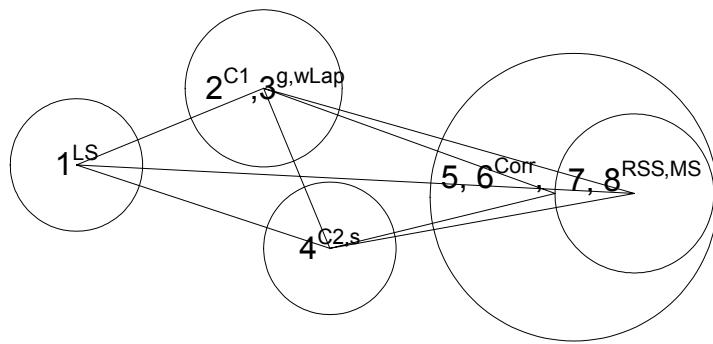
- do grupy 1 należą miary równoważne ze względu na porządek reguł (wyjątek stanowią miara *OWS* oraz zmodyfikowana miara *LS* – uzasadnienie w następnym rozdziale); do grupy 1 zaliczamy także standardową postać miary *LS* (uzasadnienie w następnym rozdziale);
- grupa miar o numerze 3 charakteryzuje się dużą korelacją wzajemną; miary *wLap* i *Laplace* są równoważne ze względu na porządek reguł (uzasadnienie w następnym rozdziale); wartości minimalnej korelacji w obrębie grupy wskazują, że najmniej reprezentatywną dla niej miarą jest *g* (*g*=2); miara *CI*, stanowiąca oddzielną grupę, jest najbardziej skorelowana z miarami tworzącymi grupę 3; wartość korelacji wzajemnej w obrębie grupy 3, rozszerzonej o miarę *CI*, to 0.85 (najmniejsza korelacja z miarą *g*);

- grupa miar o numerze 4 charakteryzuje się dużą korelacją wzajemną; wartości minimalnej korelacji w obrębie grupy wskazują, że najmniej reprezentatywną miarą dla tej grupy jest *C2*;
- miary *J* i *CN2* są miarami równoważnymi ze względu na porządek reguł (uzasadnienie w kolejnym rozdziale) porządki reguł, tworzone przez te miary, są silnie skorelowane z porządkiem tworzonymi przez miarę *Klösgen* (minimalna korelacja wynosi 0.92); miara *m* ( $m=22.433$ ) jest najmniej reprezentatywną miarą dla grupy 5;
- miary *Corr* i *Gain* tworzą podobne porządki reguł (minimalna korelacja wynosi 0.95); dla niektórych zbiorów reguł można natomiast odnotować duże różnice pomiędzy porządkiem tworzonymi przez te miary a porządkiem utworzonym na podstawie *RR*;
- miary *cFoil* i *TWS* są równoważne ze względu na porządek reguł (uzasadnienie w kolejnym rozdziale); miary te są najbardziej skorelowane z miarami tworzącymi grupę 8; wartość korelacji wzajemnej w obrębie grupy 8, rozszerzonej o miarę *TWS*, to 0.86 (najmniejsza korelacja z *accuracy*); miara *TWS*, włączona do grupy 8, jest najsilniej skorelowana z miarą *RSS* (najmniejsza korelacja wynosi 0.955);
- porządki reguł, tworzone przez miary należące do grupy 8, są silnie skorelowane; dotyczy to zarówno wzajemnej korelacji, jak i korelacji minimalnej; najmniej reprezentatywną miarą dla tej grupy jest *MS*; część z miar należących do grupy 8 jest równoważna ze względu na porządek reguł (uzasadnienie w kolejnym rozdziale).

Obniżenie wewnętrzgrupowego współczynnika korelacji wzajemnej do 0.85 powoduje połączenie grup 7 i 8 oraz 2 i 3. Obniżenie wewnętrzgrupowego współczynnika korelacji wzajemnej do 0.80 powoduje połączenie grup 5, 6, 7, 8 oraz 2 i 3. Otrzymujemy wtedy podział na cztery grupy miar podobnych ze względu na porządek reguł. Dalsze łączenie grup nie było możliwe bez znaczącego obniżenia minimalnej korelacji wewnętrzgrupowej.

Grupy 1 i 8 składają się z miar porządkujących reguły w sposób najbardziej antagonistyczny. Wynik ten jest zgodny z intuicją, gdyż grupę 1 tworzą miary przywiązuające największą wagę do dokładności reguł, a miary zawarte w grupie 8 duży (największy spośród wszystkich innych grup) nacisk kładą również na pokrycie reguły.

Na rysunku 4.10 przedstawiono diagram ilustrujący podobieństwo i niepodobieństwo porządków reguł otrzymanych przez miary należące do zidentyfikowanych grup. Na diagramie im grupy są bliżej siebie, tym większe jest podobieństwo pomiędzy reprezentacjami porządków utworzonych na podstawie zawierających je miar. Im większa jest odległość pomiędzy grupami, tym podobieństwo to jest mniejsze, a reguły oceniane są w sposób coraz bardziej antagonistyczny. Jeśli grupa złożona jest z kilku podgrup, to odległość pomiędzy numerami grup również odzwierciedla tę zasadę.



Rys. 4.10. Diagram ilustrujący podobieństwo pomiędzy grupami miar  
Fig. 4.10. Diagram reflecting similarity between the groups of measures

Nazwy miar pojawiające się przy numerach grup informują o tym, do jakiej grupy należą miary tworzące zbiór *Max*. Z diagramu można odczytać, że jedynie dwie grupy, o numerach 5 i 7, nie są reprezentowane w zbiorze *Max*. W szczególności do grupy 5 należy miara *CN2*, pojawiająca się w quasi-minimalnych zbiorach *MinAcc* i *MinBAcc*.

Miara złożona  $nMM(C1, Odds, C2, g)$  jest najbardziej skorelowana z grupą 3, a dołączając tę miarę do grupy, uzyskamy korelację wzajemną, równą 0.85.

Na podstawie badań podobieństwa miar możemy wyciągnąć wniosek, że w zbiorze rozważanych przez nas miar nie ma miar antagonistycznych, gdyż nie zidentyfikowano takiej pary miar, dla której wartość współczynnika korelacji pomiędzy reprezentacjami porządków reguł wynosiłaby zawsze -1.

## 4.4. Analiza równoważności i własności miar

### 4.4.1. Równoważność

Na podstawie stwierdzenia 3.11 możemy wnioskować, że ze względu na sposób rozstrzygania konfliktów klasyfikacji przez głosowanie równoważne są następujące miary:

- *CN2, J-measure*; mnożąc wartość *J-measure* przez liczbę dodatnią  $2(P + N)$ , uzyskamy wartość *CN2*;

- *Novelty, RI, WRA*; po prostych przekształceniach wzorów definiujących te miary otrzymamy:

$$\text{Novelty} \equiv \frac{p(P + N) - P(p + n)}{(P + N)^2}, \quad RI \equiv \frac{p(P + N) - P(p + n)}{P + N},$$

$$WRA \equiv \frac{1}{P + N} \cdot \frac{p(P + N) - P(p + n)}{P + N}; \text{ wyraźnie widać, że } \text{Novelty} \text{ to } WRA \text{ zapisana}$$

w równoważny, nieco inny sposób; mnożąc *RI* przez liczbę dodatnią  $1/(P + N)$ , uzyskamy miary *Novelty* i *WRA*;

- $cFoil$ ,  $TWS$ ; do wykazania równoważności miar  $cFoil$  i  $TWS$  wykorzystamy własność  $\log_X y = \frac{\ln y}{\ln X}$ ; obie miary definiowane są przez następujące wyrażenia:

$$TWS \equiv \frac{p}{P+N} \ln \left( \frac{p(P+N)}{(p+n)P} \right), \quad cFoil \equiv p \log_2 \left( \frac{p(P+N)}{(p+n)P} \right); \text{ mnożąc } TWS \text{ przez liczbę dodatnią } \frac{(P+N)}{\ln 2}, \text{ otrzymamy } cFoil.$$

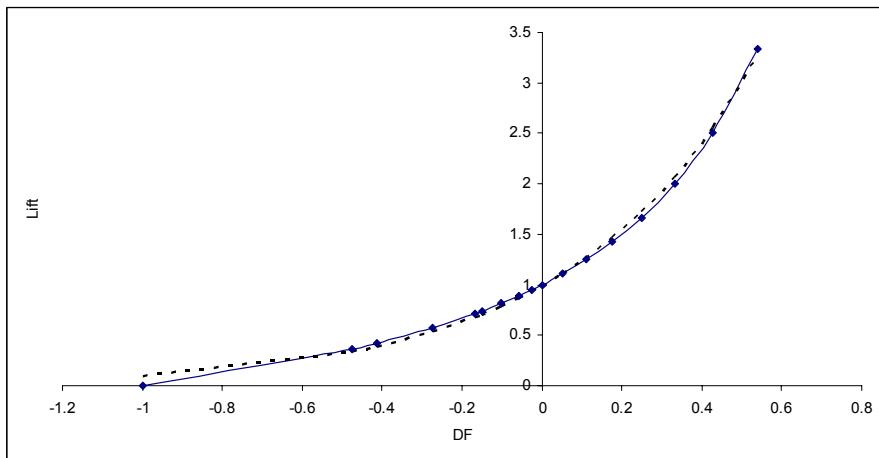
Zgodnie z definicjami 3.12 i 3.13 oraz uwagą 3.6 przedstawione pary miar są równoważne ze względu na sposób rozstrzygania konfliktów klasyfikacji w schemacie największego zaufania oraz ze względu na porządek reguł.

Wśród miar zawartych w tabeli 3.5 można zidentyfikować parę miar równoważnych ze względu na sposób rozstrzygania konfliktów w schemacie największego zaufania. Parę tę tworzą miary *precision* i *RIPPER*. Równoważnymi ze względu na sposób rozstrzygania konfliktów klasyfikacji w schemacie największego zaufania są prawdopodobnie także miary *Lift* i *DF*. Ponieważ hipotezy tej nie udało się udowodnić w sposób analityczny, poniżej przedstawiono jedynie jej empiryczne uzasadnienie.

Dowodząc równoważność miar *precision* i *RIPPER*, wystarczy zauważyć, że  $RIPPER=2precision-1$ . Istnieje zatem funkcja rosnąca, przekształcająca wartości miary *precision* na wartości miary *RIPPER*, co gwarantuje (zgodnie ze stwierdzeniem 3.10), że miary te są równoważne ze względu na sposób rozstrzygania konfliktów klasyfikacji w schemacie największego zaufania i równoważne ze względu na porządek reguł. Zależność  $RIPPER=2precision-1$  nie gwarantuje natomiast, że miary są równoważne ze względu na sposób rozstrzygania konfliktów przez głosowanie. Liczba -1, występująca we wzorze przekształcającym *precision* na *RIPPER*, może powodować, że w trakcie głosowania klasyfikatory stosujące te miary będą podejmować odmienne decyzje. Potwierdzają to wyniki badań eksperymentalnych, gdzie w 3 przypadkach (na 48) występowały różnice pomiędzy klasyfikatorami stosującymi podczas głosowania miary *precision* oraz *RIPPER*.

Dowodząc równoważności miar *Lift* i *DF*, podjęto próby znalezienia funkcji rosnącej, przekształcającej wartości *Lift* na *DF*. Problem ten pozostaje otwarty, gdyż odwzorowania takiego nie udało się znaleźć. Hipotezę o równoważności miar postawiono na podstawie analizy zależności pomiędzy ich wartościami. Analizę taką wykonano dla różnych zestawów wartości  $P, N, p, n$ . Każdorazowo zależność pomiędzy *Lift* i *DF* identyfikowano jako  $Lift = c + e^{(b_0 + b_1 \cdot DF)}$ , przy czym wartości stałych były różne w zależności od rozmiaru zbioru danych i rozkładu pomiędzy liczbą przykładów pozytywnych i negatywnych (rys. 4.11). Autorowi nie udało się powiązać wartości estymowanych parametrów  $c, b_0, b_1$  z wartościami  $P, N, p, n$ , charakteryzującymi zbiór danych i ocenianą regułę. Niemniej jednak analiza regresji każdorazowo wykazywała znakomite dopasowanie modelu do danych empirycznych.

W przeprowadzonych eksperymentach współczynnik determinacji nigdy nie był niższy niż 0.996. Na tej podstawie postawiono hipotezę, że miary *Lift* i *DF* są równoważne ze względu na sposób rozstrzygania konfliktów w schemacie największego zaufania.



Rys. 4.11. Zależność pomiędzy wartościami miar *DF* i *Lift*.

Linią ciągłą oznaczono wykres otrzymany na podstawie danych empirycznych, linią przerywaną – wykres otrzymany na podstawie analizy regresji

Fig. 4.11. Dependence between values of the *DF* and *Lift* measures.

The solid line denotes the graph obtained on the basis of empirical data, the broken line – the graph obtained based on the regression analysis

Ostatni z rozważanych przez nas typów równoważności to równoważność ze względu na porządek reguł. Tym rodzajem równoważności charakteryzują się następujące pary miar:

- *precision* i *f*; założymy, że regułę  $r_1$  pokrywa  $p_1$  przykładów pozytywnych oraz  $n_1$  przykładów negatywnych, analogiczne założenie przyjmujemy dla reguły  $r_2$ ; wówczas  $precision(r_1) = precision(r_2)$  wtedy i tylko wtedy, gdy  $p_1n_2 - p_2n_1 = 0$ , ponadto  $precision(r_1) > precision(r_2)$  wtedy i tylko wtedy, gdy  $p_1n_2 - p_2n_1 > 0$ ; równocześnie  $f(r_1) = f(r_2)$  wtedy i tylko wtedy, gdy  $2(p_1n_2 - p_2n_1) = 0$  oraz  $f(r_1) > f(r_2)$  wtedy, gdy  $2(p_1n_2 - p_2n_1) > 0$ ; wykorzystując przytoczone zależności, łatwo można pokazać, że  $precision(r_1) = precision(r_2)$  wtedy i tylko wtedy, gdy  $f(r_1) = f(r_2)$ , oraz  $precision(r_1) > precision(r_2)$  wtedy i tylko wtedy, gdy  $f(r_1) > f(r_2)$ , co dowodzi kompatybilności miar, a więc ich równoważności ze względu na porządek reguł;

- *precision* i *Lift*; zauważmy, że  $Lift = precision \cdot \frac{(P+N)}{P}$ ; porządkując reguły z ustalonej

klasy decyzyjnej, wartość *Lift* otrzymamy, mnożąc dokładność (*precision*) reguły przez stałą  $(P+N)/P$ ; spostrzeżenie to wystarcza do udowodnienia równoważności miar *precision* i *Lift* ze względu na porządek reguł; wyniki przeprowadzonych eksperymentów pokazują, że miary *precision* i *Lift* nie charakteryzują się żadnym z pozostałych typów równoważności; brak równoważności ze względu na sposób rozstrzygania konfliktów

klasyfikacji można również pokazać analitycznie; zakładając, że mamy do czynienia z dwuklasowym problemem klasyfikacji, wartości miar w ramach każdej klasy decyzyjnej przeliczane są według następujących wzorów:

$Lift_{X_1} = precision_{X_1} \cdot ((P+N)/P)$ ,  $Lift_{X_2} = precision_{X_2} \cdot ((P+N)/N)$ ; wzory te są identyczne jedynie wtedy, gdy liczebność obu klas jest identyczna;

- $precision$  i  $DF$ ; założmy, że regułę  $r_1$  pokrywa  $p_1$  przykładów pozytywnych oraz  $n_1$  przykładów negatywnych, analogiczne założenie przyjmujemy dla reguły  $r_2$ ; wówczas  $precision(r_1) = precision(r_2)$  wtedy i tylko wtedy, gdy  $p_1n_2 - p_2n_1 = 0$ , ponadto  $precision(r_1) > precision(r_2)$  wtedy i tylko wtedy, gdy  $p_1n_2 - p_2n_1 > 0$ ; równocześnie  $DF(r_1) = DF(r_2)$  wtedy i tylko wtedy, gdy  $2(P+N)P(p_1(p_2 + n_2) - p_2(p_1 + n_1)) = 0$ , skąd po prostych przekształceniach uzyskamy zależność  $DF(r_1) = DF(r_2)$  wtedy i tylko wtedy, gdy  $p_1n_2 - p_2n_1 = 0$ ; w identyczny sposób można pokazać, że  $DF(r_1) = DF(r_2)$  wtedy i tylko wtedy, gdy  $p_1n_2 - p_2n_1 > 0$ ; w ten sposób uzyskujemy równoważność miar  $precision$  i  $DF$  ze względu na porządek reguł;
- $precision$  i  $LS$ ; założmy, że regułę  $r_1$  pokrywa  $p_1$  przykładów pozytywnych oraz  $n_1$  przykładów negatywnych, analogiczne założenie przyjmujemy dla reguły  $r_2$ ; wówczas  $precision(r_1) = precision(r_2)$  wtedy i tylko wtedy, gdy  $p_1n_2 - p_2n_1 = 0$ , ponadto  $precision(r_1) > precision(r_2)$  wtedy i tylko wtedy, gdy  $p_1n_2 - p_2n_1 > 0$ ; dla podstawowej, niemodyfikowanej postaci miary  $LS$  zachodzą identyczne warunki, tzn.  $LS(r_1) = LS(r_2)$  wtedy i tylko wtedy, gdy  $p_1n_2 - p_2n_1 = 0$ , oraz  $LS(r_1) > LS(r_2)$  wtedy i tylko wtedy, gdy  $p_1n_2 - p_2n_1 > 0$ ; dowodzi to równoważności miar  $precision$  i  $LS$  ze względu na porządek reguł; miara  $LS$  nie nadaje się do rozstrzygania konfliktów klasyfikacji, gdyż jej wartość jest nieokreślona dla reguł dokładnych;
- $RSS$  i  $WRA$ ; dowodząc równoważności miar  $RSS$  i  $WRA$ , wystarczy zauważyć, że  $RSS = WRA \cdot \frac{(P+N)^2}{PN}$ ; dla reguł wskazujących na identyczną klasę decyzyjną wartość  $(P+N)^2 / PN$  jest liczbą dodatnią; zauważmy, że dla dwuklasowych problemów klasyfikacji miary  $RSS$  oraz  $WRA$  również są równoważne ze względu na sposób rozstrzygania konfliktów klasyfikacji w schematach głosowania i największego zaufania; jest tak, ponieważ liczba  $(P+N)^2 / PN$  dla każdej z dwóch klas decyzyjnych będzie taka sama; w problemach wieloklasowych miary te nie są jednak równoważne ze względu na sposób rozstrzygania konfliktów klasyfikacji; zakładając, że dane są trzy klasy decyzyjne o liczebności 10, 20 i 70 przykładów, wartość  $PN$  będzie dla tych klas wynosić odpowiednio  $10 \cdot 90$ ,  $20 \cdot 80$  i  $70 \cdot 30$ ;

- *Laplace* i *wLap*; dowodząc równoważności miar *Laplace* i *wLap*, wystarczy zauważyć, że  $wLap = Laplace \cdot \frac{(P+N)}{P}$ ; dla reguł wskazujących na identyczną klasę decyzyjną wartość  $(P+N)/P$  jest liczbą dodatnią; miary *Laplace* i *wLap* nie są równoważne ze względu na sposób rozstrzygania konfliktów klasyfikacji, uzasadnienie jest podobne jak dla miar *RSS* i *WRA*.

Ponieważ równoważność miar ze względu na porządek reguł jest relacją równoważności (stwierdzenie 3.8), miary *f*, *Lift*, *LS*, *RIPPER*, *DF* są również równoważne ze względu na porządek reguł.

W tabeli 4.27 zamieszczono przykłady pokazujące, że pary miar równoważnych ze względu na porządek reguł nie są równoważne ze względu na sposób rozstrzygania konfliktów klasyfikacji. Zauważmy, że wykazanie braku takiej równoważności sprowadza się do pokazania, że dla pewnych wartości  $p$  i  $n$ , charakteryzujących reguły opisujące identyczne klasy decyzyjne, miary te nie oceniają reguł w zgodny sposób (tzn. oceniają je w sposób antagonistyczny lub jedna miara ocenia reguły identycznie, a druga przyporządkowuje im różne wartości ocen). Ponieważ schemat największego zaufania jest szczególnym przypadkiem głosowania, przykłady zamieszczone w tabeli 4.27 pokazują, iż występujące tam pary miar nie są równoważne ani w schemacie największego zaufania, ani w głosowaniu. Wnioski wynikające z analizy tabeli 4.27 znajdują także potwierdzenie w wynikach eksperymentalnych, prezentowanych w tabelach 4.2, 4.4, 4.16, 4.17.

Tabela 4.27

Przykłady ilustrujące brak równoważności miar ze względu na sposób rozstrzygania konfliktów klasyfikacji  
( $P=60, N=40$ )

$r_1$		$r_2$		$r_1 > r_2$ lub $r_1 = r_2$	$r_2 > r_1$
$p$	$n$	$p$	$n$		
1	2	1	3	<i>precision</i> ( $>$ )	<i>f</i>
1	2	1	3	<i>precision</i> ( $>$ )	<i>DF</i>
1	2	1	3	<i>precision</i> ( $>$ )	<i>Lift</i>
30	30	20	20	<i>RIPPER</i> (=)	<i>f</i>
10	0	1	0	<i>RIPPER</i> (=)	<i>DF</i>
10	0	1	0	<i>RIPPER</i> (=)	<i>Lift</i>
10	0	1	0	<i>f</i> (=)	<i>DF</i>
10	0	1	0	<i>f</i> (=)	<i>Lift</i>

W tabeli 4.27 każda z reguł charakteryzowana jest przez liczbę pokrywających ją przykładów pozytywnych  $p$  i negatywnych  $n$ . W kolumnie  $r_1 > r_2$  lub  $r_1 = r_2$  wymieniono nazwę miary, dla której reguła  $r_1$  charakteryzuje się identyczną (=) lub wyższą (>) wartością miary niż reguła  $r_2$ . W kolumnie  $r_2 > r_1$  wymieniono nazwę miary, dla której reguła  $r_2$  charakteryzuje się wyższą wartością miary niż reguła  $r_1$ .

W grupie miar podobnych do *precision* znalazła się miara *OWS*. Dla zbiorów danych, na podstawie których przeprowadzono analizę podobieństwa miar, korelacja pomiędzy reprezentacjami porządków reguł, utworzonymi na podstawie miar *precision* i *OWS*, zawsze wynosiła 1. Wystarczy jednak spojrzeć na wykresy miar *precision* i *OWS* (patrz: dodatek A), aby przekonać się, że miary te nie są równoważne. W szczególności *OWS* nie jest monotoniczna ze względu na  $p$  (brak własności  $M_p$ ). Miara *OWS* nie jest równoważna z żadną z miar równoważnych *precision*.

Wcześniej wykazano, że miara *LS* jest równoważna z *precision* ze względu na porządek reguł. Przedstawiona w rozdziale 3.1.6 modyfikacja *LS* miała na celu umożliwianie oceny reguł dokładnych i stosowanie *LS* do nadzorowania procesu indukcji reguł. Modyfikacja ta powoduje, że nowa postać  $LS = pN / (nP + 1)$  nie jest równoważna z *precision*. Założymy, że reguły  $r_1$  i  $r_2$  pokrywa odpowiednio 5 i 10 przykładów pozytywnych oraz nie pokrywa ich żaden przykład negatywny. Zgodnie z przyjętymi założeniami otrzymamy  $precision(r_1) = precision(r_2)$  oraz  $LS(r_1) < LS(r_2)$ . Miara *precision* i zmodyfikowana miara *LS* nie są kompatybilne, zatem nie są równoważne ze względu na porządek reguł.

#### 4.4.2. Własności

W tabeli 4.28 przedstawiono własności rozważanych w niniejszej monografii miar jakości, definiowanych na podstawie tablicy kontyngencji. Inni autorzy analizowali już własności kilku z miar wymienionych w tabeli 4.28. W szczególności w pracach Bruhy [40, 42] można znaleźć analizę monotoniczności miar *C2*, *C1*, *J* i *Cohen* ze względu na zmieniającą się dokładność ocenianych reguł. Analizą monotoniczności definiowanej poprzez własności  $M1$ – $M4$  miar *f*, *s*, *DF*, *RI* zajmowali się Szczęch, Greco i Ślomiński, analizując również wymienione miary pod kątem posiadania przez nie własności konfirmacji i symetrii względem hipotezy [43, 300]. Wykazali oni też, że dla ustalonych  $P$  oraz  $N$ , przy niezmieniającym się wsparciu ( $p$ ), występuje monotoniczna zależność pomiędzy dokładnością reguł a wartością dowolnej miary definiowanej na podstawie tablicy kontyngencji i charakteryzującej się własnościami  $M1$ – $M4$ . Ponadto podali oni warunki, jakie musi spełniać miara charakteryzująca się  $M1$ – $M4$ , aby dla ustalonych  $P$  i  $N$  oraz przy niezmieniającej się dokładności być miarą monotoniczną względem wsparcia.

Nawiązując do przywołanych rezultatów, można łatwo zauważyc, że jeśli miara ma własność  $M_n$ , to przy ustalonym wsparciu (stałym  $p$ ) miara ta jest monotoniczna ze względu na zmieniającą się dokładność reguły. W dowodzie wystarczy zauważyc, że dokładność definiowana jest jako  $p / (p + n)$ . Ponieważ  $p$  jest stałe, dokładność maleje, jeśli  $n$  rośnie. Jeśli miara charakteryzuje się monotonicznością  $M_n$ , wraz ze wzrostem  $n$  maleje wartość miary.

Monotoniczny związek pomiędzy wartościami miary a wsparciem reguły przy niezmieniającej się dokładności musi być badany dla każdej miary oddzielnie. W badaniach tych wykorzystuje się zależności:  $p= support(r)$  oraz  $n=(support(r)/precision(r)) - support(r)$ . Na ich podstawie każdą miarę definiowaną na podstawie tablicy kontyngencji można zapisać jako funkcję zmiennych *support* i *precision*. Monotoniczność miary ze względu na każdą z tych zmiennych określa się na podstawie analizy wartości pochodnych cząstkowych.

Pierwsza praca autora związana z badaniem monotoniczności miar ze względu na zmieniające się wartości  $p$  i  $n$  ukazała się w 2001 roku [244]. Badano w niej m.in. monotoniczność takich miar, jak: *Cohen*, *Coleman*, *IKIB*, *WS*,  $\chi^2$  oraz *Gain*. W tabeli 4.28 przedstawiono wyniki analizy własności 32 miar jakości. Rozważano własności zidentyfikowane w rozdziale 3 jako pożądane dla miar oceniających jakość reguł decyzyjnych. Dla zestawu własności zawartego w tabeli 4.28 i większości z prezentowanych w niej miar jest to pierwsza tego typu analiza.

Dowody pokazujące, które z rozważanych miar mają określone własności, nie są trywialne, są jednak na tyle proste, że autor zdecydował się na ich pominięcie, gdyż znacząco zwiększyłoby to objętość monografii. Dowody każdorazowo opierają się na identycznych założeniach:

- monotoniczność udowadnia się wprost z definicji lub na podstawie pochodnych cząstkowych, obliczanych względem zmiennych  $p$  i  $n$ ; przydatne są także wykresy miar, na których w prosty sposób można zidentyfikować przykłady ilustrujące, że dana miara nie charakteryzuje się danym typem monotoniczności; badając monotoniczność, można korzystać ze stwierdzenia 3.9, mówiącego o monotoniczności miar równoważnych ze względu na porządek reguł;
- udowadniając  $D_{in}$ , należy na podstawie równości  $\frac{p_1}{p_1 + n_1} = \frac{p_2}{p_2 + n_2} = \frac{P}{P + N}$  dowieść, że  $q(p_1, n_1) = q(p_2, n_2)$ ; w dowodach wykorzystuje się równoważność:  $(p/(p+n)) = (P/(P+N)) \Leftrightarrow pN = nP$ , na podstawie której wylicza się  $p$  lub  $n$ , co ułatwia wykazanie równości  $q(p_1, n_1) = q(p_2, n_2)$ ; w podobny sposób dowodzi się własności  $D_>$  oraz  $D_<$ ; w dowodach  $D_>$  i  $D_<$  można również wykorzystać tezy zawarte w uwagach 3.4 i 3.5;
- w dowodach  $D_{eq}$  wykorzystuje się założenie, że  $p = n$ ; podstawiając  $p$  w miejsce  $n$ , wykazujemy, że wartość miary jest niezależna od  $p$  i  $n$ ; w dowodzie własności  $D_{eq}$  można wykorzystać także tezy uwagi 3.5;
- w dowodach  $D_{ES}$  i  $D_{HS}$  korzysta się wprost z definicji tych własności, przykładowo, jeśli dla reguły  $r$  wartość miary  $q$  obliczamy na podstawie argumentów  $p, n, P-p, N-n$ , to dla reguły  $r-$  wartość  $q$  obliczamy na podstawie argumentów  $n, p, N-n, P-p$ ; dla

tak dobranych argumentów pokazujemy, że  $q(r) = q(p, n, P - p, N - n) = -q(r \neg) = -q(n, p, N - n, P - p)$ ;

- miary mające własności  $M_p$  i  $M_n$  oraz takie, że dla  $p = P$  i  $n = 0$  oraz  $p = 0$  i  $n = N$  ich wartości są stałe niezależnie od  $P$  i  $N$ , będą charakteryzować się własnością  $D_{mwEx1}$ .

W tabeli 4.28 litera Y oznacza, że miara charakteryzuje się daną własnością, a litera N oznacza, że miara nie ma danej własności.

Spora grupa miar niemonotonicznych staje się monotoniczna, jeśli założymy, że  $p \in \{1, 2, \dots, P\}$  (czyli  $p > 0$ ) oraz  $n \in \{1, 2, \dots, N\}$  (czyli  $n > 0$ ). Dla miar stosowanych w pokryciowych algorytmach indukcji reguł decyzyjnych szczególnie uzasadnione jest ograniczenie  $p > 0$ , gdyż algorytmy te zakładają, że reguła pokrywa co najmniej jeden przykład pozytywny. Jeśli dla przyjętych ograniczeń miara niemonotoniczna zmieniała się w miarę monotoniczną, to informację tę umieszczano w nawiasie.

Analizując miary ze względu na własność  $D_{mwEx1}$ , zauważono, że pewna grupa miar przyjmuje stałą wartość jedynie dla reguł pokrywających wszystkie przykłady pozytywne i żadnego przykładu negatywnego (w tabeli 4.28 warunek ten określono jako *max*).

Tabela 4.28  
Analiza własności miar definiowanych na podstawie tablicy kontyngencji

Grupa	Miara	Monotoniczność		Opis							
		$M_p (n \geq 1)$	$M_n (p \geq 1)$	$D_{in}$	$D_>$	$D_<$	$D_{eq}$	$D_{ES}$	$D_{HS}$	$D_{WL}$	$D_{mwEx1} (max)$
	<i>coverage</i>	Y	N	N	N	N	N	N	N	N	N
	<i>support</i>	Y	N	N	N	N	N	N	N	N	N
1	<i>precision</i>	N (Y)	N (Y)	Y	Y	Y	Y	N	N	N	N
	<i>f</i>	N (Y)	N (Y)	Y	Y	Y	Y	N	Y	N	N
	<i>RIPPER</i>	N (Y)	N (Y)	Y	Y	Y	Y	N	Y	N	N
	<i>DF</i>	N (Y)	N (Y)	Y	Y	Y	Y	N	N	N	N
	<i>Lift</i>	N (Y)	N (Y)	Y	Y	Y	Y	N	N	N	N
	<i>LS</i>	N (Y)	N (Y)	Y	Y	Y	Y	N	N	N	N
	<i>OWS</i>	N	N	Y	Y	Y	Y	N	N	N	N
2	<i>Cl</i>	Y	N	Y	Y	Y	N	N	N	N	N (Y)
3	<i>wLap</i>	Y	Y	N	N	N	Y	N	N	Y	N
	<i>Laplace</i>	Y	Y	N	N	N	Y	N	N	Y	N
	<i>g</i>	Y	N (Y)	N	N	N	N	N	N	N	N
	<i>Odds</i>	N (Y)	N (Y)	Y	Y	Y	N	N	N	N	N
4	<i>C2</i>	Y	N (Y)	Y	Y	Y	N	N	N	N	N (Y)
	<i>s</i>	Y	Y	Y	Y	Y	N	Y	Y	Y	Y
	<i>YAILS</i>	Y	N (Y)	N	N	N	N	N	N	N	N (Y)
	<i>E<sup>ρ</sup></i>	Y	N (Y)	Y	Y	Y	N	N	N	N	N (Y)
5	<i>CN2-J</i>	N	N	Y	Y	N	N	N	N	N	N
	<i>Klösgen</i>	Y	Y	Y	Y	Y	N	N	Y	Y	N
	<i>m</i>	Y	Y	Y	Y	Y	N	N	N	Y	N
6	<i>Corr</i>	Y	Y	Y	Y	Y	N	Y	Y	Y	Y
	<i>Gain</i>	N	N	Y	Y	N	N	N	N	N	N
	<i>RR</i>	Y	N (Y)	N	N	N	N	N	N	N	N
7	<i>cFoil-TWS</i>	N	N (Y)	Y	Y	Y	N	N	N	N	N
8	<i>RSS</i>	Y	Y	Y	Y	Y	N	Y	Y	Y	Y
	<i>WRA-Novelty-RI</i>	Y	Y	Y	Y	Y	N	Y	Y	Y	N
	<i>Cohen</i>	Y	Y	Y	Y	Y	N	N	N	Y	N (Y)
	<i>Q2</i>	N	N	Y	Y	Y	N	N	N	N	N (Y)
	<i>MS</i>	Y	N (Y)	N	N	N	N	N	N	N	N (Y)
	<i>accuracy</i>	Y	Y	N	N	N	Y	N	Y	Y	N
	<i>F</i>	Y	Y	N	N	N	N	N	N	Y	N (Y)

Analizę zamieszczoną w tablicy 4.28 przeprowadzono dla podstawowych, niemodyfikowanych postaci miar. W przypadku miar zawierających parametr, jego wartość była taka, jak ustalono to w podrozdziale 3.1.6.

Przeprowadzona analiza wykazała, że większość miar ma pożądane, z punktu widzenia nadzorowania procesu indukcji reguły własności  $M_p$  i  $M_n$ . Wszystkie z rozważanych miar charakteryzują się monotonicznością  $M_n$ , jeśli oceniana przez nie reguła  $r$  spełnia warunek  $precision(r) \geq P/(P+N)$ . Ponadto w przypadku miar *Gain* oraz *CN2* wystarczy, aby dla reguł niespełniających warunku  $precision(r) \geq P/(P+N)$  zmienić znak wartości miary na przeciwny, a uzyskana w ten sposób modyfikacja będzie charakteryzowała się własnościami  $M_p$ ,  $M_n$ ,  $D_{in}$ ,  $D_>$ ,  $D_<$ . Dodatkowo miara *Gain* będzie także symetryczna ( $D_{HS}$ ). W eksperymentach polegających na indukcji reguł za pomocą miar *Gain* i *CN2* wykorzystano właśnie tak zmodyfikowane postaci tych miar.

Największą liczbę własności, jakimi powinny charakteryzować się miary przeznaczone do oceny jakości opisowej reguł decyzyjnych, mają miary *Corr*, *RSS* oraz *s*, a następnie *f* i *RIPPER*. Miary *Corr*, *RSS* i *s* nie charakteryzują się jedynie własnością  $D_{eq}$ , a miary *f* i *RIPPER* – własnościami  $D_{ES}$ ,  $D_{wL}$  oraz  $D_{mwEx1}$ .

Duża grupa miar charakteryzuje się własnościami  $D_{in}$ ,  $D_>$ ,  $D_<$ , jedynie część z nich to miary konfirmacji (m.in. *C1*, *C2*, *Corr*, *f*, *RSS*, *s*).

Proponowana w rozdziale 3.2.2 miara złożona  $nMM(C1, Odds, C2, g)$  charakteryzuje się własnościami  $M_p$ ,  $M_n$  (dla  $p \geq 1$ ) oraz  $D_{mwEx1}$  (dla reguł pokrywających wszystkie przykłady pozytywne i żadnego przykładu negatywnego).

#### 4.5. Omówienie zbioru miar najbardziej efektywnych

W przeprowadzonych badaniach eksperymentalnych najbardziej efektywny okazał się zbiór miar  $Max=\{C1, C2, Corr, g, LS, MS, RSS, s, wLap\}$ . Miary należące do tego zbioru zapewniały tworzenie zwycięskich klasyfikatorów na każdym z 48 rozważanych zbiorów danych. Najwyższą całkowitą dokładność klasyfikacji uzyskują miary prowadzące do indukcji największej liczby reguł. Ze względu na średnią dokładność klas decyzyjnych bezkonkurencyjna okazała się miara *wLap*. W zbiorze *Max* znalazły się również miary pozwalające na utworzenie zwycięskich klasyfikatorów ze względu na kryterium *AvC* (najlepsza miara to *RSS*). Wyniki klasyfikatorów optymalizowanych w kierunku maksymalizacji *AvC* pokazują, że kryterium to w ocenie zbioru reguł najwyższą wagę przywiązuje do liczby generowanych reguł. Jest tak, ponieważ istnieje silna zależność pomiędzy iloczynem średniej dokładności i średniego pokrycia reguł a liczbą reguł tworzących klasyfikator. Poprawę zdolności klasyfikacyjnych zbiorów reguł optymalizowanych w kierunku maksymalizacji *AvC* można uzyskać, wprowadzając do tego kryterium wagę. Zwiększenie roli całkowitej dokładności klasyfikacji lub średniej dokładności klas decyzyjnych w kryterium *AvC* spowodowałoby jednak z pewnością indukcję większej liczby reguł.

Ze względu na całkowitą dokładność klasyfikacji i średnią dokładność klas decyzyjnych zidentyfikowane quasi-minimalne zbiory zwycięskich miar są podzbiorami zbioru  $Max \cup \{CN2\}$ .

Interesujące jest to, że zbiór miar  $Max \cup \{CN2\}$  ma co najmniej jednego reprezentanta w każdej ze zidentyfikowanych grup miar podobnych ze względu na porządek reguł. Wyjątkiem jest grupa 5, zawierająca miarę  $TWS$  i równoważną do niej, ze względu na sposób rozstrzygania konfliktów klasyfikacji, miarę  $cFoil$ . Oznacza to, że zbiór  $Max$  rozszerzony o miarę  $CN2$  zawiera miary oceniające reguły z różnych punktów widzenia. Informacja ta może być użyteczna podczas wyboru kryteriów charakteryzujących jakość reguł w wielokryterialnej metodzie selekcji reguł najbardziej interesujących (patrz: rozdział 3.2.5).

Metoda adaptacyjnego doboru miar jakości, stosująca miary zawarte w zbiorze  $Max$ , pozwala każdorazowo uzyskać najlepsze wyniki zarówno ze względu na całkowitą dokładność klasyfikacji, jak i na średnią dokładność klas decyzyjnych. Po zastosowaniu tej metody zaobserwowano wyższe wartości  $Acc$  i  $BAcc$  na 34 zbiorach treningowych i 14 testowych. Dla całkowitej dokładności klasyfikacji otrzymano również dobre p-wartości (0.107 dla 34 zbiorów treningowych i 0.065 dla 14 zbiorów testowych), informujące o różnicy pomiędzy metodą adaptacyjną a najlepszą z miar, jaką jest  $C2$ . Ze względu na konieczność adaptacyjnego dobru miary czas indukcji reguł jest oczywiście wielokrotnie dłuższy. Dłuższy czas obliczeń nie jest jednak znaczącym ograniczeniem metody adaptacyjnej, gdyż doskonale nadaje się ona do zrównoleglenia. Obliczenia dla każdej z miar oraz każdego przebiegu wewnętrznej walidacji krzyżowej mogą być wykonywane niezależnie.

Ze względu na kryterium optymalności, jakim jest całkowita dokładność klasyfikacji, dobre wyniki uzyskała także miara złożona  $nMM(C1, Odds, C2, g)$ .

Wyniki będące podstawą do wyciągnięcia wymienionych wniosków otrzymano na podstawie pokryciowego algorytmu indukcji q-ModLEM, uruchamianego w dwóch trybach użycia miary (MMM i EMM). Duża liczba eksperymentów przeprowadzonych na zbiorach danych o różnorodnej charakterystyce jest, zdaniem autora, podstawą pozwalającą uogólnić wnioski na temat efektywności użycia miar na inne pokryciowe algorytmy indukcji reguł i inne zbiory danych.

Dla otrzymanych wyników nie jest również bez znaczenia przyjęty mechanizm rozstrzygania konfliktów klasyfikacji. W przeprowadzonych eksperymentach zastosowano głosowanie (dokładniej – ważone głosowanie), a siłę głosu każdej reguły odzwierciedlono za pomocą tej samej miary, która stosowana była do oceny jakości budowanej reguły. Taki schemat postępowania stanowi najprostsze i najczytelniejsze rozwiązanie. Wyniki publikacji analizujących różne schematy klasyfikacji [42, 120, 305, 306] sugerują, że przyjęte przez autora rozwiązanie jest zgodne z intuicją i przynosi dobre (często najlepsze) rezultaty.

Analiza własności miar pokazuje, że pewna grupa miar nadaje się do oceny zdolności opisowych reguł. W rozważanym zbiorze miar nie ma ani jednej miary charakteryzującej się wszystkimi pożądanymi dla tego celu własnościami. W zbiorze  $\text{Max} \cup \{\text{CN2}\}$  zawarte są miary *Corr*, *RSS* i *s*, które charakteryzują się wszystkimi własnościami poza  $D_{\text{eq}}$ . Właściwością  $D_{\text{eq}}$  charakteryzuje się należąca do zbioru *Max* miara *wLap*. Ponadto właściwość  $D_{\text{eq}}$  ma także miara *precision*, której wartość podaje się najczęściej jako jeden z kilku wskaźników charakteryzujących wyznaczone reguły.

Dalej omówiono każdą z miar należących do zbioru  $\text{Max} \cup \{\text{CN2}\}$ .

Miary *C1* i *C2* zostały zaproponowane przez Bruhę [40]. Są to miary o genezie empirycznej, łączące dwie miary o genezie statystycznej. Bruha zauważył, że miara *Coleman* [25] zbyt dużą uwagę przywiązuje do dokładności reguł, a miara *Cohen* (która znana jest przede wszystkim jako współczynnik korelacji  $\kappa$ ) [56] zbyt dużą wagę przywiązuje do pokrycia reguł. Miara *C1* łączy miary *Coleman* i *Cohen*. Miara *C2* łączy miary *Coleman* i *coverage*. Miara *Coleman* to pierwszy składnik iloczynu definiującego miary *C1* i *C2*. Mierzy ona siłę zależności pomiędzy zdarzeniami: *przykład x pokrywa regułę a przykład x należy do klasy decyzyjnej opisywanej przez regułę*. Mianownik pełni rolę normalizującą. Zakres zmienności miary to  $(-\infty, 1]$ . Miara *Cohen* mierzy siłę zależności pomiędzy dwiema parami zdarzeń: *przykład x pokrywa regułę a przykład x należy do klasy decyzyjnej opisywanej przez regułę oraz przykład x nie pokrywa reguły a przykład x nie należy do klasy decyzyjnej opisywanej przez regułę*. Zakładając, że oceniana reguła jest postaci  $\varphi \rightarrow \psi$ ,

miara *Cohen* może być zapisana jako 
$$\frac{P(\varphi \wedge \psi) + P(\neg\varphi \wedge \neg\psi) - P(\varphi)P(\psi) - P(\neg\varphi)P(\neg\psi)}{1 - P(\varphi)P(\psi) - P(\neg\varphi)P(\neg\psi)}.$$

Mianownik pełni rolę normalizującą, dzięki czemu zakres zmienności miary to  $[-1, 1]$ , przy czym wartość  $-1$  miara osiąga w przypadku identycznej liczby przykładów pozytywnych i negatywnych. Zakres zmienności miar *C1*, *C2* to przedział  $(-\infty, 1]$ . Obie miary, *C1* i *C2*, są miarami konfirmacji. W zależności od rozmiaru zbioru przykładów minimalna wartość miary będzie się zmieniała, wartość maksymalna wynosi zawsze  $1$ , niezależnie od liczebności przykładów pozytywnych i negatywnych.

Miarę *J* zaproponowali Smyth i Goldman [279], opisując algorytm indukcji reguł ITRULE. Miara *J* wywodzi się z teorii informacji. Można ją zdekomponować na dwa składniki [279], z których jeden odpowiada za złożoność reguły, a drugi mierzy różnicę pomiędzy rozkładem liczby przykładów pozytywnych i negatywnych w całym zbiorze treningowym i zbiorze przykładów pokrywanych przez regułę. W ten sposób idea miary *J* zgodna jest z wywodzącą się z teorii informacji zasadą minimalnej długości opisu, gdyż mierzy zarówno złożoność, jak i dokładność reguły. Złożoność mierzy się, obliczając  $P(\varphi \wedge \psi)$ , natomiast dokładność odzwierciedlana jest za pomocą tzw. entropii krzyżowej

(ang. *cross entropy*). Iloczyn obu tych składników prowadzi do miary  $J$ , którą w języku prawdopodobieństwa można zapisać jako  $P(\varphi \wedge \psi) \ln\left(\frac{P(\psi | \varphi)}{P(\psi)}\right) + P(\varphi \wedge \neg\psi) \ln\left(\frac{P(\neg\psi | \varphi)}{P(\neg\psi)}\right)$ .

W algorytmie CN2 równoważna postać miary  $J$  stosowana jest do eliminacji reguł nieistotnych. Miara  $J$  przyjmuje wartości w przedziale  $[0,1]$ . Za pomocą podstawowej postaci miary  $J$  nie można oceniać reguł dokładnych, gdyż dla  $n=0$  jej wartość jest nieokreślona. Nieznaczną modyfikację miary  $J$ , umożliwiającą ocenę reguł dokładnych, przedstawiono w podrozdziale 3.1.6.

Miara *Corr* mierzy korelację pomiędzy rozkładem pokrywanych przez nią przykładów pozytywnych i negatywnych (komórki w tablicy kontyngencji) a rzeczywistymi przyporządkowaniami przykładów do klas pozytywnej i negatywnej (podsumowania wierszy i kolumn w tablicy kontyngencji). Licznik *Corr* może być zapisany jako  $p(N-n)-(P-p)n$ , gdzie odjemna jest iloczynem wartości  $|TP|$  i  $|TN|$ , natomiast odjemnik jest iloczynem wartości  $|FP|$  i  $|FN|$ . W mianowniku umieszczony jest iloczyn wartości będących podsumowaniami wierszy i kolumn tablicy kontyngencji. Zakres zmienności *Corr* to przedział  $[-1,1]$ . W artykule Fürnkranza i Flacha [90] wykazano, że pomiędzy miarami *Corr* i  $\chi^2$  zachodzi następująca zależność:  $\chi^2 = (P+N)\text{Corr}^2$  [90, 307]. Miary *Corr* i  $\chi^2$  nie są równoważne ze względu na sposób rozstrzygania konfliktów klasyfikacji. Miary *Corr* używa się również do oceny reguł asocjacyjnych [97, 128, 307]. W literaturze dotyczącej oceny reguł asocjacyjnych miara *Corr* bardziej znana jest jako  $\phi$ -coefficient. Może być ona zapisana jako  $\frac{P(\varphi \wedge \psi) - P(\varphi)P(\psi)}{\sqrt{P(\varphi)P(\psi)(1-P(\varphi))(1-P(\psi))}}$  i jest miarą konfirmacji.

Geneza miar  $g$  i  $wLap$  jest podobna. Zadaniem ich jest estymacja dokładności reguły. Obie miary wprowadzają korektę na dokładność reguły, obliczoną na podstawie treningowego zbioru przykładów. Dodatkowo  $wLap$  w ocenie tej bierze pod uwagę rozkład liczby przykładów pozytywnych i negatywnych. Ze względu na współczynnik  $(P+N)/P$   $wLap$  traktowana jest jako ważona wersja miary *Laplace*. Odnosząc się do prawdopodobieństwa, miara  $g$  wprowadza korektę do prawdopodobieństwa  $P(\psi | \varphi)$ , natomiast  $wLap$  wprowadza korektę do prawdopodobieństwa  $\frac{P(\psi | \varphi)}{P(\psi)}$ . Zakresem zmienności miary  $g$  jest przedział  $[0,1]$ , miara  $wLap$  może przyjmować wartości w przedziale  $[0, \infty)$ . Miara  $g$  jest jedyną z miar należących do zbioru  $\text{Max} \cup \{\text{CN2}\}$ , która w ocenie nie bierze pod uwagę wartości  $P$  i  $N$ . Miary  $g$  i  $wLap$  mają charakter empiryczny, ocena wykonywana przez te miary zakłada m.in. to, że dokładność reguły pokrywającej małą liczbę przykładów musi być korygowana. Wraz ze wzrostem liczby przykładów pozytywnych

pokrywanych przez regułę korekcja ta powinna być coraz mniejsza. Celowość wprowadzania korekcji potwierdzają m.in. wyniki badań Fürnkranza [88], który porównywał dokładność reguł w zbiorach treningowych i testowych dla różnych, w tym generowanych w sposób syntetyczny, zbiorów danych. Podobne wyniki uzyskał Jaworski [146], podejmując próbę statystycznego oszacowania dokładności i pokrycia reguł decyzyjnych, generowanych w standardowym modelu zbiorów przybliżonych. Definiując formułę korygującą empiryczną dokładność i pokrycie reguły, Jaworski wykorzystał założenia obowiązujące w statystycznej teorii uczenia [321] oraz nierówność Hoeffdinga [133]. Wartość otrzymanej przez niego formuły korygującej uzależniona jest od liczby przykładów pozytywnych pokrywanych przez regułę oraz rozmiaru klasy decyzyjnej, którą reguła ta opisuje. Korekcja jest tym mniejsza, im więcej przykładów pozytywnych pokrywa reguła.

Miara  $LS$  mierzy następujące prawdopodobieństwo:  $\frac{P(\varphi \wedge \psi | \psi)}{P(\varphi \wedge \neg\psi | \neg\psi)}$ . Była ona stosowana zarówno do oceny atrakcyjności reguł, jak i w indukcji reguł dla celów klasyfikacyjnych [42, 70, 151, 340]. Zakres zmienności  $LS$  to przedział  $[0, \infty)$ , a wartości mniejsze od 1 oznaczają, że dokładność reguły jest mniejsza od  $P/(P + N)$ . Wartość miary jest nieokreślona dla reguł dokładnych. Modyfikację  $LS$ , umożliwiającą ocenę reguł dokładnych, przedstawiono w podrozdziale 3.1.6.

Miara  $MS$ , przedstawiona m.in. przez Yao i Zhonga [340, 341], mierzy siłę równoważności  $\varphi \leftrightarrow \psi$ , gdzie  $\varphi$  jest przesłanką reguły, a  $\psi$  jest jej konkluzją, przy czym nie chodzi tutaj o równoważność w sensie logicznym, a równoważność w sensie wzajemnego zawierania się zbiorów przykładów pokrywanych przez  $\varphi$  i  $\psi$ . W języku prawdopodobieństwa  $MS$  można zapisać jako  $\frac{P(\varphi \wedge \psi)}{P(\varphi \wedge \neg\psi) + P(\psi)}$ . Zakres zmienności  $MS$  to przedział  $[0,1]$ .

Tabela 4.29  
 Zwycięstwa miar zawartych w zbiorze  $Max \cup \{CN2\}$   
 ze względu na całkowitą dokładność klasyfikacji

Zbior	K	R	C1	C2	Corr	Cn2	g	LS	MS	RSS	S	wLap	$\Sigma$
Labor	2	15	1	1	1	1	1	1	0	1	1	1	9
Ecoli	8	30	1	1	1	1	0	0	1	1	1	1	8
Ionosphere	2	14	1	1	0	1	1	1	1	1	0	1	8
prnn-synth	2	0	1	1	1	0	1	0	1	1	1	1	8
Hepatitis	2	29	1	0	1	1	0	1	1	1	0	1	7
contact-lenses	3	29	0	1	1	1	1	0	1	1	0	0	6
breast-cancer	2	20	1	1	0	0	1	0	1	1	0	1	6
primary-tumor	21	20	1	1	1	0	1	1	0	1	0	0	6
Echokardiogram	2	17	1	1	0	1	0	0	0	1	1	1	6
heart-statlog	2	6	1	0	1	1	1	1	0	1	0	0	6
heart-c	2	4	0	1	1	1	1	0	0	1	0	1	6
Sonar	2	3	1	1	1	1	0	0	1	1	0	0	6
kdd-synthetic-control	6	0	1	1	1	1	0	0	1	0	1	0	6
sick-euthyroid	2	44	1	1	0	0	0	1	0	0	1	1	5
Zoo	7	26	1	1	0	0	1	1	1	0	0	0	5
Flag	4	22	1	1	0	0	0	1	0	0	0	1	5
Glass	6	19	0	1	1	0	0	0	1	0	1	1	5
hungarian-heart-disease	2	14	0	1	1	1	0	0	1	1	0	0	5
horse-colic	2	13	1	1	0	0	1	0	0	0	1	1	5
Soybean	19	8	1	1	0	0	1	0	0	0	1	1	5
bupa-liver-disorders	2	8	1	1	0	0	1	1	0	0	0	1	5
cylinder-bands	2	8	1	1	0	0	1	1	0	0	0	1	5
credit-a	2	6	1	1	0	1	1	0	0	0	1	0	5
Iris	3	0	1	1	0	1	1	0	0	0	1	0	5
auto-mpg	3	29	1	0	0	0	0	1	0	0	1	1	4
credit-g	2	20	1	1	0	1	1	0	0	0	0	0	4
Autos	6	16	1	1	0	0	0	0	0	0	1	1	4
Diabetes	2	15	1	1	0	1	0	0	0	1	0	0	4
Vote	2	11	0	0	0	0	1	1	0	0	1	1	4
Wine	3	7	1	0	1	0	0	0	0	1	0	1	4
mammographic-masses	2	4	1	1	0	1	0	1	0	0	0	0	4
Segment	7	0	0	0	0	0	1	1	0	0	1	1	4
Hypothyroid	4	67	0	0	0	0	1	1	1	0	0	0	3
Cleveland	5	34	0	0	1	1	0	0	1	0	0	0	3
Titanic	2	18	0	1	0	0	1	0	0	0	1	0	3
breast-w	2	16	1	0	0	1	0	0	1	0	0	0	3
Mushroom	2	2	0	0	0	0	1	1	0	0	0	1	3
Anneal	5	56	1	0	0	0	0	1	0	0	0	0	2
Lymph	4	30	0	0	0	0	0	0	0	1	0	1	2
Yeast	10	21	0	1	0	0	1	0	0	0	0	0	2
Vehicle	4	1	0	0	0	0	1	0	0	0	1	0	2
balance-scale	3	0	0	0	0	1	0	1	0	0	0	0	2
hayes-roth	3	0	0	0	1	1	0	0	0	0	0	0	2
Car	4	45	0	0	0	0	0	1	0	0	0	0	1
Audiology	24	21	0	0	0	0	0	0	0	1	0	0	1
Splice	3	19	0	0	0	0	1	0	0	0	0	0	1
tic-tac-toe	2	15	0	0	0	0	0	1	0	0	0	0	1
kr-vs-kp	2	2	0	0	0	0	0	1	0	0	0	0	1

Miara  $RSS$  [90,309] to różnica wskaźników  $TPR$  i  $FPR$ . W języku prawdopodobieństwa  $RSS$  można zapisać jako  $\frac{P(\varphi \wedge \psi)}{P(\psi)} - \frac{P(\varphi \wedge \neg\psi)}{P(\neg\psi)}$ . Zakres zmienności  $RSS$  to przedział  $[-1,1]$ ,  $RSS$  jest miarą konfirmacji.

Ostatnią z omawianych miar jest miara konfirmacji  $s$ . Zaproponowali ją Christensen [51] i Joyce [149], a do oceny atrakcyjności reguł użyli jej Greco, Pawlak i Słowiński [106]. Głębszą analizę własności  $s$  z punktu widzenia zastosowania jej do oceny atrakcyjności reguł zawierają prace Brzezińskiej, Greco i Słowińskiego [43, 300]. Do chwili obecnej nie przedstawiono wyników stosowania miary  $s$  do nadzorowania procesu indukcji reguł. Wyniki takie przedstawił autor wspólnie z Wróblem w pracach [261, 264]. W języku prawdopodobieństwa  $s$  można zapisać jako  $P(\psi | \varphi) - P(\psi | \neg\varphi)$ . Zakres zmienności  $s$  to przedział  $[-1,1]$ .

Wykresy wszystkich rozważanych miar zamieszczono w dodatku A, w dodatku B zaprezentowano natomiast wyniki wstępnych prac, polegających na badaniu efektywności miar należących do zbioru  $Max$  w nadzorowaniu indukcji reguł regresyjnych.

Jak już wspomniano na początku rozdziału, autorowi [253] oraz innym badaczom [8, 143] nie udało się powiązać wyników miar jakości z charakterystyką zbioru przykładów treningowych. Zdaniem autora adaptacyjna metoda doboru miary na podstawie zbioru  $Max$ , rozszerzonego ewentualnie zmodyfikowaną miarą  $J$  ( $CN2$ ), jest w chwili obecnej najlepszym sposobem doboru miary do konkretnego zbioru przykładów. Dla badaczy chcących dalej zajmować się tą tematyką interesującą może być zawartość tabeli 4.29. W tabeli tej zamieszczono informacje o zwycięstwach i porażkach miar należących do zbioru  $Max \cup \{CN2\}$  na wszystkich 48 rozważanych w tym rozdziale zbiorach przykładów. Symbol 0 oznacza, że na konkretnym zbiorze przykładów miara przegrała z co najmniej jedną inną miarą zawartą w zbiorze  $Max \cup \{CN2\}$ . Symbol 1 oznacza, że zastosowanie miary prowadziło do otrzymania zwycięskiego klasyfikatora. W kolumnie  $K$  zamieszczono informację o liczbie klas decyzyjnych, a w kolumnie  $R$  – informację o równomierności rozkładu przykładów pomiędzy klasami decyzyjnymi. Wartości  $R$  obliczono zgodnie ze wzorem  $R = M - (K / 100)$ , gdzie  $M$  jest wyrażonym w procentach rozmiarem najliczniejszej klasy decyzyjnej, a  $K$  jest liczbą klas decyzyjnych. W kolumnie oznaczonej jako  $\Sigma$  zamieszczono informację o tym, ile miar wygrywało na danym zbiorze przykładów.

## **5. WYBRANE METODY PRZYCINANIA REGUŁ DECYZYJNYCH I ICH ZBIORÓW**

W indukcji reguł decyzyjnych, definiowanych dla celów klasyfikacji, przycinanie realizowane jest w dwóch fazach. Pierwsza faza, zwana *prepruningiem*, realizowana jest w ramach indukcji i jej zadaniem jest zapobieganie nadmiernemu dopasowaniu reguł do danych treningowych. Ten etap przycinania może być utożsamiany z fazami wzrostu i przycinania reguł. Druga faza przycinania, zwana *postpruningiem* polega na analizie wyznaczonych reguł i przekształceniu ich do takiej postaci, która lepiej opisuje i/lub klasyfikuje dane. Jej istotą jest eliminacja i/lub zmiana postaci regułów. W *postpruningu* bardzo często pod uwagę brana jest nie tylko jakość pojedynczych regułów, ale również wzajemne interakcje pomiędzy nimi. W szczególności pod uwagę bierze się, jak zmiana postaci pojedynczej reguły wpływa na jakość opisu i klasyfikacji całego zbioru regułów. Obie fazy mogą być łączone. Przykładowo w algorytmie TDP [86], w fazie *prepruningu*, spośród hipotez (w szczególności zbiorów regułów) o podobnych zdolnościach klasyfikacyjnych wybierana jest maksymalnie złożona hipoteza, która następnie podlega przycinaniu w fazie *postpruningu*. W pracach przeglądowych Fürnkranz [86] i Bruha [41] przedstawiają typowe metody postępowania dla obu wymienionych etapów przycinania (omówiono tam m.in. algorytmy REP, Grow oraz TDP). Zasady przycinania przedstawione w przywoływanych publikacjach realizowane są do tej pory w wielu algorytmach.

W *postpruningu* oraz w algorytmach łączących *pre-* i *postpruning* kryterium optymalności zbioru regułów to najczęściej dokładność klasyfikacji, która obliczana jest na podstawie walidacyjnego zbioru przykładów. Stosowana jest również zasada minimalnej długości kodu [57, 238, 222, 297, 298]. Przycinanie przebiega w taki sposób, aby zminimalizować długość kodu, potrzebną do zapisania zbioru regułów i zbioru przykładów klasyfikowanych błędnie lub niepokrywanych przez przycinane reguły.

Z przycinaniem, zwłaszcza z fazą *postpruningu*, mamy również do czynienia w indukcji regułów decyzyjnych dla celów opisowych oraz w indukcji regułów asocjacyjnych [342]. W zadaniach tych przycinanie najczęściej polega na eliminacji regułów niespełniających wymogów minimalnej jakości lub regułów podobnych. Do identyfikacji grup regułów podobnych stosowane są techniki grupowania [129, 314]. W pracy [23] przedstawiono także metodę

generowania tzw. metareguł asocjacyjnych. Na podstawie metaregułów identyfikowane są grupy reguł podobnych, które następnie są przycinane. Metareguły definiuje się dla zbioru reguł bazowych o identycznych konkluzjach. Przesłanka i konkluzja metareguły zbudowane są z jednej reguły bazowej. Metaregula  $r_i \rightarrow r_j$  interpretowana jest jako stwierdzenie, że przykłady pokrywane przez  $r_i$  pokrywane są również przez  $r_j$ . Dokładność metareguły określa stopień zależności pomiędzy  $r_i$  i  $r_j$ . Metareguły mające część wspólną (identyczną przesłankę lub konkluzję) tworzą grupy, w obrębie których realizowany jest proces przycinania.

Do wyboru reguł najbardziej istotnych stosowano także teorię zbiorów przybliżonych. W pracy [182] przedstawiono metodę *postpruningu*, wykorzystującą znane z teorii zbiorów przybliżonych pojęcie reduktu. W metodzie tej istniejący zbiór reguł jest podstawą do zdefiniowania nowej tablicy decyzyjnej. Każda reguła  $r$  traktowana jest jako atrybut o wartościach binarnych. Jeśli przykład treningowy  $x$  pokrywa regułę, to  $r(x)=1$ ; jeśli przykład nie pokrywa reguły, to  $r(x)=0$ . Na podstawie reduktów wyznaczonych w nowej tablicy identyfikowany jest zbiór najbardziej istotnych reguł.

W indukcji reguł asocjacyjnych stosowana jest również faza *prepruningu*, polegająca na rezygnacji z rozszerzania tych koniunkcji warunków elementarnych, które nie mają szans, aby na ich podstawie uzyskać reguły spełniające wymogi minimalnej jakości. Zabieg taki znacznie przyspiesza proces indukcji [285]. Inne podejście polega na indukcji tzw. reguł optymalnych (k-optymalnych, Pareto-optymalnych), spełniających zadane przez użytkownika ograniczenia [181, 328]. Proces przycinania jest wtedy nierozerwalną częścią algorytmu indukcji.

Faza *prepruningu* łączy się bezpośrednio z konkretnym algorytmem indukcji reguł, natomiast faza *postpruningu* może być realizowana niezależnie od algorytmu indukcji. Jeśli obie fazy się przeplatają, to są one również związane z konkretnym algorytmem indukcji reguł (np. rozdział 2, opis algorytmu RIPPER).

Niniejszy rozdział związany jest z metodami *postpruningu*, które są niezależne od algorytmu indukcji reguł. Dzięki temu prezentowane metody mogą zostać użyte do przycinania dowolnego zbioru reguł decyzyjnych. Pierwsze dwie metody związane są z generalizacją, czyli modyfikacją reguł, mającą na celu zwiększenie liczby pokrywanych przez nie przykładów pozytywnych. W ostatniej części rozdziału przedstawiono kilka propozycji algorytmów filtracji zbiorów reguł decyzyjnych, przeznaczonych do klasyfikacji. Algorytmy filtracji nie ingerują w budowę reguł, ich nadzorem celem jest usunięcie reguł zbędnych. Zbędnymi mogą być reguły pokrywające podobne zbiory przykładów (zwłaszcza przykładów pozytywnych) lub niemające większego wpływu na wyniki klasyfikacji.

## 5.1. Agregacja reguł

Agregacja reguł decyzyjnych, zwana również sklejaniem lub łączeniem reguł, jest jedną z metod przycinania. Polega ona na łączeniu reguł charakteryzujących się pewnymi podobnymi cechami. Na problem agregacji można spojrzeć jako na problem uogólniania reguł.

Agregacja to łączenie dwóch lub większej liczby reguł podobnych. Najprostsze podejście do agregacji polega na sklejaniu odpowiadających sobie zakresów warunków elementarnych, znajdujących się w łączonych regułach. W pracy [248] autor przedstawił iteracyjny algorytm sklejania reguł decyzyjnych. Algorytm ten ma następujące cechy charakterystyczne: po sklejeniu nie zmienia się język reprezentacji reguł, sklejaniu mogą podlegać jedynie reguły zawierające proste warunki elementarne i podlegają mu reguły o podobnej budowie syntaktycznej. Dokładniej algorytm ten zostanie zaprezentowany w dalszej części rozdziału.

W pracach Latkowskiego i Mikołajczyka [174, 197] przedstawiono podobną metodę, przy czym przed sklejaniem reguły są grupowane [197] lub oblicza się ich wzajemne podobieństwo (algorytm *LRJ – Linear Rule Joining*) [174]. Sklejaniu podlegają reguły należące do tej samej grupy lub wystarczająco do siebie podobne. Algorytm *LRJ* dopuszcza, aby w łączonych warunkach elementarnych pojawiła się tzw. wewnętrzna alternatywa [189]. Oznacza to, że po złączeniu warunków  $a \in [v_{11}, v_{12}]$  oraz  $a \in [v_{21}, v_{22}]$ , gdzie  $v_{12} < v_{21}$ , w sklejonej regule może wystąpić warunek ( $a \in [v_{11}, v_{12}] \vee a \in [v_{21}, v_{22}]$ ). Dla algorytmu *LRJ* nie jest istotne, czy warunki elementarne sklejanych reguł zbudowane są z tego samego podzbioru atrybutów warunkowych. Takie założenie oznacza, że przesłanka sklejonej reguły może składać się z większej liczby warunków elementarnych niż dłuższa z reguł podlegających sklejaniu.

Szczególny sposób agregacji reguł przedstawia algorytm zaproponowany przez Pindura, Stefanowskiego i Susmagę [224]. W algorytmie tym autorzy wprowadzają do przesłanek złożone (skośne) warunki elementarne. Warunki skośne są zbudowane na podstawie atrybutów tworzących warunki proste w regułach podlegających agregacji. Algorytm wykorzystuje granice zakresów warunków elementarnych do wyznaczenia wartości współczynników definiujących warunki skośne. Proponowany algorytm nie nadaje się do agregacji klasycznych reguł decyzyjnych. Można go jedynie zastosować do reguł, w których zakresy wszystkich warunków elementarnych są ograniczone albo od góry, albo od dołu (w zamyśle autorów była agregacja reguł wyznaczanych w modelu zbiorów przybliżonych, definiowanym na podstawie relacji dominacji [104]).

Warto wspomnieć, chociaż nie ma to bezpośredniego związku z agregacją reguł, że istnieją także nieliczne algorytmy umożliwiające indukcję reguł zawierających skośne warunki elementarne. Przykładem takiego algorytmu jest ADReD [233], który tworzy warunki skośne, będące kombinacją liniową wszystkich atrybutów warunkowych. Otrzymane

w ten sposób reguły są trudne do interpretacji, algorytm charakteryzuje się dużą złożonością obliczeniową, a jego efektywność została przebadana jedynie na kilku zbiorach danych. Częściej można zetknąć się z indukcją drzew decyzyjnych, zawierających w węzłach warunki skośne [21, 161, 205]. W celu indukcji drzew zawierających warunki skośne stosowano różnego rodzaju heurystyczne strategie przeszukiwania, algorytm genetyczny [161, 205] oraz liniowe maszyny wektorów podpierających [21]. Maszyny wektorów podpierających były również stosowane do indukcji reguł zawierających elipsoidalne warunki elementarne [64]. Inne podejście pozwalające na uzyskanie warunków skośnych polega na wprowadzeniu do zbioru atrybutów warunkowych nowych atrybutów, których wartości są kombinacjami liniowymi wartości atrybutów istniejących. Przeprowadzenie, na podstawie tak rozszerzonego zbioru atrybutów, standardowej indukcji reguł umożliwia wyznaczenie reguł zawierających warunki skośne [28, 302].

W dalszej części rozdziału przedstawione zostaną dwa algorytmy agregacji reguł. Pierwszy skleja reguły o podobnej budowie syntaktycznej i nie dopuszcza do zmiany języka reprezentacji reguł. Drugi dokonuje agregacji reguł na podstawie informacji o otoczce wypukłej, zawierającej łączone reguły.

### **5.1.1. Agregacja przez łączenie zakresów warunków elementarnych**

Podstawowym założeniem algorytmu agregacji reguł przez łączenie zakresów warunków elementarnych jest to, że łączone są reguły zbudowane z prostych warunków elementarnych oraz wskazujące na identyczną klasę decyzyjną. Założymy, że dane są dwie reguły:  $r_1 \equiv \varphi_1 \rightarrow \psi$ ,  $r_2 \equiv \varphi_2 \rightarrow \psi$ . Przez  $attr(r)$  oznaczmy zbiór atrybutów, na podstawie którego zbudowano warunki elementarne, należące do  $Ec(r)$ . Algorytm podejmie próbę połączenia reguł  $r_1$  i  $r_2$  jedynie wtedy, gdy  $attr(r_1) \subseteq attr(r_2)$  lub  $attr(r_2) \subseteq attr(r_1)$ . Innymi słowy, łączone są jedynie te reguły, których przesłanki zbudowane są na podstawie podobnego zbioru atrybutów warunkowych.

Agregacja polega na łączeniu zakresów wartości odpowiadających sobie warunków. Jeśli  $a op Za^{r^1}$  jest elementem przesłanki reguły  $r_1$  oraz  $a op Za^{r^2}$  występuje w części warunkowej reguły  $r_2$ , to po połączeniu tych warunków powstanie  $a op Za^{r^{1r^2}}$  o własności  $Za^{r^1} \subseteq Za^{r^{1r^2}}$  i  $Za^{r^2} \subseteq Za^{r^{1r^2}}$ . Ponieważ wynikiem agregacji jest reguła zbudowana z prostych warunków elementarnych, operator relacyjny  $op$  będziemy utożsamiać z  $\in$ . Przykładowo,  $a = v$  może zostać zapisany jako  $a \in \{v\}$ , a  $a > v$  można zapisać jako  $a \in (v, \infty)$ .

W zagregowanej regule zakres  $Za^{r1r2}$  definiowany jest następująco:

- jeśli  $a$  jest typu symbolicznego lub porządkowego i złączeniu podlegają warunki  $a \in Za^{r1}$  oraz  $a \in Za^{r2}$ , to  $Za^{r1r2} = Za^{r1} \cup Za^{r2}$ ,
- jeśli atrybut jest typu ciągłego i złączeniu podlegają warunki  $a \in (v_a^1, v_a^2)$  oraz  $a \in (v_a^3, v_a^4)$ , to wynikiem złączenia jest  $a \in (v_a^{\min}, v_a^{\max})$ , gdzie  $v_a^{\min} = \min\{v_a^1, v_a^3\}$ ,  $v_a^{\max} = \max\{v_a^2, v_a^4\}$ .

Pozostałe założenia dotyczą sterowania procesem łączenia reguł. Można je przedstawić w postaci następujących punktów:

- reguły agregowane są parami; regułą bazową, stanowiącą podstawę dla nowej zagregowanej reguły, jest reguła o lepszej jakości (wyższej wartości miary  $q$ ),
- warunki elementarne sklejane są sekwencyjnie, o kolejności sklejania decyduje wartość miary oceniającej nowo tworzoną regułę; aby wskazać najlepszy w danej chwili warunek do sklejenia, stosuje się strategię wspinaczki,
- proces sklejania kończy się z chwilą, gdy nowa, zagregowana reguła pokrywa wszystkie przykłady pozytywne, pokrywane przez reguły podlegające agregacji,
- założymy, że agregacji podlegają reguły  $r_1$  i  $r_2$  oraz  $r_1$  jest regułą bazową, a  $r$  jest regułą będącą wynikiem agregacji; wówczas reguły  $r_1$  i  $r_2$  zastępowane są przez regułę  $r$  wtedy i tylko wtedy, gdy  $q(r) \geq q(r_1) - \lambda$  (standardowo  $\lambda=0$ ).

Na podstawie przedstawionych założeń można zaproponować następujący algorytm:

Algorytm agregacji reguł decyzyjnych przez łączenie zakresów warunków elementarnych

Wejście:  $U$  – zbiór przykładów

$RUL$  – zbiór reguł wskazujących na klasę decyzyjną  $X_i$

$q$  – miara jakości reguł decyzyjnych

$\lambda$  – parametr decydujący o tym, o ile może zmaleć jakość reguł

Wyjście:  $aggRUL$  – wyjściowy zbiór reguł

1. **Begin**
2.  $aggRUL := \emptyset;$
3. posortuj reguły ze zbioru  $RUL$  nierosnąco, zgodnie z wartością miary  $q$ , obliczoną na zbiorze przykładów  $U$ , wynik umieść na liście  
 $RUL\_l := <r_1, r_2, \dots, r_{|RUL|}>;$
4. **Repeat**
5.  $minq := q(r_1) - \lambda;$
6. **ForEach**  $i > 1$
7.     **If**  $(attr(r_i) \subseteq attr(r_1) \text{ or } attr(r_1) \subseteq attr(r_i))$
8.         **then**
9.             **If**  $Sklej(r_1, r_i, q, minq)$  **then**  $RUL := RUL - \{r_i\}$
10.         **end foreach**
11.      $RUL := RUL - \{r_1\};$
12.      $aggRUL := aggRUL \cup \{r_1\};$
13.     przenuumeruj reguły na liście  $RUL\_l$ , wynik zapisz jako  
 $RUL\_l := <r_1, r_2, \dots, r_{|RUL|}>;$
14. **Until**  $RUL \neq \emptyset$
15. **end.**

Podstawą agregacji jest zawsze pierwsza reguła znajdująca się na liście, do niej doklejane są kolejne reguły z listy (linia 9). Jeśli proces agregacji przebiegnie pomyślnie, doklejona reguła usuwana jest ze zbioru reguł, a nowa zagregowana reguła staje się regułą bazową dla kolejnych kandydujących do sklejenia reguł (pętla 6–10). Po rozpatrzeniu wszystkich reguł z listy reguła będąca wynikiem agregacji usuwana jest z wejściowego zbioru reguł (linia 11) i dodawana do zbioru zawierającego zagregowane reguły (linia 12). Oczywiście może zdarzyć się tak, że dla ustalonej reguły bazowej każda próba agregacji zakończy się niepowodzeniem, wówczas reguła bazowa przekazywana jest do wynikowego zbioru bez żadnych modyfikacji. Postępowanie takie zapewnia, że jeśli tylko wejściowy zbiór reguł pokrywał wszystkie przykłady należące do rozpatrywanej klasy  $X_i$ , to wyjściowy zbiór reguł również pokrywa te przykłady.

### Funkcja sklejająca dwie reguły

```

Parametry:      r1, r2 - reguły podlegające sklejeniu
                q - miara jakości reguł
                minq - minimalna jakość sklejonej reguły
Założenie:      q(r1) ≥ q(r2),
Function Sklej(r1, r2, q, minq): Boolean
1. Begin
2.   rtemp:=r1;
    // attr zawiera zbiór atrybutów, na podstawie których zbudowane są warunki
    // elementarne reguł r1 i r2
3.   attr:=attr(r1) ∪ attr(r2);
4.   Repeat
5.     qmax:=-∞;           // minimalna wartość miary q
6.     wmax:=null; amax:=∅    // pusty warunek elementarny, brak atrybutu amax

    // wspinaczka - sprawdzenie, które warunki najlepiej w danej chwili skleić
7.   Foreach a∈attr do
8.     Zastąp w regule r1 warunek w, zbudowany na podstawie a, warunkiem
        powstały przez połączenie zakresów w oraz odpowiadającego mu
        warunku w regule r2
9.     If qmax<q(r1) then
10.       wmax:=w; amax:=a;
11.       qmax:=q(r1);
12.       wróć do pierwotnej postaci warunku w w regule r1
13.   end foreach

    // sklejenie najlepszych w danej chwili warunków elementarnych
14.   Zastąp w regule r1 warunek w warunkiem wmax;
15.   attr:=attr-amax;
16. Until (([rtemp] ∪ [r2]) ⊆ [r1]) // sklejona reguła musi pokrywać wszystkie
    // przykłady pozytywne pokrywane przez
    // reguły podlegające sklejaniu

    // sprawdzenie, czy reguła spełnia wymóg minimalnej jakości
17. If q(r1) ≥ minq then return true
18. else
19.   r1:=rtemp;
20.   return false
21. end.

```

Zadaniem funkcji  $\text{Sklej}(r_1, r_2, q, \min q)$  jest połączenie reguł  $r_1$  i  $r_2$ . Proces łączenia polega na doklejeniu reguły  $r_2$  do reguły bazowej  $r_1$ . Proces ten kończy się sukcesem, jeśli zagregowana reguła charakteryzuje się jakością wyższą niż  $\min q$ . Zauważmy, że w szczególnym przypadku reguła  $r_1$  może składać się z mniejszej liczby warunków elementarnych niż reguła  $r_2$ . W czasie działania algorytmu ze zbioru  $\text{attr}$  będą wybierane również takie atrybuty  $a$ , które nie występują w żadnym z warunków reguły  $r_1$  (pętla 7–13). W takiej sytuacji zakłada się, że reguła  $r_1$  zawiera warunek uniwersalny  $a \in Va$  (gdzie  $Va$  jest zbiorem wszystkich wartości atrybutu  $a$ ).

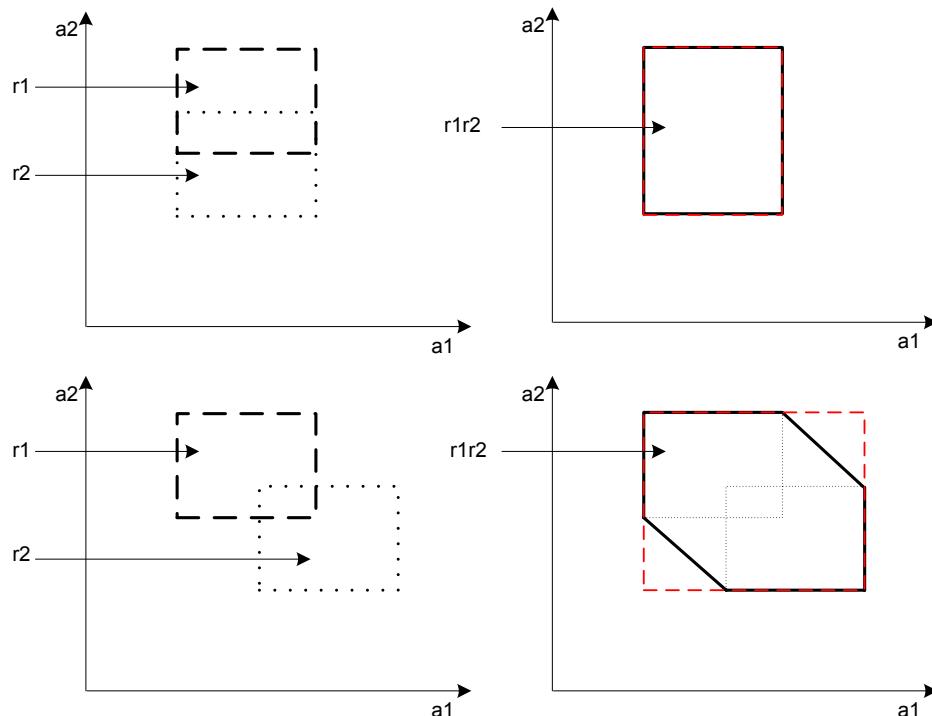
Algorytm można poddać modyfikacji polegającej na tym, iż w pierwszej kolejności agregowane będą reguły pokrywające jak najbardziej rozłączne zbiory przykładów pozytywnych i możliwie identyczne zbiory przykładów negatywnych. Wiąże się to jednak z wydłużeniem czasu działania algorytmu. Złożoność obliczeniowa algorytmu zależy od złożoności metody wspinaczki, stosowanej w trakcie próbnego sklejania reguł. Jeśli przez  $L$  oznaczymy liczbę reguł, przez  $n$  – liczbę przykładów, a przez  $m$  – liczbę atrybutów warunkowych, to złożoność obliczeniowa algorytmu jest rzędu  $O(Lm^3n)$ .

### 5.1.2. Agregacja na podstawie otoczki wypukłej, zawierającej łączone reguły

Wynikiem działania algorytmu dokonującego agregacji reguł na podstawie otoczki wypukłej może być reguła zawierająca skośne warunki elementarne. Otrzymujemy w ten sposób reguły o większych możliwościach dyskryminacyjnych, które mogą obejmować większe i nieregularne (skośne) obszary w przestrzeni cech. Chcąc maksymalnie ograniczyć złożoność agregowanych reguł, algorytm tworzy jedynie warunki skośne (nie są stosowane elipsoidy ani inne krzywe wyższych stopni). Główna idea agregacji polega na wykorzystaniu algorytmu wyznaczania otoczki wypukłej [14] na podstawie punktów tworzących granice warunków elementarnych, z których zbudowane są reguły poddawane agregacji. Na podstawie punktów należących do otoczki wypukłej tworzone są równania hiperpłaszczyzn ograniczających wszystkie przykłady pozytywne pokrywane przez reguły. Warunek skośny tworzą jedynie atrybuty ciągłe, agregacja warunków zbudowanych na podstawie atrybutów symbolicznych i porządkowych odbywa się tak jak w algorytmie agregacji – przez łączenie zakresów warunków elementarnych. Reguły agregowane są parami. Prezentowany algorytm nosi nazwę CHIRA i został opracowany przez autora wspólnie z Gudsiem (ang. Convex Hull Based Iterative Algorithm of Rules Aggregation) [265].

Na rysunku 5.1 przedstawiono różnice pomiędzy algorytmem dokonującym agregacji poprzez łączenie warunków elementarnych a algorytmem CHIRA.

Otoczka wypukła wyznaczana jest dla zbioru punktów definiujących granice zakresów ciągłych warunków elementarnych. Oznacza to, że agregacja dotyczy od razu wszystkich ciągłych warunków elementarnych, należących do przesłanek agregowanych reguł. Aby wyznaczyć te punkty, każda reguła poddana agregacji musi zostać przekształcona do postaci hipersześcianu w  $l$ -wymiarowej przestrzeni ( $l$  jest liczbą atrybutów numerycznych, występujących w przesłance reguły). Reguła ma postać hipersześcianu, jeśli każdy z jej ciągłych warunków elementarnych jest obustronnie ograniczony. Przykładowo, warunki  $a \geq v$  lub  $a \leq v$  zastępowane są warunkami  $a \in [v, \max_a]$ ,  $a \in [\min_a, v]$ , gdzie  $\min_a$  i  $\max_a$  oznaczają odpowiednio minimalną i maksymalną wartość atrybutu  $a$  w rozważanym zbiorze przykładów. Wartości znajdujące się na początku i końcu przedziału definiującego zakres warunku elementarnego nazwiemy punktami granicznymi warunku elementarnego. Regułę  $r$ , będącą w postaci hipersześcianu, można domknąć, co będzie oznaczało dodanie do jej przesłanki uniwersalnych warunków elementarnych  $a \in [\min_a, \max_a]$ , gdzie  $a$  jest atrybutem ciągłym, takim że  $a \notin \text{attr}(r)$ .



Rys. 5.1. Przykład agregacji reguł zbudowanych z ciągłych warunków elementarnych.

Po lewej stronie znajdują się zarysy reguł podlegających agregacji, po stronie prawej widoczne są zarysy reguły wynikowej. Linia przerywana oznacza regułę utworzoną przez proste złączenie zakresów warunków elementarnych, linia ciągła oznacza regułę utworzoną przez algorytm CHIRA

Fig. 5.1. The example of joining rules that include two *continuous* elementary conditions.

On the left there are outlines of rules subject to aggregation, on the right there are outlines of the output rule. The broken line indicates the rule created by simple joining ranges of elementary conditions, the solid line means the rule obtained through the CHIRA algorithm

Algorytm CHIRA może działać zarówno na uporządkowanym, jak i na nieuporządkowanym zbiorze reguł. Agregacji podlegają reguły wskazujące na identyczną klasę decyzyjną, a kolejność agregacji determinowana jest przez:

- kolejność reguł (dot. zbiorów uporządkowanych),
- ranking reguł, ustanowiony przez wybraną miarę jakości (dot. nieuporządkowanych zbiorów reguł).

W obecnej wersji algorytmu agregacji podlegają jedynie te reguły, w których sumaryczna liczba warunków zbudowanych na podstawie atrybutów ciągłych jest nie większa niż podana przez użytkownika wartość  $\max\text{-dim}$ . Wartość  $\max\text{-dim}$  ogranicza również liczbę atrybutów tworzących skończony warunek elementarny. Ograniczenie to znacznie przyspiesza działanie algorytmu i powoduje, że algorytm wyznaczania otoczki wypukłej uruchamiany jest jednokrotnie.

Można sobie wyobrazić taką konkretyzację algorytmu, w której otoczka wypukła wyznaczana jest jedynie dla pewnego podzbioru ciągłych warunków elementarnych, podczas gdy pozostałe warunki byłyby łączone w sposób prosty (złączenia zakresów) lub pozostawałyby niezmienione. Zakładając, że liczba atrybutów numerycznych w agregowanych regułach wynosi  $p$  oraz dopuszczalna liczba atrybutów w złożonym warunku elementarnym może wynosić  $s$  ( $s \leq p$ ), w pierwszym kroku agregacji należałyby sprawdzić  $\sum_{k=2}^s \binom{p}{k}$  podzbiorów atrybutów i tyle samo razy uruchamiać algorytm wyznaczania

otoczki wypukłej oraz próbować połączyć pozostałe zakresy warunków elementarnych. Takie postępowanie byłoby zbyt czasochłonne i znacznie ograniczałoby możliwości zastosowania algorytmu.

Przyjmijmy następujące oznaczenia:

- $\text{numattr}(r)$  jest zbiorem atrybutów ciągłych, na podstawie których zbudowane są warunki elementarne reguły  $r$ ;
- $\text{symattr}(r)$  jest zbiorem atrybutów symbolicznych i porządkowych, na podstawie których zbudowane są warunki elementarne reguły  $r$ ;
- $H(r)$  jest liczbą hiperpłaszczyzn, których równania występują w domkniętej postaci reguły  $r$ ; hiperpłaszczyzny te tworzą hipersześcian w  $|\text{numattr}(r)|$ -wymiarowej przestrzeni;
- $V(r)$  jest liczbą wierzchołków hipersześcianu opisanego zbiorem hiperpłaszczyzn  $H(r)$ .

Dysponując zbiorem hiperpłaszczyzn  $H(r)$ , można wyznaczyć zbiór wierzchołków  $V(r)$ , możliwe jest również działanie odwrotne. Dla reguł z prostymi warunkami elementarnymi wyznaczenie zbioru  $V(r)$  jest operacją banalną, polegającą na odczytaniu punktów

granicznych ciągłych warunków elementarnych. Dokładniej, jeśli w zbiorze atrybutów znajduje się  $l$  atrybutów ciągłych, to reguła domknięta zawiera  $2l$  punktów granicznych, a zbiór  $V(r)$  uzyskujemy, tworząc wszystkie możliwe ciągi  $\langle lp_1, lp_2, \dots, lp_l \rangle$ , takie że  $\forall i \in \{1, 2, \dots, l\} \quad lpi \in \{zi1, zi2\}$ . Wartości  $zi1$  i  $zi2$  stanowią odpowiednio punkty graniczne początkowy i końcowy w  $i$ -tym warunku elementarnym, zbudowanym na podstawie atrybutu ciągłego.

**Przykład 5.1.** Założmy, że dane są dwie reguły:  $r_1$  i  $r_2$ . Dla uproszczenia przyjmiemy, że nie zawierają one symbolicznych warunków elementarnych. Postaci reguł są następujące:

$$r_1 \equiv a_1 \in [z11, z12] \wedge a_2 \geq z21 \rightarrow d = 1, \quad r_2 \equiv a_3 \in [z31, z32] \wedge a_2 \leq z21 \rightarrow d = 1.$$

Odpowiadające im reguły w postaci hipersześcianów będą miały postaci:

$$a_a \in [z11, z12] \wedge a_2 \in [z21, \max_{a_2}] \rightarrow d = 1,$$

$$a_3 \in [z31, z32] \wedge a_2 \in [\min_{a_2}, z21] \rightarrow d = 1.$$

Domykając regułę  $r_1$  ze względu na atrybut  $a_3$ , uzyskamy:

$$a_a \in [z11, z12] \wedge a_2 \in [z21, \max_{a_2}] \wedge a_3 \in [\min_{a_3}, \max_{a_3}] \rightarrow d = 1.$$

Zbiór  $V(r_1)$  składa się z punktów  $\langle z11, z21 \rangle$ ,  $\langle z11, \max_{a_2} \rangle$ ,  $\langle z12, z21 \rangle$ ,  $\langle z12, \max_{a_2} \rangle$ , natomiast w przypadku domkniętej postaci  $r_1$  do zbioru tego należą punkty  $\langle z11, z21, \min_{a_3} \rangle$ ,  $\langle z11, z21, \max_{a_3} \rangle$ ,  $\langle z11, \max_{a_2}, \min_{a_3} \rangle$ ,  $\langle z11, \max_{a_2}, \max_{a_3} \rangle$ ,  $\langle z12, z21, \min_{a_3} \rangle$ ,  $\langle z12, z21, \max_{a_3} \rangle$ ,  $\langle z12, \max_{a_2}, \min_{a_3} \rangle$ ,  $\langle z12, \max_{a_2}, \max_{a_3} \rangle$ .

Algorytm agregacji reguł na podstawie informacji o otoczce wypukłej, zawierającej agregowane reguły, można przedstawić w następującej, uproszczonej postaci:

Algorytm agregacji reguł na podstawie otoczki wypukłej  
 Wejście:  $U$  – treningowy zbiór przykładów  
 $V$  – walidacyjny zbiór przykładów  
 $RUL$  – zbiór lub lista reguł  
 $q$  – miara jakości  
 $max\text{-}dim$  – maksymalny wymiar agregacji  
 $q\text{-}drop$ ,  $class\text{-}drop$ ,  $pruning\text{-}drop$  – progowe wartości parametrów dostrajania i przycinania reguł  
 Wyjście:  $aggRUL$  – zagregowany zbiór lub lista reguł

1. **Begin**
2.   **Foreach**  $r \in RUL$  **do**  
     doprowadź  $r$  do postaci hipersześciennu
4.   **If**  $RUL$  jest zbiorem reguł **then**  
     Posortuj  $RUL$  w porządku leksykograficznym, najpierw nierośnaco ze względu na wartość miary  $q$ , a następnie nierośnaco ze względu na pokrycie reguł
5.   **Foreach**  $r_i \in RUL$  **do**  
     // agregacja reguł  
     **Foreach**  $r_j \in RUL, j > i$  **do**  
        $r := aggreguj(r_i, r_j, U, max\text{-}dim);$   
       **If**  $r \neq null$  **then**  
          $dostroj\text{-}regule(r, U, q);$   
         **If**  $wymogi\text{-}jakości(r, U, V, q, q\text{-}drop, class\text{-}drop)$  **then**  
            $r_i := r;$   
            $RUL := RUL - \{r_j\}$
13.   // usuwanie reguł pokrywających przykłady pokrywane przez zagregowaną regułę  
     **Foreach**  $r_k, k > j$  **do**  
       **If**  $[r_k] \subseteq [r]$  **then**  $RUL := RUL - \{r_k\};$
15.   **end foreach**
16. **end foreach**
17.   // przycinanie reguł – tylko dla listy reguł  
   **If**  $RUL$  jest listą reguł **then**  $przytnij(RUL, pruning\text{-}drop)$
18.    $aggRUL := RUL;$
19. **end.**

Począwszy od najlepszej (lub pierwszej na liście) reguły podejmowana jest próba agregacji tej reguły ze wszystkimi pozostałymi regułami wskazującymi na tę samą co ona klasę decyzyjną. Jeśli próba połączenia nie zakończy się powodzeniem, wynikiem agregacji jest reguła pusta (linie 7, 8). Jeśli reguły zostały zagregowane, to uruchamiana jest faza dostrajania, polegająca na równoległym przesuwaniu hiperpłaszczyzn ograniczających reguły (linia 9). Faza dostrajania zostanie opisana w dalszej części rozdziału. Po zakończeniu dostrajania algorytm sprawdza, czy reguła spełnia wymogi minimalnej jakości (linia 10). W algorytmie wykorzystano dwa kryteria jakości. Dla listy reguł kryterium tym jest minimalna dokładność klasyfikacji, jaką charakteryzuje się lista, w której agregowane reguły zastąpiono regułą będącą wynikiem agregacji. Dokładność klasyfikacji może być obliczana zarówno na treningowym, jak i walidacyjnym zbiorze przykładów. Jeśli agregacji poddawany jest nieuporządkowany zbiór reguł, to zagregowana reguła musi spełniać wymóg minimalnej

jakości. Domyślnie przyjmuje się, że jakość zagregowanej reguły nie może być gorsza od jakości lepszej z reguł poddawanych złączeniu. Jeśli zagregowana reguła spełnia wymogi nałożone przez funkcję *wymogi-jakosci* (linia 10), to reguła bazowa zastępowana jest regułą zagregowaną, a reguła dołączana do reguły bazowej jest usuwana. Przed próbą dołączenia do reguły bazowej kolejnej reguły sprawdzane jest także, czy zmieniona reguła bazowa przypadkiem nie pokrywa wszystkich przykładów pozytywnych, pokrywanych przez inne, nierozważane jeszcze reguły (linie 13, 14). Jeśli taka sytuacja ma miejsce, reguły „pokrywane przez regułę bazową” są usuwane.

#### 5.1.2.1. Agregacja pary reguł

Przez  $numattr(r)$  oznaczmy zbiór atrybutów ciągłych, na podstawie których zbudowane są warunki elementarne reguły  $r$ . Założymy, że  $r_i$  i  $r_j$  są agregowanymi regułami oraz że wynikiem ich agregacji jest reguła  $r$ .

Procedura agregacji uruchamiana jest jedynie dla reguł wskazujących na identyczną klasę decyzyjną. Procedura agregacji uruchamiana jest oddziennie dla warunków elementarnych, zbudowanych na podstawie atrybutów ciągłych, oraz dla warunków elementarnych zbudowanych na podstawie atrybutów symbolicznych lub porządkowych.

Procedura łączenia ciągłych warunków elementarnych jest następująca:

1. Niech  $numattr(r) = numattr(r_i) \cup numattr(r_j)$  oraz  $|numattr(r)| = k$ . Jeśli  $k > max\_dim$ , to reguły nie są agregowane.
2. Przekształć reguły  $r_i$  i  $r_j$  do postaci hipersześcianu i domknij  $r_i$  ze względu na zbiór atrybutów  $numattr(r_j) \setminus numattr(r_i)$  oraz  $r_j$  ze względu na zbiór atrybutów  $numattr(r_i) \setminus numattr(r_j)$ . Dla domkniętych reguł wyznacz zbiory wierzchołków  $V(r_i)$ ,  $V(r_j)$ .
3. Utwórz  $V(r) = V(r_i) \cup V(r_j)$ . Za pomocą algorytmu Qhull [14] wyznacz  $k$ -wymiarową otoczkę wypukłą punktów należących do  $V(r)$ . Wynikiem działania algorytmu jest zbiór  $F(r)$ , zawierający wszystkie „ściany” otoczki. Każda „ściana” jest reprezentowana przez zbiór  $k$  wierzchołków należących do zbioru  $V(r)$ .
4. Na podstawie punktów definiujących daną „ścianę” wyznacz (rozwiązuając odpowiedni układ równań liniowych) równanie hiperpłaszczyzny zawierającej tę „ścianę”. Uzyskane hiperpłaszczyzny zapisz do zbioru  $H(r)$ .

W przedstawionej procedurze wykorzystano algorytm wyznaczania otoczki wypukłej, zawarty w pakiecie obliczeń inżynierskich Matlab.

Warunki elementarne, zbudowane na podstawie atrybutów symbolicznych i porządkowych, łączone są według takiej samej zasady jak w algorytmie agregacji reguł: przez łączenie zakresów warunków elementarnych.

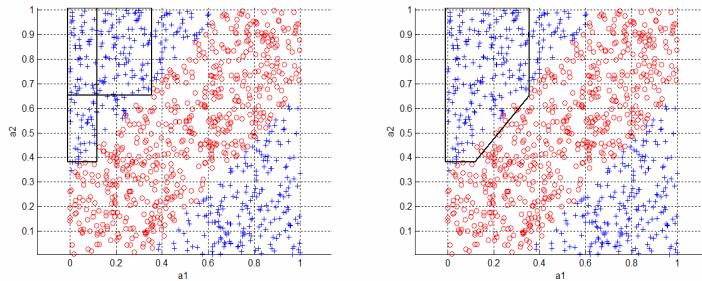
#### 5.1.2.2. Dostrajanie po agregacji

Dostrajanie zagregowanych reguł odbywa się w dwóch etapach. W pierwszym kroku łączone są „ściany” leżące na tej samej hiperpłaszczyźnie. Zastosowany w  $k$ -wymiarowej przestrzeni algorytm Qhull opisuje każdą ze ścian za pomocą  $k$  punktów. Oznacza to, że np. w przypadku sześciianu (przestrzeń 3-wymiarowa), który złożony jest z 6 ścian będących czworokątami, wynikiem działania algorytmu Qhull będzie 12 trójkątnych ścian. Aby pozbyć się nadmiarowych „ścian”, algorytm CHIRA identyfikuje je, łączy i usuwa zduplikowane równania hiperpłaszczyzn.

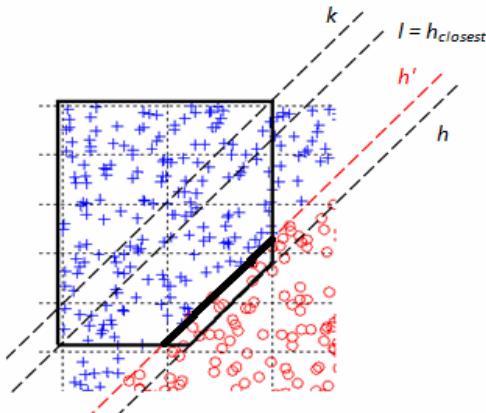
Drugi krok dostrajania polega na modyfikacji równań hiperpłaszczyzn, zawartych w złożonych warunkach elementarnych. Celem modyfikacji jest podniesienie jakości reguł. Na rysunku 5.2 przedstawiono przykładową zagregowaną regułę. Widać, że przesunięcie skośnego warunku elementarnego spowodowałoby, że reguła ta pokrywałaby mniej przykładów negatywnych. W celu dostrojenia skośnych warunków elementarnych zastosowano następującą procedurę:

1. Wybierz jeszcze niedostrojoną hiperpłaszczyznę  $h \in H(r)$ , taką że  $h \notin H(r_i)$  i  $h \notin H(r_j)$  ( $r_i, r_j$  są agregowanymi regułami).
2. Znajdź „ścianę”  $f$ , leżącą na płaszczyźnie  $h$ , oraz zbiór wierzchołków z  $V(r)$ , sąsiednich do  $f$ . Wierzchołek jest sąsiadni do  $f$ , jeśli leży na „ścianie”  $g$ , przylegającej do  $f$ , oraz nie leży na ścianie  $f$ .
3. Wyznacz równania hiperpłaszczyzn równoległych do  $h$  i przechodzących przez co najmniej jeden z wierzchołków sąsiednich do  $f$ . Oznacz przez  $h_{closest}$  hiperpłaszczyznę najbliższą do  $h$ . Hiperpłaszczyzny  $h$  oraz  $h_{closest}$  ograniczają przestrzeń prowadzonych poszukiwań (rys. 5.3).
4. W obszarze ograniczonym przez  $h$  i  $h_{closest}$  zastosuj metodę Fibonacciego [78] do znalezienia optymalnej hiperpłaszczyzny  $h'$ , równoległej do  $h$ . Ponieważ poszukiwana jest hiperpłaszczyzna równoległa do  $h$ , metoda Fibonacciego stosowana jest do ustalenia optymalnej wartości wyrazu wolnego. Celem optymalizacji jest maksymalizacja wartości miary jakości  $q$ .
5. W zbiorze  $H(r)$  zastąp  $h$  przez  $h'$ . Wykorzystując  $h'$  oraz równania „ścian” przylegających do  $f$ , oblicz nowe współrzędne wierzchołków „ściany”  $f$  (rozwiązuje odpowiedni układ równań liniowych).

Zastosowana metoda Fibonacciego nie gwarantuje tego, że w przeszukiwanym zakresie znalezione zostanie globalne maksimum funkcji celu (czyli reguła o najwyższej jakości), gdyż nie wiadomo, czy w zakresie tym funkcja celu ma jedno maksimum lokalne. Jednakże, jak pokazują prace związane z optymalizacją reguł rozmytych [216], metoda Fibonacciego pozwala uzyskać dobre wyniki bez dużego kosztu obliczeniowego. Dzieje się tak zwłaszcza wtedy, kiedy wyjściowe rozwiązanie jest już rozwiązaniem stosunkowo dobrym.



Rys. 5.2. Reguły przed agregacją (strona lewa) i po agregacji (strona prawa)  
Fig. 5.2. Rules before (left-hand side) and after (right-hand side) aggregation



Rys. 5.3. Wizualizacja procesu dostrajania zagregowanej reguły  
Fig. 5.3. Visualization of the process of the aggregated rule tuning

#### 5.1.2.3. Przycinanie po agregacji

Końcowe przycinanie reguł odbywa się w dwóch etapach. W pierwszym usuwane są wszystkie domknięcia warunków elementarnych dodane na etapie domykania reguł. Sposób realizacji drugiego etapu uzależniony jest od tego, czy agregowane są reguły tworzące listę. Jeśli tak jest, z przesłanek usuwane są te skośne warunki elementarne, które nie wpływają na wyniki klasyfikacji. Podczas ich usuwania stosowana jest strategia wspinaczki. Warunki usuwane są dopóty, dopóki dokładność klasyfikacji obliczana na zbiorze walidacyjnym  $V$  nie spadnie poniżej zadanego przez użytkownika poziomu (linia 10, parametr `class-drop`). Wartość parametru `class-drop` podawana jest procentowo względem dokładności listy wejściowej, zawierającej niezagregowane reguły. Gdy `class-drop=0`, oznacza to, że

użytkownik nie godzi się na spadek dokładności klasyfikacji. Ze względu na to, iż zbiór reguł jest uporządkowany, algorytm sprawdza, czy na etapie przycinania z jakiejś reguły nie usunięto przypadkiem wszystkich przesłanek. Jeśli taka sytuacja zaistnieje, wszystkie reguły znajdujące się na liście za taką regułą są usuwane.

Gdy agregacji podlega nieuporządkowany zbiór reguł, drugi etap przycinania polega na usuwaniu z przesłanek reguł zbędnych warunków skośnych. Etap ten realizowany jest za pomocą strategii wspinaczki i jest odpowiednikiem procedury skracania reguł. Podczas przycinania jakość reguł weryfikowana jest za pomocą miary jakości, której wartość obliczana jest na podstawie treningowego zbioru przykładów U (linia 10, parametr  $q\text{-drop}$ ). Dla nieuporządkowanego zbioru reguł dokładność klasyfikacji nie jest zatem bezpośrednim celem optymalizacji.

**Przykład 5.2.** Poniżej zaprezentowano zagregowaną regułę przed eliminacją zbędnych warunków elementarnych i po niej:

- reguła przed skracaniem:

$$\begin{aligned} & (-a_1 + 0.602 \geq 0) \wedge (-a_1 + 0.432a_2 + 0.196 \geq 0) \wedge (-a_1 + 0.63a_2 + 0.022 \geq 0) \wedge \\ & (-0.975a_1 + a_2 - 0.277 \geq 0) \wedge (-0.491a_1 + a_2 - 0.329 \geq 0) \wedge (-0.991a_2 + 1 \geq 0) \wedge \\ & (a_2 - 0.348 \geq 0) \wedge (a_1 + 0.01 \geq 0) \rightarrow \text{class}=+ \end{aligned}$$

- reguła po skróceniu:

$$(-0.975a_1 + a_2 - 0.277 \geq 0) \rightarrow \text{class}=+$$

### 5.1.3. Badanie efektywności algorytmów agregacji

Badanie efektywności proponowanych algorytmów agregacji podzielono na dwie części. Najpierw zbadano działanie algorytmów na zbiorach danych, pochodzących z bazy UCI, a w drugiej części przeanalizowano działanie algorytmu CHIRA na zbiorach reguł wyznaczonych na podstawie syntetycznie wygenerowanych zbiorów przykładów.

Analizie poddano następujące zbiorów danych: *australian credit*, *balance scale*, *breast-wisc.*, *bupa*, *ecoli*, *glass*, *heart clev.*, *heart-statlog*, *ionosphere*, *iris*, *parkinsons*, *pendigits*, *pima-diabetes*, *Ripley's*, *sonar*, *yeast*, *vehicle*, *wine*, *yeast*. Charakterystyka tych zbiorów, poza *parkinsons* i *pendigits*, prezentowana była w rozdziale 4. Charakterystyka zbioru *parkinsons* jest następująca: 195 przykładów, 2 klasy decyzyjne, 22 atrybuty, a charakterystyka zbioru *pendigits* to: 10992 przykładów, 10 klas decyzyjnych, 16 atrybutów.

Wyniki prezentowane w tabelach 5.1–5.5 uzyskano, stosując metodę 10-krotnej warstwowej walidacji krzyżowej (wyjątek stanowił zbiór *pendigits*, do którego zastosowano metodę *train and test*). Porównania pomiędzy różnymi konfiguracjami algorytmów wykonano na podstawie tych samych zbiorów treningowych i testowych. Do eksperymentów wykorzystano implementację algorytmu RIPPER, zawartą w programie Weka. W regułach

wyznaczanych za pomocą algorytmu q-ModLEM do ustalenia zakresów warunków elementarnych wykorzystano entropię warunkową, a do oceny reguły w fazie przycinania i podczas klasyfikacji wykorzystano miarę  $g$  ( $g = 2$ ). Jak wynika z badań opisanych w rozdziale 4, taki sposób indukcji zapewnia wyznaczenie zbiorów reguł o umiarkowanej liczebności i wysokiej dokładności klasyfikacji.

Eksperymenty przeprowadzono dla różnych konfiguracji parametrów. Efektywność algorytmów badano ze względu na: całkowitą dokładność klasyfikacji (Acc), średnią dokładność klas decyzyjnych (BAcc), procentowy wzrost/spadek średniej dokładności klas decyzyjnych przed agregacją i po niej, procentowy spadek liczby reguł po agregacji, procentowy spadek liczby atrybutów tworzących warunki elementarne reguł. Liczbę atrybutów tworzących przesłanki reguł obliczano tak, jak definiuje to miara (3.26). Badania wykonano dla dwóch wartości parametru  $\text{max-dim}$ , wynoszących 3 i 5. W tabelach 5.1 i 5.2 zawarto szczegółowe wyniki agregacji reguł utworzonych przez algorytm RIPPER. W tabelach 5.3, 5.4, 5.5 zaprezentowano rezultaty agregacji reguł otrzymanych przez algorytm q-ModLEM. Dla reguł tych podano także rezultaty agregacji, będące wynikiem stosowania algorytmu łączącego zakresy warunków elementarnych. W tabelach 5.3, 5.4, 5.5 algorytm ten nazwano algorytmem sklejania. Wektory liczb o postaci  $(x \ y \ z)$ , zamieszczone w tabelach, informują o parametrach algorytmu i mają następujące znaczenie:

- x – dopuszczalny procentowy spadek jakości zagregowanej reguły w stosunku do lepszej (mającej większą jakość) z agregowanych reguł,
- y – dopuszczalny procentowy spadek dokładności klasyfikacji, wyznaczany na podstawie zbioru treningowego (dotyczy fazy przycinania reguł i algorytmu RIPPER),
- z – dopuszczalny procentowy spadek jakości przycinanej reguły (dotyczy fazy przycinania i algorytmu q-ModLEM).

Dokładność klasyfikacji i średnia dokładność klas decyzyjnych podawane są w procentach.

Tabela 5.1  
Wyniki agregacji reguł wyznaczonych przez algorytm  
RIPPER (5 3 -), max-dim=5

Zbiór danych	Dokładność klasyfikacji			Informacje na temat reguł		
	Reguły wejściowe Acc	Reguły po agregacji Acc	Wzrost/ spadek BAcc	Liczba reguł	Redukcja liczby reguł [%]	Redukcja liczby atrybutów [%]
Australian	85.6±4.2	85.2±4.4	-0.4	4.7±1.7	2.1	11.2
Balance-scale	81.1±3	85.4±2.6	5.4	11.1±2.1	68.5	41.5
Breast-wisc	94.9±2	95.4±2.2	1.8	6.1±0.7	52.5	28.2
Bupa	67±6.5	68.5±7.5	3.3	3.8±0.9	0.0	5.4
Ecoli	84±3.7	84±3.4	0.0	8.7±1.3	1.1	10.7
Glass	64.8±8.7	66.1±9.3	1.7	7.8±1.0	2.6	7.1
Heart-C.	79.2±4.6	78.9±4.7	-0.2	4.9±1.1	0	9.8
Heart-Statlog	81.1±5.4	79.6±6.1	-1.8	4.1±0.9	29.3	-11.8
Ionosphere	90.0±5.9	90.0±6	0.0	4.9±1.4	6.1	8.8
Iris	95.3±5.5	95.3±5.5	0.0	3.6±0.5	0.0	0.0
Parkinsons	88.2±7.9	88.2±7.9	0.0	4.2±1.1	0.0	1.5
Pendigits	92	90.5	-1.6	66	10.6	11
Pima	74.7±3.6	73.2±5	-0.5	3.9±1	10.3	16.2
Ripley`s	89.5±4.1	89.1±3.9	-0.4	6±1.7	35.0	48
Sonar	75.4±7	75.4±7	0.0	4.5±1.4	0	1.5
Vehicle	66.8±3	67.6±2.8	1.3	14.2±2.1	4.2	22.7
Wine	89.9±6.8	89.9±6.8	0.0	3.9±0.6	0.0	0.0
Yeast	58.2±3.1	57.8±3.4	5.2	16.2±1.8	7.4	25.2
<b>Średnia</b>	<b>81.0</b>	<b>81.2</b>	<b>0.8</b>	<b>9.9</b>	<b>12.8</b>	<b>13.2</b>

Tabela 5.2  
Rezultaty agregacji reguł wyznaczonych  
przez algorytm RIPPER

Algorytm	Acc	BAcc	Średnia liczba reguł	Średnia liczba atrybutów
RIPPER	81.0	78.0	9.9	25.3
CHIRA; (5 2 -) dim=3	81.0	78.4	9.3	22.5
CHIRA; (5 2 -) dim = 5	81.2	78.8	8.6	22.0
CHIRA; (10 8 -); dim=3	80.7	78.0	8.9	21.2
CHIRA; (10 8 -); dim=5	80.4	77.7	8.1	19.7

Tabela 5.3

Wyniki agregacji reguł wyznaczonych przez algorytm  
q-ModLEM (0 - 0), max-dim=3

Zbiór danych	Reguły wejściowe Acc	Reguły po agregacji Acc	Wzrost/ spadek BAcc	Reguły po sklejaniu Acc
Australian	85.5±3.5	85.5±3.4	0.0	85.2±2.0
Balance-scale	84.8±3.9	85.7±4.0	-2.7	85.7±3.3
Breast-wisc	95.5±2.5	95.5±2.3	-0.8	95.7±2.6
Bupa	69.0±7.5	69.8±5.8	1.2	69.5±7.4
Ecoli	82.2±7.2	82.3±6.5	2.0	82.1±7.3
Glass	66.9±10.4	65.9±10.4	-0.3	67.9±10.4
Heart	80.2±7.5	81.9±6.2	1.9	80.0±7.5
Heart-statlog	80.6±7.2	81.6±7.1	1.5	80.6±7.3
Ionosphere	91.1±4.6	92.6±3.6	1.5	91.2±4.6
Iris	93.3±3.1	94.2±3.2	0.7	93.3±3.1
Parkinsons	88.0±11.6	87.4±11.0	-1.4	88.0±11.6
Pendigits	91.4	91.6	0.0	91.6
Pima	74.9±4.2	74.7±3.5	-0.8	74.9±4.2
Ripley's	88.8±2.6	89.0±2.7	0.1	88.8±2.6
Sonar	75.4±8.4	76.5±7.8	2.1	75.4±8.3
Vehicle	72.1±5.5	73.2±6.2	1.6	72.5±5.4
Wine	94.5±6.7	94.5±5.3	0.0	94.5±6.7
Yeast	57.4±4.0	57.5±3.9	0.9	57.5±4.1
<b>Średnia</b>	<b>81.8</b>	<b>82.2</b>	<b>0.4</b>	<b>81.9</b>

Tabela 5.4  
Redukcja liczby reguł wyznaczonych przez algorytm  
q-ModLEM (0 - 0), max-dim=3

Zbiór danych	Liczba reguł	Redukcja liczby reguł [%]	Redukcja liczby symboli [%]	Redukcja liczby reguł po sklejaniu [%]
Australian	20.9±1.4	7.1	6.3	3.9
Balance-scale	104.4±5.1	22.9	13.5	19.5
Breast-wisc	23.7±2.0	28.4	21.8	12.7
Bupa	59.7±4.9	12.7	11.1	5.5
Ecoli	45.0±8.3	26.2	19.3	14.7
Glass	29.1±2.3	8.3	3.9	7
Heart	17.3±2.5	11.3	8.5	9.6
Heart-statlog	16.5±3.6	25.5	17.7	6.2
Ionosphere	16.8±1.4	14.5	7.9	2.8
Iris	7.9±1.4	30.5	20.2	5.5
Parkinsons	14.4±1.9	3.5	3.6	0
Pendigits	224	15.4	7.0	2.8
Pima	45.1±4.9	12.6	6.9	3
Ripley's	29.6±3.4	42.2	33.6	30.5
Sonar	26.9±3.2	6.9	5.5	4.6
Vehicle	92.7±8.8	7.7	5.1	3.1
Wine	9.9±0.8	2.1	1.8	2.1
Yeast	202.1±16.2	8.4	7.7	5.5
<b>Średnia</b>	<b>54.8</b>	<b>15.9</b>	<b>11.2</b>	<b>7.7</b>

Tabela 5.5  
Rezultaty agregacji reguł wyznaczonych przez algorytm  
q-ModLEM

Algorytm	Acc	BAcc	Średnia liczba reguł	Średnia liczba symboli
q-ModLEM	81.8	77.8	54.8	123.4
CHIRA; (0 - 0) max-dim=3	82.2	78.2	46.1	109.6
CHIRA; (0 - 0) max-dim = 5	82.0	77.8	46.8	111.2
CHIRA; (5 - 0); max-dim=3	82.1	78.1	44.2	106.5
CHIRA; (5 - 0); max-dim=5	81.9	77.7	43.8	108.9

Analizując wyniki, można zauważyc prawidłowość polegającą na tym, że redukcja reguł jest we wszystkich przypadkach większa niż redukcja atrybutów tworzących warunki elementarne. Agregacja reguł wyznaczonych przez algorytm RIPPER wymaga nieznacznego obniżenia dokładności klasyfikacji podczas przycinania (parametry 5 2 -). Większe obniżenie dokładności klasyfikacji (parametry 10 8 -) prowadzi już do znacznego spadku dokładności klasyfikacji. Lepsze wyniki kompresji otrzymano dla 5-wymiarowych warunków złożonych. Agregacja reguł wyznaczonych przez algorytm RIPPER jest niezwykle trudna, gdyż algorytm ten generuje niewiele reguł. Zauważmy, że w przypadku zbiorów danych,

w których rzeczywiście można spodziewać się skośnych warunków elementarnych (np. *Balance-scale*, *Ripley's*) redukcja reguł jest znaczna. Dużą redukcję zanotowano również dla zbiorów *Breast-wisc* oraz *Heart-statlog*, chociaż w tym przypadku zwiększała się liczba atrybutów pojawiających się w przesłankach reguł. W pracy [265] przedstawiono wyniki zastosowania miary  $m$  w algorytmie CHIRA. Zgodnie z oczekiwaniami pozwoliło to na większą redukcję liczby reguł i atrybutów, jednakże odbyło się to kosztem nieznacznego obniżenia dokładności klasyfikacji.

Algorytm q-ModLEM wyznacza znacznie więcej reguł niż RIPPER. Nie biorąc pod uwagę reguł wyznaczanych dla najliczniejszej klasy decyzyjnej, można powiedzieć, że q-ModLEM generuje dwukrotnie więcej reguł niż RIPPER. Większa liczba reguł wejściowych prowadzi do większej redukcji po agregacji. Największe redukcje otrzymano dla zbiorów *Balance-scale*, *Ripley's* oraz *Breast-wisc* i *Heart-statlog*. Dla parametru `max-dim=3` oraz wektorów (0 - 0) i (5 - 0) nieznacznie podniosła się także dokładność klasyfikacji. Redukcja liczby reguł wynosiła od 16% do 20%. Zwiększenie wartości `max-dim` do 5 powoduje redukcję od 15% do 25%.

Algorytm CHIRA pozwala na większe ograniczenie liczby reguł, niż robi to algorytm sklejania. W przeprowadzonych eksperymentach CHIRA redukuje ponaddwukrotnie więcej reguł niż algorytm sklejania. Należy jednak pamiętać, że reguły zawierające skośne warunki elementarne są trudniejsze do interpretacji niż reguły zawierające jedynie warunki proste.

Wałąną cechą algorytmu agregacji jest to, że nie zmienia on (a nawet nieznacznie polepsza) zdolności klasyfikacyjnych zredukowanego zbioru reguł. Dotyczy to zarówno całkowitej dokładności klasyfikacji, jak i średniej dokładności klas decyzyjnych.

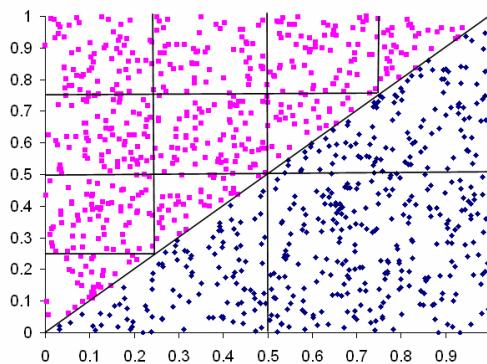
Ze względu na różny sposób dostrajania zagregowanych reguł stopnie kompresji, uzyskane dla algorytmów RIPPER i q-ModLEM, nie mogą być porównywane.

W algorytmie CHIRA najbardziej czasochłonną operacją jest wyznaczenie otoczki wypukłej, zawierającej agregowane reguły. Dla większości z analizowanych zbiorów reguł czas agregacji w trybie 10-krotnej walidacji krzyżowej nie był dłuższy niż 1 minuta. Agregacja zbiorów złożonych z większej liczby reguł może być jednak czasochłonna. Przykładowo czas agregacji reguł opisujących zbiory danych *Balance-scale*, *Pendigits* i *Yeast* wynosił odpowiednio 10, 18 i 37 minut (eksperyment wykonano na komputerze wyposażonym w procesor Intel Core Duo 3.5 GHz). Ze względu na czas obliczeń przed agregacją konieczna może być wstępna eliminacja reguł za pomocą jednego z algorytmów filtracji, prezentowanych w podrozdziale 5.3.

W dalszej części badań, aby sprawdzić, czy algorytm zachowuje się zgodnie z oczekiwaniemi, wygenerowano trzy zbiory danych syntetycznych. Pierwszy zbiór (d2d) składał się z dwóch numerycznych atrybutów warunkowych i dwóch klas decyzyjnych, które można oddzielić za pomocą dwóch prostych równaniach  $a_1 - a_2 + 0.33 = 0$ ,

$a_1 - a_2 - 0.33 = 0$ . Rozkład przykładów pomiędzy klasami był równomierny, wygenerowano 1000 przykładów, tak aby równomiernie pokryć obszar  $[0,1] \times [0,1]$ . Drugi zbiór danych (2d) ma takie same parametry jak zbiór d2d, jedyną różnicą jest to, że klasy decyzyjne rozgranicza jedna prosta o równaniu  $a_1 - a_2 = 0$  (rys. 4.4). Trzeci zbiór (3d) jest zbiorem złożonym z trzech atrybutów numerycznych, w których dwie klasy decyzyjne rozdzielono płaszczyzną o równaniu  $a_1 + a_2 - a_3 = 0$ ; rozkład przykładów pomiędzy klasami decyzyjnymi wynosił 165 (przykłady pozytywne) i 835 (przykłady negatywne).

W pierwszej części eksperymentu w sposób syntetyczny wygenerowano reguły, które dla zbiorów d2d oraz 2d były prostokątami idealnie pokrywającymi przykłady pozytywne i negatywne, a w przypadku zbioru 3d były sześciokątnymi. Na rysunku 5.4 zaprezentowano ideę generowania syntetycznych reguł dla zbioru 2d.



Rys. 5.4. Zasada tworzenia reguł w zbiorze danych syntetycznych – zbiór 2d  
Fig. 5.4. Rule defining in the synthetic data – 2d data set

Pierwsza reguła ma postać  $(a_1 \leq 0.5) \wedge (a_2 \geq 0.5) \rightarrow \text{class}=+$ , a druga –  $(a_1 \geq 0.5) \wedge (a_2 \leq 0.5) \rightarrow \text{class}=-$ . Przesłanki dwóch kolejnych reguł dla klasy + mają postaci:  $(a_1 \leq 0.25) \wedge (a_2 \geq 0.25)$  oraz  $(a_1 \leq 0.75) \wedge (a_2 \geq 0.75)$ . Reguły tworzy się zgodnie z przedstawioną tutaj zasadą dopóty, dopóki nie zostanie pokryty cały zbiór przykładów.

Algorytm CHIRA, stosujący jedynie kryterium jakości agregowanej reguły i niedopuszczający do spadku jakości reguł, dla wszystkich trzech zbiorów pozwolił na uzyskanie pożądanych wyników. Dokładniej, w utworzonych skośnych warunkach elementarnych znalazły się równania, według których rozdzielane są klasy decyzyjne. W rezultacie dla zbioru 2d otrzymano dwie reguły, dla zbioru d2d – trzy reguły, dla zbioru 3d otrzymano również dwie reguły. Wyniki te potwierdzają, że w warunkach sztucznych algorytm działa zgodnie z założeniami.

Druga część eksperymentu polegała na wygenerowaniu przez algorytmy RIPPER i q-ModLEM reguł opisujących zbiory syntetyczne i poddaniu tych reguł agregacji, przy czym dokładność klasyfikacji wyznaczonych reguł weryfikowano na podstawie całego zbioru przykładów (nie wydzielano części testowej). Oba algorytmy wygenerowały reguły różniące się od tych utworzonych w sposób sztuczny. W szczególności część z wyznaczonych reguł

była niedokładna. Taki wynik nie dziwi, gdyż zarówno RIPPER, jak i q-ModLEM dopuszczają do indukcji reguł niedokładnych, jeśli tylko pokrywają one wystarczająco dużo przykładów pozytywnych. Szczegółowe wyniki eksperymentów zamieszczone w tabeli 5.6.

Algorytm q-ModLEM nie poradził sobie dobrze z wyznaczeniem reguł dla zbioru d2d. Dla zbiorów d2d i 3d wyznaczono dużo reguł, a agregacja pozwoliła na znaczne ograniczenie ich liczby. W wynikowym zbiorze nie znaleziono jednak reguł, których skośne warunki elementarne zawierały równania, według których rozdzielane są klasy decyzyjne. Takie reguły udało się wyznaczyć jedynie dla zbioru 2d. Algorytm RIPPER uzyskał znacznie lepsze wyniki. Postać skośnych warunków elementarnych jest podobna do równań rozgraniczających klasy decyzyjne zbiorów 2d i d2d.

Tabela 5.6  
Wynik agregacji reguł wyznaczonych na podstawie syntetycznych zbiorów danych

Zbiór danych	Algorytm	Przed agregacją Acc	Po agregacji Acc	Liczba reguł	
				Przed agregacją	Po agregacji
d2d	q-ModLEM	73.0	76.5	28	14
	RIPPER	99.6	95.6	15	3
2d	q-ModLEM	99.2	99.8	128	4
	RIPPER	98.7	97.6	10	2
3d	q-ModLEM	100.0	99.9	58	11
	RIPPER	98.6	95.6	9	3

Po agregacji każdy ze zbiorów syntetycznych udało się opisać znaczaco mniejszą liczbą reguł; stało się to kosztem niewielkiej utraty dokładności klasyfikacji. Z przeprowadzonej analizy można wyciągnąć wniosek, że postać wynikowych reguł w dużej mierze uwarunkowana jest postacią reguł wejściowych.

W świetle przeprowadzonych badań można stwierdzić, że algorytmy agregacji pozwalają na ograniczenie liczby reguł, nie powodując jednocześnie spadku ich zdolności klasyfikacyjnych. Większą redukcję liczby reguł otrzymano, stosując do agregacji algorytm CHIRA. W zależności od metody indukcji redukcja liczby reguł wynosiła od 6 do 21% (algorytm RIPPER) oraz od 15 do 25% (algorytm q-ModLEM). Odnutowana redukcja atrybutów była mniejsza, gdyż w zagregowanych regułach znajdowały się skośne warunki elementarne, zbudowane z kilku atrybutów. Algorytm agregacji przez łączenie zakresów warunków elementarnych pozwala na ograniczenie liczby reguł o około 8%. W pracy [248] przedstawiono wyniki zastosowania tego algorytmu do zbiorów reguł wyznaczanych na podstawie reduktów względnych. Dla tak wyznaczonych reguł redukcja wynosiła 25%.

Analiza wyników otrzymanych przez algorytm CHIRA dostarcza jeszcze jednego ważnego spostrzeżenia. Jeśli w zbiorze danych występują zależności, które można opisać za pomocą skośnych warunków elementarnych, to przy odpowiednim zbiorze reguł wejściowych algorytm CHIRA jest w stanie takie warunki znaleźć.

Można próbować poprawić wyniki algorytmu CHIRA przez wyposażenie go w procedurę umożliwiającą obracanie hiperplaszczyzn w fazie dostrajania reguł. Procedura

taka mogłaby być realizowana w sposób podobny jak proponuje Murthy [205], w algorytmie indukcji skośnych drzew decyzyjnych.

## 5.2. Redefinicja reguł na podstawie oceny ważności warunków elementarnych

W dotychczas opisanych metodach i algorytmach dokonywano oceny jakości reguł. Obecnie omówione zostanie zagadnienie oceny warunków elementarnych, z których zbudowane są reguły.

Publikacji dotyczących oceny ważności warunków elementarnych w kontekście wyznaczonego zbioru reguł jest niewiele. W zasadzie jedynie prace Greco, Stefanowskiego i Słowińskiego [105, 110] poruszają ten problem. Rozważają oni ważność warunków elementarnych z punktu widzenia dokładności zawierających je reguł. W analizie pod uwagę brana jest dokładność reguł zawierających oceniane warunki i reguł ich niezawierających. Ważność nie jest oceniana z punktu widzenia statystycznej istotności, dlatego też w przywoływanych pracach nie stosuje się terminu *istotność*, a właśnie terminu *ważność*. W definicji wskaźników przeznaczonych do oceny ważności warunków elementarnych kluczową rolę odgrywają indeksy Shapleya [243] i Banzhafa [12], które w teorii gier stosowane są do oceny siły graczy i koalicji. Autorzy prac [105, 110] wskazują na dużą złożoność obliczeniową, związaną z wyznaczaniem wartości indeksów Shapleya i Banzhafa.

Indeksy te można zastosować do utworzenia wskaźników oceniających ważność zbiorów warunków elementarnych. Dodatnie wartości tych wskaźników oznaczać będą, że oceniane warunki wpływają na zwiększenie się dokładności reguł, natomiast wartości ujemne – że oceniane warunki są (przynajmniej częściowo) redundantne. W dalszej części rozdziału wykorzystane zostaną jedynie wskaźniki umożliwiające ocenę ważności pojedynczych warunków elementarnych.

### 5.2.1. Ocena ważności warunków elementarnych

Załóżmy, że dane są: reguła decyzyjna  $r$  oraz zbiór warunków elementarnych  $Ec(r) = \{w_1, w_2, \dots, w_n\}$ . W standardowej postaci wartości indeksów Shapleya (5.1) i Banzhafa (5.2) obliczane są na podstawie informacji o średnim wpływie ocenianego warunku elementarnego na dokładność reguł będących wszystkimi możliwymi uogólnieniami reguły  $r$ :

$$\phi_S(w, r) = \sum_{Y \subseteq Ec(r) - \{w\}} \frac{(n - |Y| - 1)! |Y|!}{n!} [precision(Y \cup \{w\}, r) - precision(Y, r)] \quad (5.1)$$

$$\phi_B(w, r) = \frac{1}{2^{n-1}} \sum_{Y \subseteq Ec(r) - \{w\}} [precision(Y \cup \{w\}, r) - precision(Y, r)]. \quad (5.2)$$

We wzorach (5.1) i (5.2)  $precision(Y, r)$  oznacza dokładność reguły  $r$ , której przesłanka zbudowana jest jedynie z warunków zawartych w zbiorze  $Y$ . Dodatkowo przyjmujemy następujące założenia:  $precision(\emptyset, r) = 0$ ,  $precision(Ec(r), r) = precision(r)$ .

Przez  $RUL_X$  oznaczmy zbiór wszystkich reguł wskazujących na klasę decyzyjną  $X$ . Ważność warunku elementarnego w zbiorze  $RUL_X$  obliczana jest na podstawie ważności tego warunku we wszystkich regułach należących do  $RUL_X$ , a także na podstawie ważności tego warunku we wszystkich regułach wskazujących na klasę decyzyjną inną niż  $X$ . Ocena ważności warunku  $w$  dla klasy decyzyjnej  $X$  wyraża się wzorem (5.3):

$$G(w, RUL_X) = \sum_{r \in RUL_X} (\phi_{\_}(w, r) \cdot coverage(r)) - \sum_{r \notin RUL_X} (\phi_{\_}(w, r) \cdot coverage(r)). \quad (5.3)$$

We wzorze (5.3)  $\phi_{\_}$  jest jednym z indeksów (5.1) lub (5.2).

Zgodnie z (5.3) w globalnej ocenie warunku elementarnego pod uwagę brany jest jego wkład do dokładności każdej z zawierających go reguł oraz wszystkich jej uogólnień, a także brane jest pod uwagę pokrycie reguł zawierających oceniany warunek.

Pokryciowe algorytmy indukcji reguł dokonują bieżącej oceny warunków elementarnych w fazach wzrostu i przycinania. Ocena ta dokonywana jest zazwyczaj za pomocą miary jakości. Naturalnym uogólnieniem wzorów (5.1)–(5.3) jest zatem zastąpienie miary  $precision$  miarą jakości, definiowaną na podstawie tablicy kontyngencji. Taka zamiana spowoduje, że w każdym uogólnieniu  $r$  badany będzie wpływ danego warunku nie tylko na dokładność reguły, ale również na jej pokrycie. Wzory (5.1) i (5.2) przyjmą wtedy odpowiednio postaci (5.4) i (5.5), a ocena warunku elementarnego w zbiorze reguł będzie wyrażać się wzorem (5.6):

$$\phi_{qS}(w, r) = \sum_{Y \subseteq Ec(r) - \{w\}} \frac{(n - |Y| - 1)! |Y|!}{n!} [q(Y \cup \{w\}, r) - q(Y, r)], \quad (5.4)$$

$$\phi_{qB}(w, r) = \frac{1}{2^{n-1}} \sum_{Y \subseteq Ec(r) - \{w\}} [q(Y \cup \{w\}, r) - q(Y, r)], \quad (5.5)$$

$$G_q(w, RUL_X) = \sum_{r \in RUL_X} \phi_{q\_}(w, r) - \sum_{r \notin RUL_X} \phi_{q\_}(w, r). \quad (5.6)$$

We wzorach (5.4) i (5.5) miarę  $precision$  zastąpiono miarą jakości  $q$ . Miara ta powinna brać pod uwagę zarówno dokładność, jak i pokrycie ocenianych reguł. We wzorze (5.6) usunięto ocenę pokrycia reguł zawierających oceniany warunek elementarny, gdyż jest ona wykonywana już w trakcie wyznaczania wartości  $\phi_{q\_}(w, r)$ .

Wyznaczając wartości indeksów  $\phi_{\_}$  lub  $\phi_{q\_}$  dla reguły  $r$ , złożonej z  $n$  warunków elementarnych, na całym zbiorze przykładów musimy  $2^{n-1}$  razy wyznaczyć wartość miary  $precision$  lub innej miary jakości  $q$ , tyle bowiem otrzymamy uogólnień reguły

r. Gdy liczba warunków elementarnych jest duża lub duży jest zbiór przykładów, na podstawie którego obliczana jest jakość reguły, operacja taka będzie czasochłonna. Aby skrócić czas obliczeń, związany z wyznaczaniem wartości indeksów  $\phi_{q\_}$  i  $\phi\_$ , można przedstawić ich uproszczone formy, składające się jedynie z tych składników sum występujących we wzorach (5.1), (5.2), (5.4), (5.5), które wnoszą najwięcej do oceny ważność warunku elementarnego. W przedstawionych dalej uproszczonych postaciach indeksów Shapleya i Banzhafa przyjęto, że najwięcej informacji o ważności warunku elementarnego wnoszą:

- reguły bazowe, a więc te, które zawierają oceniany warunek elementarny,
- reguły bazowe, z których usunięto jedynie oceniany warunek elementarny,
- reguły, których przesłanki zawierają jedynie oceniany warunek elementarny.

Uproszczone postaci indeksów Shapleya i Banzhafa, przeznaczone do oceny ważności warunków elementarnych, będą wyrażały się wzorami (5.7)–(5.10). Wyrażenie (5.7) jest uproszczoną postacią wyrażenia (5.1), a (5.8) jest uproszczoną postacią (5.4). Podobnie (5.9) przedstawia uproszczoną postać (5.2), wreszcie (5.10) – uproszczoną postacią indeksu (5.5):

$$\phi_{ss}(w, r) = \frac{1}{n} (precision(r) - precision(Ec(r) - \{w\}, r) + precision(\{w\}, r)), \quad (5.7)$$

$$\phi_{sqS}(w, r) = \frac{1}{n} (q(r) - q(Ec(r) - \{w\}, r) + q(\{w\}, r)), \quad (5.8)$$

$$\phi_{sB}(w, r) = \frac{1}{2^{n-1}} (precision(r) - precision(Ec(r) - \{w\}, r) + precision(\{w\}, r)), \quad (5.9)$$

$$\phi_{sqB}(w, r) = \frac{1}{2^{n-1}} (q(r) - q(Ec(r) - \{w\}, r) + q(\{w\}, r)). \quad (5.10)$$

W dalszej części rozdziału indeksy wyrażone wzorami (5.1) i (5.2) będziemy nazywać indeksami podstawowymi, indeksy (5.4) i (5.5) – indeksami zmodyfikowanymi, wreszcie indeksy (5.7)–(5.10) – indeksami uproszczonymi. W zależności od zastosowanej miary jakości będą to indeksy podstawowe uproszczone (5.7), (5.9) lub indeksy zmodyfikowane uproszczone (5.8, 5.10).

Warto zauważyć, że w przypadku indeksów Shapleya podstawowego i zmodyfikowanego największa waga  $1/n$  przyporządkowywana jest właśnie tym składnikom, które występują we wzorach (5.7), (5.8), definiujących uproszczone postaci tego indeksu. W indeksie Banzhafa wagi przyporządkowane każdemu składnikowi sum (5.2), (5.5) są identyczne. Indeks Banzhafa można traktować jako średnią arytmetyczną, a indeks Shapleya – jako średnią ważoną. We wzorach (5.7), (5.8), a także (5.9), (5.10), wagi  $1/n$  i  $1/2^{n-1}$  mogą zostać pominięte lub zastąpione przez  $1/2$ . Zmiany te skutkować będą jedynie zmianami wartości ocen generowanych dla poszczególnych warunków elementarnych. Nie wpłyną one

ani na porządek tych warunków, ani na proporcje pokazujące, o ile jeden warunek jest lepszy (gorszy) od drugiego. Dlatego jeśli zależy nam jedynie na uporządkowaniu warunków elementarnych, obecna postać indeksów uproszczonych może pozostać taka jak we wzorach (5.7)–(5.10). Przedstawione rozumowanie prowadzi również do wniosku, że porządkи warunków, ustanowione przez uproszczone indeksy Shapleya i Banzhafa, są identyczne. Dowód tej własności jest prosty, wystarczy wyrażenia (5.7), (5.8) pomnożyć przez  $n$ , a wyrażenia (5.9), (5.10) – przez  $2^{n-1}$ . Wówczas otrzymamy wyrażenia identyczne. Korzystamy tutaj z takiej własności, że mnożenie każdego wyrazu ciągu przez liczbę dodatnią nie zmienia kolejności wyrazów w tym ciągu.

Nie można udowodnić tego, że porządkи warunków elementarnych, utworzone za pomocą indeksów (5.1), (5.7); (5.4), (5.8); (5.2), (5.8); (5.5), (5.10), będą identyczne. Można jednak empirycznie sprawdzić, czy i w jakim stopniu porządkи te są skorelowane. Eksperymenty takie opisano w dalszej części rozdziału.

Warto również wspomnieć o dwóch własnościach omawianych tutaj indeksów. Po pierwsze, wartości indeksów podstawowych (5.1), (5.2) oraz indeksów podstawowych uproszczonych (5.7), (5.9) są liczbami nieujemnymi. Dowód tej własności wynika z faktu, że po dodaniu nowego warunku elementarnego do przesłanki reguły jej dokładność rośnie albo pozostaje niezmieniona. Po drugie, wartości indeksów zmodyfikowanych (5.4), (5.5) oraz zmodyfikowanych uproszczonych (5.8), (5.10) mogą być liczbami ujemnymi. Jeżeli do wyznaczenia wartości indeksów zmodyfikowanych użyto miary oceniającej równocześnie dokładność i pokrycie reguł, wówczas po usunięciu warunku elementarnego reguła może być oceniana wyżej niż reguła zawierająca oceniany warunek.

Na zakończenie należy zaznaczyć, że pojęcie uproszczonych indeksów Shapleya i Banzhafa zdefiniowano jedynie dla celów oceny ważności warunków elementarnych w regułach. Analiza zachowania i celowość wprowadzania takich indeksów dla teorii gier nie była zamiarem autora.

### 5.2.2. Algorytm redefinicji

Obliczając wartości wskaźników prezentowanych w poprzednim podrozdziale, dla każdej klasy decyzyjnej  $X$  i każdego zbioru  $RUL_X$  możemy utworzyć ranking warunków elementarnych. Jeżeli  $RUL_X$  jest opisem klasy decyzyjnej, wówczas ranking ten odzwierciedla wpływ warunków elementarnych na jakość opisu klasy decyzyjnej  $X$ . Warunki znajdujące się na początku rankingu mają większe znaczenie dla opisu klasy  $X$  niż warunki znajdujące się na jego końcu.

Dalej zaprezentowano algorytm, który na podstawie rankingu warunków elementarnych dokonuje indukcji reguł. Algorytm działa na zasadzie generowania pokrycia zbioru treningowego. Podczas indukcji używane są jedynie te warunki elementarne, które zawarte są

w regułach tzw. bazowego zbioru reguł. Prezentowany algorytm optymalizuje bazowy zbiór reguł, tworząc je od nowa (stąd słowo *redefinicja* w tytule algorytmu). Warunki elementarne, które stanowią podstawowy budulec reguły, nie są w żaden sposób modyfikowane. Kolejność dodawania warunków do przesłanek reguły jest zgodna z kolejnością warunków w rankingu ważności. W algorytmie dodano zabezpieczenia przed umieszczeniem w przesłance reguły warunków redundantnych. Po pierwsze, po dodaniu kolejnego warunku do przesłanki sprawdzana jest jakość nowej, rozszerzonej reguły. Jeśli jakość reguły rośnie, warunek pozostaje w przesłance. Po drugie, po zakończeniu tworzenia kolejnej reguły, a przed indukcją następnej ranking warunków tworzony jest od nowa. Nowy ranking tworzony jest w zbiorze niepokrytych jeszcze przykładów treningowych.

Prezentowany algorytm umożliwia również umieszczenie w przesłance zanegowanych warunków elementarnych. Warunki te tworzone są na podstawie założenia, że warunki proste, ważne dla opisu jednej klasy decyzyjnej, są nieważne dla opisów innych klas. Innymi słowy, na podstawie warunku  $w$  ważnego dla klasy  $X$  można utworzyć warunek  $\neg w$ , który będzie ważny dla opisu innej klasy decyzyjnej. Argumentacja ta ma szczególne uzasadnienie w przypadku problemów dwuklasowych.

Prezentowany algorytm nie dokonuje zmian w zakresach warunków elementarnych; wynika to z wyników badań przeprowadzonych przez autora. Dopuszczenie do takich zmian prowadziło do pogorszenia jakości klasyfikacji. Nie jest to zresztą wynik zaskakujący, gdyż zmiana zakresu warunku elementarnego powodowała, że uzyskiwał on zupełnie inną ocenę ważności, a tym samym – inną pozycję w rankingu.

Algorytm redefinicji zbioru reguł na podstawie oceny ważności warunków elementarnych

Wejście:  $U$  – zbiór przykładów,

$A \cup \{d\}$  – zbiór atrybutów

$RUL$  – zbiór reguł podzielony na opisy klas decyzyjnych  $RUL_{Xj}$

$q$  – miara oceniająca jakość reguł

$Ec_j$  – zbiory warunków elementarnych, zawartych w regułach  $RUL_{Xj}$

$\langle w_{j1}, w_{j2}, \dots, w_{jm} \rangle$  ranking warunków elementarnych, zawartych w  $Ec_j$

Wyjście:  $RUL_{red}$  – wyjściowy zbiór reguł

```

1. Begin
2.    $RUL_{out} = \emptyset;$ 
3.   Foreach klasa decyzyjna  $X_j, j \in \{1, \dots, k\}$  do
4.     Rozpocznij od rankingu warunków  $\langle w_{j1}, w_{j2}, \dots, w_{jm} \rangle$ 
5.      $G := X_j;$ 
6.     While ( $G \neq \emptyset$  or dowolny warunek z  $Ec_j$  pokrywa jakiś przykład z  $G$ ) do
7.        $Ec := \emptyset;$ 
8.        $pr := \emptyset;$  // rozpoczęt od pustej przesłanki
9.        $q(pr \rightarrow X_j) = -\infty;$  // ustaw minimalną jakość reguły  $pr \rightarrow X_j$ 
          //  $pr \rightarrow X_j$  oznacza regułę wskazującą na klasę  $X_j$ 

        // dodawanie do przesłanki niezanegowanych warunków elementarnych
10.      Foreach  $i=1, \dots, m_j$  do
11.        If  $q(pr \wedge w_{ji} \rightarrow X_j) > q(pr \rightarrow X_j)$  then  $pr := pr \wedge w_{ij};$ 

        // przygotowanie rankingu warunków z innych niż  $X_j$  klas decyzyjnych
12.      Foreach ( $s=1, \dots, k$  and  $s \neq j$ ) do
13.         $Ec := Ec \cup Ec_s$ 
14.        Posortuj warunki elementarne, należące do  $Ec$ , malejąco ze względu
            na ich ważność (niezależnie od tego która klasę opisują) i umieść
            je na liście  $\langle w_1, w_2, \dots, w_{|Ec|} \rangle$ 

        // dodawanie do przesłanki zanegowanych warunków elementarnych
15.      Foreach  $s=1, \dots, |Ec|$  do //
16.        If  $q(pr \wedge \neg w_s \rightarrow X_j) > q(pr \rightarrow X_j)$  then  $pr := pr \wedge \neg w_s;$ 

        // skracanie reguły, przygotowanie nowego rankingu
17.       $sh\_r := Skroc-regule(pr \rightarrow X_j, U, q);$ 
18.       $RUL_{out} = RUL_{out} \cup \{sh\_r\};$ 
19.       $G := G - [sh\_r];$  // [sh_r] ozn. przykłady pozytywne pokrywane przez sh_r
20.      Foreach  $s=1, \dots, k$  do
21.        Przelicz ranking warunków elementarnych  $Ec_s$ , biorąc pod uwagę zbiór
            przykładów  $GU(U-X_j)$ 
22.      end while
23.    end for
24.  end.

```

Procedura  $Skroc-regule(pr \rightarrow X_j, U, q)$  (linia 17) realizuje zadanie skracania reguły w sposób identyczny z opisanym w rozdziale 2.4. Skracanie reguł wyznaczonych na podstawie rankingu ważności warunków elementarnych jest szczególnie uzasadnione, gdyż reguły te są zazwyczaj bardzo długie.

Pozytywne (niezanegowane) warunki elementarne są bardziej pożądane w przesłankach niż warunki zanegowane. Algorytm w pierwszej fazie wzrostu dodaje do przesłanki warunki pozytywne (linie 10, 11), a w drugiej kolejności – warunki zanegowane (linie 15, 16). Zauważmy, że jeśli pozytywne warunki elementarne wystarczają do indukcji reguły dokładnej, to w regule nie pojawi się żaden warunek zanegowany. Wynika to z faktu, że

dodanie do reguły dokładnej kolejnego dowolnego warunku powoduje jedynie ograniczenie jej pokrycia. Ze specyfiki większości z rozpatrywanych w rozdziale 3 miar jakości wynika, że w takiej sytuacji jakość reguły może się jedynie pogorszyć.

Po pokryciu wszystkich przykładów treningowych uzyskujemy wynikowy zbiór reguł, który następnie jest dodatkowo ograniczany za pomocą prezentowanego w kolejnym rozdziale algorytmu *Coverage*. Algorytm ten umożliwia usunięcie reguł pokrywających przykłady pokrywane również przez inne, lepsze reguły.

Na problem redefinicji reguł można spojrzeć również jako na problem indukcji reguł w zbiorze przykładów opisanych za pomocą atrybutów utożsamianych z warunkami elementarnymi i ich negacjami. Atrybuty mają charakter binarny. Jeśli warunek elementarny pokrywa przykład, to odpowiadający mu atrybut przyjmuje wartość 1, w przeciwnym przypadku atrybut przyjmuje wartość 0. Tak zdefiniowany problem redefinicji można traktować jako rodzaj konstruktywnej indukcji sterowanej hipotezami [338].

### 5.2.3. Badanie efektywności algorytmu redefinicji

Aby zbadać efektywność algorytmu redefinicji, analizie poddano następujące zbiorów danych: *australian credit*, *balance scale*, *breast-wisc.*, *bupa*, *dermatology*, *glass*, *heart clev.*, *heart-statlog*, *ionosphere*, *iris*, *lymphography*, *mushrooms*, *parkinsons*, *pima-diabetes*, *post-operative*, *segment*, *sonar*, *splice*, *vehicle*, *yeast*. Charakterystyka tych zbiorów, poza *dermatology*, *parkinsons* i *post-operative*, prezentowana była w rozdziale 4. Nieomawiane w rozdziale 4 zbiorów mają następującą charakterystykę: *dermatology* – 366 przykładów, 6 klas decyzyjnych, 34 atrybuty; *parkinsons* – 195 przykładów, 2 klasy decyzyjne, 22 atrybuty; *post-operative* – 90 przykładów, 3 klasy decyzyjne, 8 atrybutów. Zbiór *post-operative* charakteryzuje się mocno niezrównoważonym rozkładem liczby przykładów reprezentujących kolejne klasy decyzyjne.

Eksperymenty wykonano jedynie dla różnych postaci indeksu Banzhafa. Postępowanie takie można wytlumaczyć wynikami porównań, jakie zamieszczono w pracach [105, 110]. Autorzy tych prac informują, że indeksy Shapleya i Banzhafa generują bardzo podobne lub wręcz identyczne rankingi warunków elementarnych, a różnice dotyczą oceny warunków mniej istotnych.

Obliczenia przeprowadzono, stosując 10-krotną warstwową walidację krzyżową. Wszystkie analizowane warianty algorytmu realizowano, opierając się na tych samych podziałach zbiorów. Bazowe reguły wyznaczano za pomocą algorytmu q-ModLEM, w którym do ustalenia zakresów warunków elementarnych wykorzystano entropię warunkową, a do oceny reguły w fazie przycinania i podczas klasyfikacji użyto miary  $g$  ( $g = 2$ ). Do obliczenia zmodyfikowanej postaci indeksu Banzhafa również użyto miary  $g$ . Badano sześć konfiguracji algorytmu:

1. ocena warunków za pomocą podstawowej postaci indeksu Banzhafa ( $\phi_B$ ), reguły bez warunków negatywnych,
2. ocena warunków za pomocą podstawowej postaci indeksu Banzhafa ( $\phi_B$ ), reguły z warunkami negatywnymi,
3. ocena warunków za pomocą zmodyfikowanej postaci indeksu Banzhafa ( $\phi_{qB}$ ), reguły bez warunków negatywnych,
4. ocena warunków za pomocą zmodyfikowanej postaci indeksu Banzhafa ( $\phi_{qB}$ ), reguły z warunkami negatywnymi,
5. ocena warunków za pomocą uproszczonej postaci zmodyfikowanego indeksu Banzhafa ( $\phi_{sqB}$ ), reguły bez warunków negatywnych,
6. ocena warunków za pomocą uproszczonej postaci zmodyfikowanego indeksu Banzhafa ( $\phi_{sqB}$ ), reguły z warunkami negatywnymi.

Badano również dwie konfiguracje procedury skracania reguł. W pierwszej nie dopuszciano do pogorszenia jakości skracanych reguł, w drugiej dopuszciano do 10% spadku jakości podczas skracania.

W tabeli 5.7 przedstawiono szczegółowe wyniki dla każdego z analizowanych zbiorów danych. Wyniki przytoczono dla ustawień algorytmu, pozwalających na największą redukcję liczby reguł. W tabeli tej zaprezentowano dokładność klasyfikacji (Acc) oraz średnią dokładność klas decyzyjnych (BAcc). Dla porównania zamieszczono również wyniki uzyskane przez bazowy zbiór reguł (q-ModLEM) oraz przez algorytm RIPPER. W tabeli 5.8 zamieszczono informacje o liczbie wyznaczonych reguł, a w tabeli 5.9 przedstawiono uśrednione (dla wszystkich rozważanych zbiorów danych) wyniki dla innych ustawień algorytmu.

Tabela 5.7  
Rezultaty redefinicji reguł – dokładność klasyfikacji

Zbiór danych	q-ModLEM		Redefinicja bez negacji				Redefinicja z negacjami				RIPPER	
			$\phi_{qB}$		$\phi_{sqB}$		$\phi_{qB}$		$\phi_{sqB}$			
	Acc	BAcc	Acc	BAcc	Acc	BAcc	Acc	BAcc	Acc	BAcc	Acc	BAcc
Australian	85.6 $\pm 3.0$	85.6 $\pm 3.2$	84.1 $\pm 2.7$	83.9 $\pm 2.8$	84.5 $\pm 3.0$	84.4 $\pm 3.1$	86.0 $\pm 3.6$	85.3 $\pm 4.0$	85.8 $\pm 3.7$	85.2 $\pm 4.0$	85.6 $\pm 4.1$	86.0 $\pm 4.7$
Balance scale	82.5 $\pm 3.1$	59.7 $\pm 2.2$	83.7 $\pm 3.2$	60.6 $\pm 2.2$	82.2 $\pm 2.8$	59.5 $\pm 1.9$	83.6 $\pm 2.9$	60.5 $\pm 2.0$	81.6 $\pm 2.8$	58.9 $\pm 1.8$	80.9 $\pm 4.0$	61.0 $\pm 7.0$
Breast-Wisc.	95.3 $\pm 2.5$	95.4 $\pm 3.1$	94.8 $\pm 2.8$	94.6 $\pm 3.2$	95.6 $\pm 2.7$	95.7 $\pm 3.3$	95.5 $\pm 2.2$	95.8 $\pm 2.4$	96.0 $\pm 2.2$	96.4 $\pm 2.6$	94.8 $\pm 2.8$	92.8 $\pm 2.8$
Bupa	67.8 $\pm 6.0$	64.6 $\pm 7.0$	62.8 $\pm 5.3$	60.4 $\pm 6.1$	64.4 $\pm 6.0$	61.9 $\pm 6.6$	63.8 $\pm 5.9$	61.5 $\pm 7.2$	69.6 $\pm 9.5$	67.1 $\pm 10.1$	68.0 $\pm 6.7$	66.9 $\pm 6.5$
Dermatology	96.4 $\pm 1.7$	95.7 $\pm 2.7$	95.4 $\pm 4.0$	93.3 $\pm 5.8$	95.6 $\pm 3.9$	94.3 $\pm 5.7$	95.9 $\pm 3.5$	95.0 $\pm 5.5$	95.9 $\pm 3.5$	95.4 $\pm 5.2$	92.1 $\pm 1.0$	91.7 $\pm 1.9$
Glass	67.6 $\pm 9.2$	56.1 $\pm 12.1$	64.1 $\pm 9.6$	53.5 $\pm 7.3$	64.0 $\pm 11.0$	53.1 $\pm 8.4$	68.9 $\pm 9.6$	61.4 $\pm 12.0$	69.7 $\pm 8.2$	61.0 $\pm 11.1$	68.6 $\pm 9.2$	62.2 $\pm 12.5$
Heart clev.	80.6 $\pm 5.9$	80.3 $\pm 6.3$	82.0 $\pm 5.0$	82.1 $\pm 5.0$	81.6 $\pm 5.1$	81.7 $\pm 5.0$	78.9 $\pm 5.8$	78.6 $\pm 6.0$	79.5 $\pm 5.6$	79.0 $\pm 5.8$	80.1 $\pm 7.6$	79.5 $\pm 7.7$
Heart statlog	80.6 $\pm 7.2$	79.4 $\pm 5.8$	81.0 $\pm 6.5$	79.4 $\pm 6.1$	80.6 $\pm 7.2$	79.1 $\pm 6.7$	79.9 $\pm 7.3$	78.8 $\pm 6.6$	80.2 $\pm 6.0$	78.9 $\pm 6.2$	80.8 $\pm 5.8$	80.2 $\pm 6.3$
Ionosphere	91.3 $\pm 4.2$	89.9 $\pm 5.1$	88.2 $\pm 3.9$	86.4 $\pm 5.9$	88.3 $\pm 4.0$	86.2 $\pm 5.9$	89.6 $\pm 4.5$	89.0 $\pm 6.5$	89.6 $\pm 4.5$	88.9 $\pm 6.4$	89.6 $\pm 4.5$	88.1 $\pm 6.7$
Iris	93.3 $\pm 3.1$	93.3 $\pm 3.1$	92.7 $\pm 6.2$	92.7 $\pm 6.2$	92.7 $\pm 6.2$	92.7 $\pm 6.2$	94.7 $\pm 6.0$	94.7 $\pm 6.0$	94.7 $\pm 6.0$	94.7 $\pm 6.0$	94 $\pm 3.5$	94 $\pm 3.5$
Lymphography	80.2 $\pm 9.9$	70.9 $\pm 14.2$	76.8 $\pm 12.0$	70.9 $\pm 20.2$	76.9 $\pm 12.7$	69.7 $\pm 20.1$	80.2 $\pm 10.0$	70.1 $\pm 21.3$	79.8 $\pm 11.0$	69.6 $\pm 21.8$	79.6 $\pm 7.5$	56.5 $\pm 12.8$
Mushrooms	99.9 $\pm 0.1$	99.9 $\pm 0.1$	99.6 $\pm 0.2$	99.6 $\pm 0.2$	99.6 $\pm 0.2$	99.6 $\pm 0.2$	99.7 $\pm 0.2$	99.7 $\pm 0.2$	99.7 $\pm 0.2$	99.7 $\pm 0.2$	100.0 $\pm 0.0$	100.0 $\pm 0.0$
Parkinsons	86.3 $\pm 10.6$	80.0 $\pm 15.0$	86.4 $\pm 9.5$	79.0 $\pm 9.0$	85.9 $\pm 12.0$	80.4 $\pm 12.5$	87.9 $\pm 9.5$	79.8 $\pm 13.8$	87.9 $\pm 9.5$	79.8 $\pm 13.8$	86.2 $\pm 9.2$	82.8 $\pm 12.7$
Pima	74.9 $\pm 3.8$	71.0 $\pm 3.8$	75.5 $\pm 3.6$	69.2 $\pm 3.9$	75.6 $\pm 3.3$	69.7 $\pm 4.4$	74.5 $\pm 3.5$	69.5 $\pm 5.6$	74.6 $\pm 3.3$	69.5 $\pm 4.8$	74.6 $\pm 3.5$	68.6 $\pm 4.4$
Post-operative	66.0 $\pm 8.8$	43.9 $\pm 8.4$	69.2 $\pm 7.5$	48.5 $\pm 11.4$	69.2 $\pm 7.5$	48.3 $\pm 10.3$	69.2 $\pm 7.6$	47.6 $\pm 11.0$	67.5 $\pm 9.8$	47.4 $\pm 10.1$	70.0 $\pm 7.6$	33.4 $\pm 3.2$
Segment	93.2 $\pm 1.5$	93.2 $\pm 1.5$	92.2 $\pm 2.1$	92.2 $\pm 2.1$	92.1 $\pm 2.0$	92.1 $\pm 2.0$	94.7 $\pm 1.7$	95.0 $\pm 1.4$	95.0 $\pm 1.4$	94.3 $\pm 1.8$	94.3 $\pm 1.8$	94.3 $\pm 1.8$
Sonar	75.4 $\pm 8.4$	75.3 $\pm 8.1$	75.9 $\pm 9.5$	75.7 $\pm 9.6$	74.9 $\pm 9.6$	74.7 $\pm 9.9$	75.9 $\pm 10.2$	76.1 $\pm 10.6$	74.5 $\pm 10.2$	74.7 $\pm 10.8$	75.4 $\pm 7.9$	75.5 $\pm 7.8$
Splice	92.6 $\pm 1.5$	92.2 $\pm 2.0$	92.9 $\pm 1.4$	91.7 $\pm 1.7$	93.1 $\pm 2.0$	92.0 $\pm 2.4$	93.9 $\pm 1.4$	93.3 $\pm 2.1$	94.2 $\pm 1.2$	93.8 $\pm 1.5$	93.5 $\pm 1.7$	93.2 $\pm 1.9$
Vehicle	72.0 $\pm 4.8$	72.2 $\pm 4.5$	71.4 $\pm 4.2$	71.5 $\pm 4.2$	71.5 $\pm 4.7$	71.7 $\pm 4.8$	72.0 $\pm 4.0$	72.3 $\pm 3.9$	73.1 $\pm 5.0$	73.2 $\pm 4.9$	66.9 $\pm 3.0$	67.5 $\pm 3.0$
Wine	94.0 $\pm 7.4$	94.1 $\pm 7.2$	95.6 $\pm 5.4$	95.7 $\pm 5.4$	95.6 $\pm 5.4$	95.7 $\pm 5.4$	95.6 $\pm 5.4$	95.7 $\pm 5.4$	95.6 $\pm 5.4$	95.7 $\pm 5.4$	90.9 $\pm 6.1$	91.4 $\pm 5.5$
Yeast	55.8 $\pm 3.2$	43.6 $\pm 5.2$	57.1 $\pm 3.4$	44.7 $\pm 6.0$	56.8 $\pm 2.8$	43.6 $\pm 5.8$	56.9 $\pm 3.2$	44.6 $\pm 5.7$	56.8 $\pm 4.6$	43.6 $\pm 3.9$	56.4 $\pm 3.3$	40.2 $\pm 5.6$
<b>Średnia</b>	<b>82.4</b>	<b>77.9</b>	<b>82.0</b>	<b>77.4</b>	<b>81.9</b>	<b>77.4</b>	<b>82.7</b>	<b>78.3</b>	<b>82.9</b>	<b>78.4</b>	<b>82.0</b>	<b>76.5</b>

Tabela 5.8

## Rezultaty redefinicji reguł – liczba reguł

Zbiór danych	q-ModLEM	Redefinicja bez negacji		Redefinicja z negacjami		RIPPER
		$\phi_{qB}$	$\phi_{sqB}$	$\phi_{qB}$	$\phi_{sqB}$	
Australian	22.5	15	15.3	17.6	16.8	7.7
Balance scale	104.1	30	28.6	29.4	27.1	24.2
Breast-Wisc.	19	9.5	10	9.4	10.3	6.9
Bupa	32.9	10.5	11.5	12.9	12	15.9
Dermatology	20.2	13.5	13	12.9	13.6	9.2
Glass	28.9	19.1	19.5	20.5	20.5	7.9
Heart clev.	17.3	9.9	9.9	9.7	9.7	7.8
Heart statlog	16.5	9.5	9.4	10.2	10.3	7.6
Ionosphere	16.8	10.3	10.0	10.4	9.8	4.9
Iris	7.8	5.2	5.2	6.0	6.0	3.0
Lymphography	20.5	11.7	11.4	10.0	11.0	7.9
Mushrooms	27.8	12.0	12.0	13.4	13.4	8.9
Parkinsons	14.6	8.1	9.4	8.1	8.3	4.0
Pima	45.8	19.1	18.4	21.1	19.9	21.2
Post-operative	17.6	6.8	7.3	6.3	6.5	3.0
Segment	65.9	50.4	51.6	47.6	49.8	21.0
Sonar	26.4	17.3	17.6	11.4	11.9	6.2
Splice	72.4	48.3	52.2	50.7	55.0	20.6
Vehicle	92.7	40.5	39.1	43.5	41.7	16.7
Wine	9.8	8.9	8.9	8.8	8.8	6.7
Yeast	207.4	107.6	103.7	117.1	111.0	137.8
<b>Średnia</b>	<b>42.2</b>	<b>22.1</b>	<b>22.1</b>	<b>22.7</b>	<b>22.5</b>	<b>16.6</b>
<b>Mediana</b>	<b>22.5</b>	<b>12.0</b>	<b>12.5</b>	<b>12.0</b>	<b>12.0</b>	<b>7.9</b>

Aby sprawdzić, czy główną przyczyną redukcji liczby reguł w algorytmie redefinicji nie jest przypadkiem zastosowana na końcu filtracja, algorytm filtracji zastosowano również do wejściowego, nieredefiniowanego zbioru reguł (tabela 5.9, drugi wiersz).

Analizę porównawczą algorytmów wykonano parami, stosując test zgodności par Wilcooxona. Porównywano dokładność klasyfikacji (Acc), średnią dokładność klas decyzyjnych (BAcc) oraz liczbę reguł. Wyniki referencyjne stanowiły te otrzymane przez q-ModLEM.

Analizę wyników rozpoczęmy o porównania liczby reguł. W przypadku algorytmu redefinicji następuje znacząca redukcja liczby reguł, przy niepogorszonych zdolnościach klasyfikacyjnych, zarówno jeśli chodzi o dokładność klasyfikacji, jak i średnią dokładność klas decyzyjnych. Różnice pomiędzy poszczególnymi konfiguracjami algorytmu redefinicji są tutaj niewielkie i statystycznie nieistotne. Różnice pomiędzy liczbą reguł wyznaczonych przez q-ModLEM i algorytm redefinicji wynoszą około 50% i są statystycznie istotne ( $p$ -wartość=0.00006).

Tabela 5.9  
Rezultaty redefinicji reguł dla różnych konfiguracji algorytmu

Algorytm	Acc	BAcc	Liczba reguł
q-ModLEM	82.4	77.9	42.2
q-ModLEM + filtracja	82.1	75.0	30.6
RIPPER	82.0	76.5	16.6
$\phi_B$ 0%	82.7	78.2	28.1
$\phi_B$ 10%	81.8	77.5	23.0
$\phi_B$ 0% neg	83.2	79.4	28.9
$\phi_B$ 10% neg	82.5	77.8	23.4
$\phi_{sB}$ 0%	81.4	77.8	27.5
$\phi_{sB}$ 10%	81.8	77.2	22.9
$\phi_{sB}$ 0% neg	82.8	78.9	29.5
$\phi_{sB}$ 10% neg	82.7	78.3	23.7
$\phi_{qB}$ 0%	82.6	77.9	26.2
$\phi_{qB}$ 10%	82.0	77.4	22.1
$\phi_{qB}$ 0% neg	82.9	78.8	27.8
$\phi_{qB}$ 10% neg	82.7	78.3	22.7
$\phi_{sqB}$ 0%	82.0	77.5	26.1
$\phi_{sqB}$ 10%	81.9	77.4	22.1
$\phi_{sqB}$ 0% neg	82.6	78.2	27.7
$\phi_{sqB}$ 10% neg	82.9	78.4	22.5

Kolejnym punktem odniesienia dla liczby wyznaczanych reguł był algorytm RIPPER. Użyta implementacja tego algorytmu (pochodząca z pakietu Weka [333]) nie generuje reguł dla najliczniejszej klasy decyzyjnej. Algorytm RIPPER generuje mniej reguł niż algorytm redefinicji i różnica ta jest statystycznie istotna (p-wartość od 0.002 do 0.005 w zależności od konfiguracji algorytmu redefinicji). Porównując jedynie liczbę reguł generowanych dla wszystkich klas, poza klasą najliczniejszą, otrzymujemy podobną liczbę reguł i statystycznie nieistotne różnice pomiędzy algorytmami (p-wartość  $\approx 0.2$ ). Wyniki te należy uznać za bardzo dobre, gdyż RIPPER jest jednym z najefektywniejszych – pod względem liczby wyznaczanych reguł – algorytmów indukcji. Warto wspomnieć o tym, że po zastosowaniu redefinicji liczba warunków elementarnych maleje średnio o 36%, a wyznaczone reguły pokrywają unikalnie więcej przykładów niż reguły wejściowe.

W tabelach 5.8 i 5.9 widać także, zgodnie z oczekiwaniemi, że liczba reguł maleje, jeśli w fazie przycinania dopuszcza się do spadku jakości reguł. Różnice pomiędzy liczbą reguł generowanych przez różne konfiguracje algorytmu redefinicji są nieznaczne.

W analizie dokładności utworzonych klasyfikatorów wykonano porównania pomiędzy algorytmem q-ModLEM a algorytmem redefinicji oraz algorytmem RIPPER a algorytmem redefinicji. Statystycznie istotne różnice wystąpiły pomiędzy:

- algorytmem redefinicji z parametrami ( $\phi_{sqB}$  10% neg) a algorytmem RIPPER (p-wartość=0.029);
- algorytmem redefinicji z parametrami ( $\phi_{qB}$  0% neg) a algorytmem RIPPER (p-wartość=0.026);
- algorytmem redefinicji z parametrami ( $\phi_B$  0% neg) a algorytmem RIPPER (p-wartość=0.017)
- algorytmem redefinicji z parametrami ( $\phi_B$  0% neg) a algorytmem q-ModLEM (p-wartość=0.032).

O tym, jakie będą różnice pomiędzy czasami działania algorytmu redefinicji, w którym stosuje się podstawowe i uproszczone formy indeksu Banzhafa, decyduje: liczba reguł wejściowych, liczba warunków elementarnych, z jakich zbudowane są reguły wejściowe, oraz liczba przykładów. Czas obliczeń niezbędnych do wyznaczenia wartości indeksu Banzhafa w formie podstawowej rośnie wykładniczo wraz ze wzrostem liczby warunków elementarnych, tworzących regułę. Dla analizowanych danych wyraźnie zauważalne różnice w czasie obliczeń odnotowano dla zbiorów: *Australian*, *Pima-diabetes*, *Segment*, *Yeast*. Największą różnicę odnotowano dla zbioru *Yeast*. Redefinicja za pomocą indeksów podstawowych i zmodyfikowanych trwała około 3 razy dłużej niż redefinicja za pomocą indeksów uproszczonych. Czas obliczeń, niezbędny do wykonania pełnej, 10-krotnej walidacji krzyżowej dla zbioru *Yeast* i podstawowej postaci indeksu Banzhafa, nie był dłuższy niż 10 minut (eksperyment wykonano na komputerze wyposażonym w procesor Intel Core Duo 3.5 GHz). Pośród analizowanych zbiorów danych znajdował się również, liczniejszy niż *Yeast*, zbiór *Mushrooms*. Jednakże ze względu na niewielką liczbę warunków elementarnych, tworzących przesłanki reguł wejściowych (średnio 2 warunki), nie odnotowano zauważalnych różnic pomiędzy różnymi konfiguracjami algorytmu. W zbiorach *Australian*, *Pima-diabetes*, *Segment*, *Yeast* średnia liczba warunków elementarnych w wejściowych regułach wynosiła odpowiednio 3, 3.7, 3.5, 4.3.

Przeprowadzono także eksperyment na zbiorze *Nursery*. Zbiór ten złożony jest z 12960 przykładów i 8 atrybutów symbolicznych. Algorytm q-ModLEM tworzył w tym zbiorze średnio 362 reguły, złożone średnio z 4.9 warunków elementarnych. Czas trwania jednego eksperymentu w ramach walidacji krzyżowej dla uproszczonej postaci indeksu Banzhafa wynosił 29 minut. Dla podstawowej postaci indeksu czas ten wynosił 6 godzin i 50 minut. Różnice pomiędzy dokładnością klasyfikacji klasyfikatorów tworzonych za pomocą uproszczonej i podstawowej wersji indeksu Banzhafa wynosiły mniej niż 1.5%, a w odniesieniu do liczby reguł – mniej niż 10%.

Przyjrzymy się jeszcze podobieństwom pomiędzy rankingami warunków, tworzonymi za pomocą różnych postaci indeksu Banzhafa. W tablicy 5.10 zaprezentowano wartości współczynnika korelacji  $\tau$  Kendalla dla rankingów otrzymanych przez: podstawowy indeks

Banzhafa i uproszczony indeks Banzhafa oraz podstawowy indeks Banzhafa i zmodyfikowany indeks Banzhafa. W przypadku pierwszego porównania wartość korelacji wynosiła co najmniej 0.88. Rankingi różniły się nieznacznie, różnice dotyczyły warunków umiejscowionych w ich środkowej części. Oznacza to, że porównywane indeksy podobnie oceniały najlepsze i najgorsze warunki elementarne. Średnia rozbieżność w ocenie warunków wynosiła 2 pozycje w rankingu (np. indeks podstawowy umieszczał dany warunek na 9 miejscu, a indeks uproszczony – na 11 miejscu).

Różnice pomiędzy rankingami otrzymanymi na podstawie indeksu podstawowego i zmodyfikowanego (wykorzystującego miarę  $g$ ) były większe. Porównując rankingi, można było zauważyć pewną prawidłowość. Im większa była liczba warunków elementarnych, tworzących przesłanki reguł, tym różnice pomiędzy rankingami były większe.

Tabela 5.10  
Wartość współczynnika korelacji  $\tau$  Kendalla pomiędzy rankinami warunków elementarnych, utworzonych przez podstawowe, uproszczone i zmodyfikowane postaci indeksu Banzhafa

Zbiór danych	Wartość współczynnika korelacji $\tau$ Kendalla	
	$\phi_B$ vs. $\phi_{SB}$	$\phi_B$ vs. $\phi_{qB}$
Australian	0.91	0.87
Balance scale	0.98	0.93
Breast-Wisc.	0.91	0.87
Bupa	0.93	0.74
Dermatology	0.99	0.93
Glass	0.94	0.82
Heart clev.	1.00	0.91
Heart statlog	0.88	0.84
Ionosphere	0.99	0.89
Iris	0.93	0.93
Lymphography	0.96	0.87
Mushrooms	0.89	0.55
Parkinsons	0.88	0.81
Pima	0.95	0.66
Post-operative	0.96	0.80
Segment	0.97	0.86
Sonar	0.96	0.88
Splice	0.99	0.90
Vehicle	0.91	0.76
Wine	1.00	0.91
Yeast	0.97	0.75
<b>Średnia</b>	<b>0.95</b>	<b>0.83</b>

Wyniki przeprowadzonych badań pokazują, że algorytm redefinicji pozwala na znaczne ograniczenie liczby reguł tworzących klasyfikator. Redukcji ulega także liczba warunków elementarnych. Dokładność i średnia dokładność klas decyzyjnych zbiorów reguł po redefinicji pozostają niepogorszone lub są lepsze niż dokładności zbiorów wejściowych.

Ze względu na złożoność obliczeniową, związaną z obliczaniem wartości podstawowego indeksu Banzhafa, w przypadku większych zbiorów danych celowe jest stosowanie jego postaci uproszczonej. Stosując uproszczoną postać indeksu, należy liczyć się z możliwością nieznacznego spadku dokładności klasyfikacji. Zastosowanie indeksów zmodyfikowanych prowadzi do tworzenia najmniejszych zbiorów reguł.

Ostatecznie jako standardowy zestaw parametrów dla algorytmu redefinicji można rekomendować: stosowanie indeksu podstawowego (lub podstawowego uproszczonego w przypadku większych – kilka tysięcy rekordów – zbiór danych); użycie negacji warunków elementarnych w redefiniowanych regułach; niedopuszczanie do spadku jakości reguł podczas skracania.

Przedstawiony algorytm może być stosowany do redefinicji dowolnego zbioru reguł decyzyjnych. Reguły te mogą zawierać zarówno proste, jak i złożone warunki elementarne. W przeprowadzonych eksperymentach analizowano jedynie nieuporządkowane zbiory reguł. W zbiorach uporządkowanych o ważności danego warunku decyduje również to, w której kolejnej (kolejnych) regule (regułach) warunek ten się pojawia.

Za wadę algorytmu należy uznać czas niezbędny do wyznaczenia rankingu ważności warunków elementarnych. W przypadku zbioru *nursery* czas związany z wykonaniem jednego eksperymentu dla indeksów uproszczonych wynosił około 30 minut. Celowe i stosunkowo proste byłoby opracowanie równoległej wersji algorytmu, w której wartość indeksu Banzhafa obliczana byłaby dla różnych warunków elementarnych w niezależnych procesach.

Rezultaty badań prezentowanych w niniejszym rozdziale autor zaważył w publikacji [266].

### **5.3. Filtracja reguł**

Filtracja polega na usuwaniu ze zbioru tych reguł, które ze względu na wartość ustalonego kryterium jakości okazują się regułami zbędnymi. Z założenia filtracja jest techniką przycinania, która nie ingeruje w postać wyznaczonych reguł. W literaturze takie podejście do filtracji reguł decyzyjnych prezentowane było m.in. przez Ägotnesa [5], autora [244, 249, 254, 264] oraz Gambergera i Lavrac [93]. Problem znalezienia minimalnego podzbioru reguł, maksymalizującego wartość zadanego kryterium jakości, jest problemem NP-zupełnym. Z tego też powodu podczas filtracji stosowane są heurystyczne metody przeszukiwania. Ägotnes do znalezienia quasi-minimalnego zbioru reguł stosuje m.in. algorytm genetyczny. W populacji każdy osobnik to klasyfikator, a wystąpienie danej reguły w klasyfikatorze sygnalizowane jest pojawieniem się wartości 1 na odpowiedniej pozycji zakodowanego osobnika. Optymalizacja prowadzona jest w kierunku maksymalizacji funkcji będącej sumą ważoną, odzwierciedlającą zdolności klasyfikacyjne (dokładność klasyfikacji) oraz opisowe (liczba reguł) klasyfikatora. Algorytm genetyczny stosuje również

Ishibuchi wraz ze współpracownikami [139], wybierając ze zbioru reguł rozmytych grupę reguł, które następnie tworzą klasyfikator. Gamberger i Lavrac [93] proponują algorytm eliminacji reguł pokrywających podobne zbiory przykładów. Zakładają oni, że wejściowy zbiór zawiera pewną liczbę reguł podobnych. Reguły te niekoniecznie muszą być wyznaczone przez jeden algorytm indukcji. Możliwe jest łączenie zbiorów reguł tworzonych w sposób automatyczny z regułami definiowanymi przez eksperta. Eliminacja reguł zbędnych odbywa się, począwszy od tych, które pokrywają najmniejszą liczbę przykładów pozytywnych. Inne podejście do filtracji prezentuje autor [244, 254], przedstawiając strategie filtracji: wstępującą i zstępującą. Zostaną one opisane w dalszej części rozdziału.

Gdy zdolności klasyfikacyjne nie stanowią nadzawanego celu indukcji lub indukcja odbywa się jedynie dla celów opisowych, filtracja polega na eliminacji reguł niespełniających wymogów minimalnej jakości. W zadaniach filtracji użyteczne będą wtedy przywoływane już wyniki badań Bayardo i Agrawal [15] oraz Brzezińskiej, Greco i Słowińskiego [43, 300], dotyczące Pareto-optymalnych zbiorów reguł ze względu na wsparcie i dokładność oraz wsparcie i miary konfirmacji i s.

Filtrację można także zrealizować poprzez: usunięcie reguł statystycznie nieistotnych [44, 257, 329], grupowanie i wybór reguł reprezentujących wyznaczone grupy [153], zdefiniowanie miary złożonej i wybór reguł najlepszych ze względu na wartości tej miary [117, 257].

W dalszej części rozdziału omówione zostaną cztery metody filtracji. Autor stosował je w realizacji wielu zadań analitycznych, polegających zarówno na indukcji reguł dla celów opisowych (np. [257]), jak i indukcji dla celów klasyfikacyjnych (np. [258]).

W przedstawionych dalej algorytmach podstawę do filtracji stanowią: jakość pojedynczych reguł oraz zdolności klasyfikacyjne zbioru reguł.

Najprostsza metoda filtracji polega na usunięciu wszystkich reguł niespełniających wymagań minimalnej jakości. Jak już wspomiano, najczęściej definiuje się minimalną dokładność i minimalne pokrycie (lub wsparcie), jakimi muszą charakteryzować się reguły. Do mierzenia jakości reguł mogą być stosowane dowolne miary zdefiniowane w rozdziale 3 (zakładamy, że będą one miarami korzyści). Ten rodzaj filtracji nie uwzględnia ani zdolności klasyfikacyjnych, ani warunków definicji 2.5 (definicji opisu klasy decyzyjnej). Przy zbyt wysokich wymaganiach dotyczących jakości reguł klasyfikator utworzony na podstawie przefiltrowanego zbioru reguł może charakteryzować się niewielką zdolnością rozpoznawania przykładów testowych.

Drugi sposób filtracji polega na sprawdzeniu, czy dla ustalonej reguły  $r_i$  istnieje reguła  $r_j$  o jakości nie gorszej niż  $r_i$  (tzn.  $q(r_i) \leq q(r_j)$ ) i pokrywająca wszystkie przykłady pozytywne, pokrywane przez  $r_i$  ( $[r_i] \subseteq [r_j]$ ). Jeśli reguła taka istnieje, to  $r_i$  jest usuwana ze

zbioru reguł. Filtracja dokonuje się oddzielnie w obrębie reguł wskazujących na identyczną klasę decyzyjną. Ten rodzaj filtracji nazywany będzie *Inclusion*.

Trzecia metoda filtracji polega na posortowaniu reguł opisujących klasę decyzyjną nierośnaco ze względu na ustalone kryterium jakości i wykorzystaniu tak otrzymanego porządku do budowy pokrycia przykładów pokrywanych przez wejściowy zbiór reguł. Pokrycie budowane jest zgodnie z kolejnością reguł w rankingu. Po dodaniu kolejnej reguły do przefiltrowanego zbioru reguł ze zbioru przykładów usuwane są wszystkie przykłady pokrywające tę regułę. Przefiltrowany zbiór reguł rozrasta się dopóty, dopóki reguły w nim zawarte nie pokryją wszystkich przykładów pokrywanych przez wejściowy (poddawany filtracji) zbiór reguł. Przedstawiony sposób filtracji nazywany będzie *Coverage* [244]. Algorytm podobny do *Coverage* przedstawiono ostatnio w [185], gdzie filtracja odbywa się na podstawie tzw. domknięcia grupy reguł pokrywających podobne zbiory przykładów. Eliminację reguł przeprowadza się na podstawie porównania pokryć reguł należących do danego domknięcia. Metoda nadaje się do redukcji zarówno reguł decyzyjnych, jak i asocjacyjnych. Autorzy przeprowadzili badania na 15 benchmarkowych zbiorach danych, otrzymując średnio 15% redukcję liczby reguł bez istotnych strat w dokładności klasyfikacji.

Kolejne metody filtracji biorą pod uwagę dokładność klasyfikacji wyjściowego zbioru reguł. Pierwsza metoda (*Forward*) jest metodą wstępującą, w której zbiory reguł, opisujące klasy decyzyjne, rozszerzane są iteracyjnie. Reguła dopisywana jest do wynikowego zbioru reguł, jeśli jej dodanie spowoduje wzrost jakości (linie 10,11) tego zbioru. Jakość zbioru reguł obliczana jest na podstawie zbioru walidacyjnego  $V$ . Jakość ta może być definiowana w różny sposób. Może ona uwzględniać zarówno opisowe, jak i klasyfikacyjne zdolności zbioru reguł. Przykładowo, w przytoczonych w dalszej części rozdziału wynikach eksperymentów o jakości zbioru reguł decydowała dokładność klasyfikacji utworzonego na jego podstawie klasyfikatora. Proces filtracji kończy się z chwilą rozpatrzenia wszystkich reguł lub jest przerywany, jeśli wynikowy zbiór reguł charakteryzuje się jakością taką samą jak zbiór wejściowy (linie 12, 13). Reguły dodawane są do wynikowego zbioru zgodnie z porządkiem ustanowionym przez miarę jakości  $q$ . W algorytmie ważna jest również kolejność klas decyzyjnych. W każdym zasadniczym kroku algorytmu (pętla, linie 8–14) do opisu klasy decyzyjnej dodawana jest jedna reguła. Podczas filtracji w pierwszej kolejności zwiększa się liczliwość zbiorów reguł wskazujących na te klasy decyzyjne, które przez wejściowy zbiór reguł klasyfikowane były z najmniejszą dokładnością. Zbiór walidacyjny  $V$  może być zbiorem niezależnym lub zależnym od zbioru treningowego  $U$ . W najprostszym przypadku  $V=U$ .

Wstępujący algorytm filtracji zbioru reguł - Forward

Wejście:  $U$  - zbiór przykładów treningowych

$V$  - zbiór przykładów walidacyjnych

$q$  - miara jakości reguł

$X_j$  -  $j$ -ta klasa decyzyjna

$RUL$  - zbiór reguł podzielony na opisy klas decyzyjnych  $RUL_{X_j}$

$\text{Opt}(R, D)$  - kryterium jakości zbioru reguł  $R$ , obliczone na podstawie zbioru przykładów  $D$

Wyjście:  $\text{filtrRUL}$  - przefiltrowany zbiór reguł

Założenie: Klasa decyzyjne ponumerowane są tak, aby

$\forall i < j \text{ } \text{Opt}(RUL_{X_i}, X_i) \leq \text{Opt}(RUL_{X_j}, X_j)$

```

1. Begin
2.   Oblicz jakość każdej reguły z  $RUL$ , używając miary  $q$  oraz zbioru przykładów  $U$ .
    Posortuj reguły znajdujące się w zbiorach  $RUL_{X_j}$ ,  $j \in \{1, \dots, k\}$ , w porządku
    leksykograficznym, niemalejaco najpierw ze względu na wartość miary
     $q$ , a następnie ze względu na pokrycie reguł.
3.    $\text{filtrRUL} := \emptyset;$ 
4.    $\text{Opt}(\text{filtrRUL}, V) := 0;$ 
5.   ForEach  $X_j$ ,  $j \in \{1, \dots, k\}$  do  $\text{rules}_{X_j} := \text{True}$ 
6.    $i := 1;$ 
7.   While ( $\text{rules}_{X_1}$  or  $\text{rules}_{X_2}$  or .. or  $\text{rules}_{X_k}$ ) do
8.     ForEach  $j := 1, \dots, k$  do
9.       If  $\text{rules}_{X_j}$  then

        // dodawanie kolejnych reguł do zbioru wynikowego
        //  $r_{ji}$  oznacza  $i$ -tą regułę z opisu  $j$ -tej klasy
10.      If  $\text{Opt}(\text{filtrRUL}, V) < \text{Opt}(\text{filtrRUL} \cup \{r_{ji}\}, V)$  then
11.         $\text{filtrRUL} := \text{filtrRUL} \cup \{r_{ji}\};$ 
12.        If  $i = |RUL_{X_j}|$  then  $\text{rules}_{X_j} := \text{False};$ 
13.        If  $\text{Opt}(\text{filtrRUL}, V) \geq \text{Opt}(RUL, V)$  then  $\forall j \in \{1, \dots, k\} \text{ } \text{rules}_{X_j} := \text{False};$ 
14.      end foreach
15.       $i := i + 1;$ 
16.   end while
17. end.

```

Algorytm działający według zasady zstępującej (*Backwards*) w sposób iteracyjny usuwa reguły z każdej klasy decyzyjnej. Jeśli usunięcie reguły nie powoduje spadku jakości klasyfikatora, to jest ona usuwana.

Standardowo w algorytmach filtracji *Inclusion*, *Coverage*, *Forward* i *Backwards* stosowana jest jedna z obiektywnych miar jakości, definiowanych na podstawie tablicy kontyngencji. Nie ma jednak przeszkód, aby porządek reguł w obrębie każdego zbioru  $RUL_{X_j}$  tworzony był w inny, zgodny z intencjami użytkownika, sposób.

Algorytmy *Inclusion* i *Coverage* zdefiniowano na podstawie spostrzeżenia, że pomimo stosowania paradygmatu kolejnych pokryć zbiory reguł tworzone za pomocą algorytmów pokryciowych zawierają reguły nadmiarowe. Zauważmy, że dla obu tych algorytmów, jeśli wejściowy zbiór reguł spełnia warunki opisu klas decyzyjnych (zgodnie z definicją 2.5), to zbiór wyjściowy również będzie spełniał te warunki. Zastosowanie algorytmów *Forward* i *Backwards* takiej gwarancji nie daje (nawet jeśli przyjmiemy założenie, że  $V := U$ ). Zauważmy również, że jeśli zbiór reguł składa się z opisów minimalnych, to algorytmy *Inclusion* i *Coverage* nie usuną żadnej reguły. Algorytmy *Forward* i *Backwards* mogą

usuwać reguły z opisów minimalnych, co spowoduje utratę zdolności rozpoznawania przykładów treningowych.

Aby oszacować złożoność obliczeniową przedstawionych algorytmów, przyjmijmy następujące oznaczenia:  $L$  to liczba reguł podlegających filtracji,  $m$  to liczba atrybutów warunkowych,  $n$  to liczba przykładów w tablicy treningowej. Przyjmijmy także, że filtracja *Forward* i *Backwards* odbywa się na podstawie treningowego zbioru przykładów.

Złożoność obliczeniowa algorytmów filtracji, biorących pod uwagę jedynie jakości wyznaczonych reguł (*Inclusion, Coverage*), jest rzędu  $O(Lnm)$ . Dominującą operacją w tych algorytmach jest obliczanie jakości reguł. Złożoność obliczeniowa algorytmów *Forward* i *Backwards* zależy od przyjętego sposobu klasyfikacji. Jeśli przyjmiemy, że klasyfikacja odbywa się przez głosowanie, a zaufanie do reguł określone jest na podstawie ich jakości, to złożoność obliczeniowa obu algorytmów jest rzędu  $O(L^2nm)$ .

Dla celów praktycznych zastosowania algorytmów *Forward* i *Backwards* mogą zostać sparametryzowane. Poprzez parametry można określić maksymalną wynikową liczbę reguł oraz dopuszczalny spadek jakości klasyfikatora. Dla tych algorytmów możliwe jest również narysowanie wykresu ilustrującego, jak w kolejnych krokach działania algorytmu wzrasta (algorytm *Forward*) lub spada (algorytm *Backwards*) wartość kryterium oceniającego jakość zbioru reguł (w szczególności dotyczy to dokładności klasyfikacji). Na wykresie oś odciętych informowałaby o liczbie reguł, a oś rzędnych – o związanej z tymi regułami jakości. Wykresy takie można utworzyć oddzielnie dla każdej klasy decyzyjnej.

Efektywność przedstawionych algorytmów filtracji zbadano na 34 zbiorach danych, stosowanych w rozdziale 4 do badania efektywności miar jakości, definiowanych na podstawie tablicy kontyngencji. W tabelach 5.11 i 5.12 przedstawiono wyniki filtracji zbiorów reguł, utworzonych przez algorytm q-ModLEM. W algorytmie tym użyto miar należących do zbioru miar najbardziej obiecujących, oznaczonego w rozdziale 5 jako *Max*. W tabeli 5.11 przedstawiono wyniki filtracji zbiorów reguł otrzymanych przez algorytm q-ModLEM(MMM), w tabeli 5.12 umieszczono wyniki filtracji reguł otrzymanych za pomocą algorytmu q-ModLEM(EMM).

Znakiem (-) oznaczano statystycznie gorsze wyniki (test kolejności par Wilcooxona, poziom istotności 0.05), znakiem (+) oznaczono wyniki statystycznie lepsze. Zbiorem referencyjnym był wejściowy zbiór reguł. W algorytmach *Forward* i *Backwards* w czasie filtracji jako kryterium optymalności zbioru reguł przyjęto całkowitą dokładność klasyfikacji treningowego zbioru przykładów.

Analizując wyniki zamieszczone w tabelach 5.11, 5.12, można zauważyc, że niezależnie od miary jakości i sposobu wyznaczania reguł algorytm *Inclusion* redukuje liczbę reguł w najmniejszym stopniu, a algorytm *Backwards* – w stopniu najwyższym. Niezależnie od sposobu indukcji widoczna jest również prawidłowość, że filtracja reguł utworzonych na

podstawie miar przywiązujących dużą wagę do dokładności reguł ( $C1$ ,  $C2$ ,  $g$ ,  $wLap$ ,  $LS$ ) powoduje spadek dokładności klasyfikacji.

Tabela 5.11

Wyniki filtracji zbiorów reguł utworzonych  
przez algorytm q-ModLEM(MMM)

Miara	Wejściowy		Inclusion		Coverage		Forward		Backwards	
	Acc	Rul	Acc	Rul	Acc	Rul.	Acc.	Rul	Acc	Rul
$C2$	82.3	127	82.1	103	81.8(-)	79	81.6	68	80.8(-)	52
$C1$	82.0	151	81.6(-)	119	81.0(-)	90	80.8(-)	83	80(-)	63
$g$	81.4	145	81.5	122	81.3	89	81.2	79	80.2(-)	54
$wLap$	81.1	157	81.0	127	80.5(-)	93	80.6	87	79.5(-)	63
$LS$	80.7	172	80.6	131	79.9(-)	99	79.9(-)	97	78.8(-)	71
$s$	80.3	86	80.4	65	80.0	52	80.7	47	80.5	37
$RSS$	78.2	34	78.1	30	78.0	26	79.6	12	80.2(+)	24
$Corr$	79.9	44	79.8	39	79.5	35	80.8	17	81.1	28
$MS$	77.8	33	77.8	29	77.6	26	80.1(+)	11	80.0(+)	22

Tabela 5.12

Wyniki filtracji zbiorów reguł utworzonych  
przez algorytm q-ModLEM(EMM)

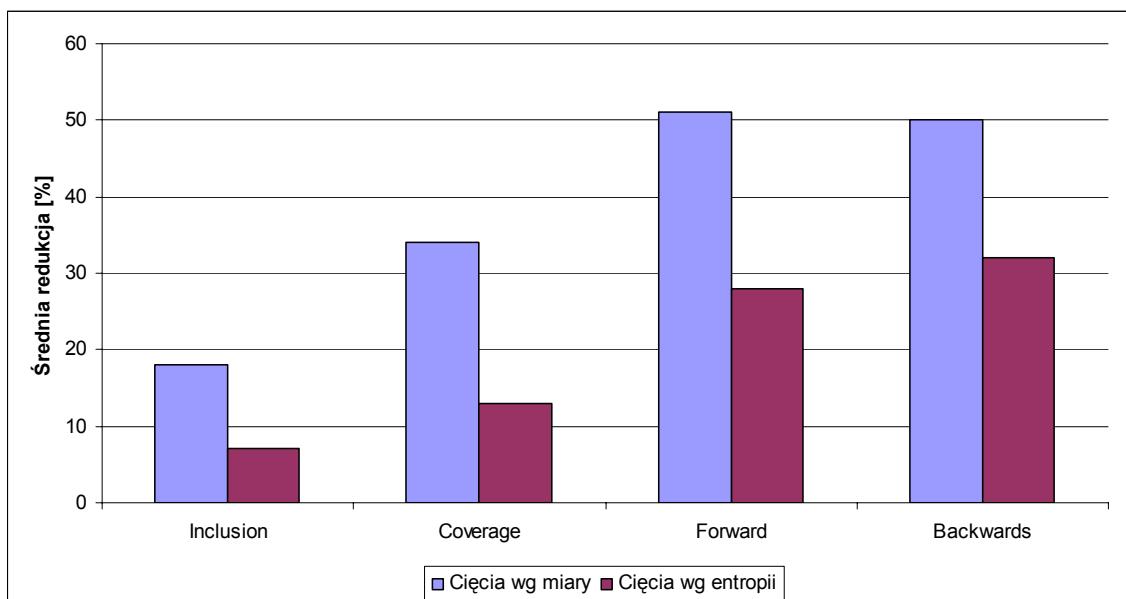
Miara	Wejściowy		Inclusion		Coverage		Forward		Backwards	
	Acc	Rul	Acc	Rul	Acc	Rul.	Acc.	Rul	Acc	Rul
$C2$	81.6	67	81.5	63	81.4	58	81.4	53	80.5	49
$C1$	80.5	73	80.5	69	80.3	64	80.4	61	79.9 (-)	56
$g$	81.8	51	81.7	49	81.7	48	81.5	40	81.0 (-)	34
$wLap$	80.3	65	80.4	63	80.2	60	80.3	56	79.5 (-)	49
$LS$	80.1	78	80.0	74	79.8(-)	68	79.8(-)	68	79.5 (-)	62
$s$	79.7	66	79.6	60	79.5	57	79.8	54	80.0	52
$RSS$	77.5	18	77.7	16	77.6	15	78.7 (+)	9	80.0 (+)	9
$Corr$	80.0	26	80.1	22	80.1	21	81.1 (+)	13	81.7 (+)	13
$MS$	77.3	15	77.3	14	77.3	13	79.8 (+)	8	80.0 (+)	9

Dla miar preferujących reguły o wysokim pokryciu notujemy wzrost zdolności klasyfikacyjnych. Zdolności klasyfikacyjne reguł otrzymanych w efekcie stosowania algorytmu *Backwards* są statystycznie gorsze od zdolności klasyfikacyjnych zbiorów niepoddawanych filtracji, przy czym dotyczy to jedynie miar  $C1$ ,  $C2$ ,  $g$ ,  $wLap$ ,  $LS$  i jest niezależne od metody indukcji reguł (q-ModLEM(EMM) lub q-ModLEM(MMM)). Algorytmy *Coverage* i *Forward* pogarszają zdolności klasyfikacyjne reguł utworzonych

przez q-ModLEM(MMM) (tabela 5.11). W pozostałych sytuacjach filtracja pozwala na ograniczenie liczby reguł bez utraty zdolności klasyfikacyjnych.

Z przeprowadzonych badań wynika również, że spadek (wzrost) średniej dokładności klas decyzyjnych (*balanced accuracy*) jest prawie identyczny ze spadkiem (wzrostem) dokładności klasyfikacji.

Niezaprzeczelną zaletą algorytmów filtracji jest redukcja liczby reguł opisujących dane i biorących udział w klasyfikacji. Stopień redukcji jest zawsze (poza nielicznymi przypadkami użycia algorytmu *Inclusion*) statystycznie istotny. Prezentowane na rysunku 5.5 średnie wartości redukcji otrzymano, obliczając dla każdego algorytmu filtracji średnią arytmetyczną ze średniej redukcji każdej z 9 rozważanych miar jakości.



Rys. 5.5. Średnia procentowa redukcja liczby reguł po zastosowaniu filtracji  
Fig. 5.5. Average percentage reduction of the number of rules after filtration

Interesujące jest również to, że tak znaczna redukcja liczby reguł nie odbywa się kosztem obniżenia możliwości rozpoznawania przykładów testowych. Niezależnie od tego, w jaki sposób dokonano indukcji reguł, zredukowane zbiorы reguł rozpoznają do 3% mniej przykładów testowych niż zbior wejściowe. Największy spadek (maksymalnie 3%) zanotowano dla algorytmu *Backwards*. Wyniki dla pozostałych algorytmów są następujące: *Inclusion* – do 0.6%, *Coverage* – do 0.7%, *Forward* – do 1.7%.

Filtracja pozwala nie tylko na ograniczenie liczby reguł, ale także powoduje, że podnoszą się wartości takich parametrów wynikowego zbioru reguł, jak: średnia jakość reguł, średni poziom statystycznej istotności oraz stosunek liczby przykładów pozytywnych, pokrywanych unikalnie przez regułę, do ogólnej liczby pokrywanych przez nią przykładów pozytywnych (w tabeli 5.13 wartość ta oznaczana jest jako U. i podawana jest w procentach). Ponadto po filtracji spadają: liczba konfliktów klasyfikacji (ozn. V) oraz liczba konfliktów

rozstrzyganych w sposób błędny (ozn. negV). Dla algorytmu q-ModLEM(MMM) wartość V przed filtracją wynosi średnio 42%, a negV – średnio 9.2%. Po filtracji średnie spadki V i negV to odpowiednio 15% i 4%. Dla algorytmu q-ModLEM(EMM) średnie spadki V i negV wynoszą odpowiednio 15.6% i 4.6%. Największe spadki odnotowano dla algorytmu *Forward*.

Tabela 5.13

Wyniki filtracji – zmiany liczby przykładów pokrywanych unikalnie oraz poziomu statystycznej istotności reguł

Miara	Wejściowy		Inclusion		Coverage		Forward		Backwards	
	U	p <sub>val</sub>	U	p <sub>val</sub>	U	p <sub>val</sub>	U	p <sub>val</sub>	U	p <sub>val</sub>
Cięcia według miary (MMM)										
C2	4.7	0.04	6.9	0.03	7.6	0.03	8.4	0.03	11.7	0.03
C1	4.0	0.05	5.8	0.04	6.5	0.04	6.9	0.04	8.8	0.05
g	4.6	0.01	5.8	0.01	6.6	0.017	7.3	0.00	9.1	0.00
wLap	3.9	0.02	5.4	0.01	6.1	0.02	6.3	0.01	8.2	0.02
LS	3.8	0.06	5.4	0.05	6.0	0.05	6.1	0.05	8.0	0.05
S	5.2	0.04	7.2	0.03	7.9	0.03	10.6	0.03	12.7	0.03
RSS	12.0	0.01	14.0	0.01	14.6	0.01	24.1	0.01	19.6	0.01
Corr	8.3	0.01	10.3	0.01	10.6	0.01	19.0	0.01	17.1	0.01
MS	8.1	0.03	10.2	0.03	10.5	0.03	25.0	0.02	16.4	0.03
Cięcia według entropii (EMM)										
C2	8.6	0.08	9.0	0.07	9.7	0.07	12.1	0.05	14.9	0.06
C1	8.3	0.07	8.9	0.07	9.2	0.07	11.2	0.05	12.9	0.06
g	8.3	0.03	8.9	0.03	9.1	0.03	11.0	0.01	13.5	0.02
wLap	8.1	0.05	8.7	0.05	8.9	0.04	10.4	0.03	12.3	0.04
LS	8.3	0.07	8.8	0.07	9.1	0.06	9.9	0.06	10.4	0.07
S	8.9	0.08	10.0	0.07	10.3	0.07	14.3	0.06	13.7	0.06
RSS	14.1	0.06	15.9	0.05	16.3	0.05	32.1	0.01	26.1	0.01
Corr	10.7	0.06	12.4	0.05	12.8	0.05	24.0	0.02	22.2	0.02
MS	12.5	0.10	13.9	0.10	14.4	0.10	35.5	0.04	25.6	0.07

Przedstawione wyniki badań pokazują, że podjęcie próby filtracji jest działaniem jak najbardziej pożdanym, które może prowadzić do znacznej redukcji liczby reguł. Złożoność algorytmów filtracji jest umiarkowana. W przeprowadzonych badaniach czas działania najbardziej złożonych algorytmów filtracji *Forward* i *Backwards* nie wynosił więcej niż kilka minut.

Efektywność algorytmów filtracji sprawdzono również dla reguł generowanych za pomocą algorytmu RMatrix. Otrzymane wyniki przedstawiono m.in. w [254]. Ponieważ algorytm RMatrix generuje większą liczbę reguł, większy był stopień redukcji liczby reguł.

W pracy [249] autor przedstawił jeszcze jedną propozycję filtracji, polegającą na zamianie reguł decyzyjnych w reguły rozmyte i stosowaniu wnioskowania rozmytego

(konstruktywny model Mamdamiego [186]) do rozstrzygania konfliktów klasyfikacji. Zamiarem autora było dalsze ograniczenie liczby reguł bez utraty możliwości rozpoznawania przykładów testowych. Zaproponowano dwa sposoby rozmywania: arbitralny [65] i za pomocą algorytmu genetycznego [249]. Idea rozmywania polegała na zamianie warunków elementarnych, zbudowanych na podstawie atrybutów ciągłych, w warunki rozmyte. Każdy warunek zbudowany na podstawie atrybutu ciągłego może zostać zapisany jako  $a \in [v_1^a, v_2^a]$ , a ten z kolei – jako  $a \in [v_1^a, v_{11}^a, v_{21}^a, v_2^a]$ , gdzie  $v_1^a = v_{11}^a$ ,  $v_2^a = v_{21}^a$ . Taki zapis jest podstawą do zdefiniowania rozmytej postaci warunku  $a \in [v_1^a, v_{11}^a, v_{21}^a, v_2^a]$ , w której  $v_1^a \leq v_{11}^a$ ,  $v_{11}^a \leq v_{21}^a$  oraz  $v_{21}^a \leq v_2^a$ . Do nadzorowania procesu rozmywania wykorzystano algorytm genetyczny. Każda reguła kodowana była jako ciąg warunków elementarnych, a zakodowany zbiór reguł był osobnikiem w populacji. Reguły przekazywane do rozmywania wybierane były arbitralnie przez użytkownika. Kryterium jakości osobnika była dokładność klasyfikacji rozmytego zbioru reguł. Efektywność algorytmu zbadano na kilku zbiorach danych [249]. Algorytm zastosowano także do redukcji liczby reguł w systemie diagnostyki pomp głębinowych, pracujących w podziemnych stacjach odwadniania kopalń węgla kamiennego [252]. Ze względu na czas działania oraz konieczność arbitralnego wyboru reguł poddawanych rozmywaniu algorytm nie nadaje się do filtracji większych zbiorów reguł. Podobnie jak *Inclusion*, *Coverage*, *Forward* i *Backwards*, algorytm filtracji przez selekcję i rozmycie reguł może być stosowany do dowolnego zbioru reguł decyzyjnych. Warunki elementarne, zbudowane na podstawie atrybutów symbolicznych lub porządkowych, nie są rozmywane. Pokrywanie przykładu przez taki warunek rozpatrywane jest w sposób binarny – warunek pokrywa przykład lub go nie pokrywa.

## **6. PRZYKŁADY ZASTOSOWAŃ ALGORYTMÓW INDUKCJI REGUŁ DECYZYJNYCH**

W literaturze przedmiotu opisano dużo przykładów praktycznych zastosowań algorytmów indukcji reguł decyzyjnych. Wielu autorów definiuje indukcję reguł jako proces iteracyjny, w którym podczas rozwiązywania konkretnego zadania testowane są różne zbiorы atrybutów i przykładów oraz różne metody tworzenia i przycinania reguł. Postępowanie takie dotyczy nie tylko budowy systemów regułowych, ale również dowolnej metody analitycznej, tworzącej model danych na podstawie zbioru przykładów. Wszystkie główne metodyki eksploracji danych (np. SEMMA, CRISP-DM) zalecają, aby w fazie modelowania testować efektywność różnych konfiguracji metod analitycznych i na tej postawie dokonywać wyboru metody najbardziej adekwatnej dla rozważanego problemu.

Czy eksploracja danych jest zatem zadaniem czysto inżynierskim? Odpowiedź na to pytanie uzależniona jest od rodzaju analizowanych danych oraz od dziedziny zastosowania. Dla wielu problemów proces ten może być z sukcesem realizowany w sposób rutynowy, za pomocą jednego z dostępnych pakietów analitycznych (np. Rapid-Miner, Statistica Data Miner, SAS Enterprise Miner, GhostMiner, Weka, itd.) lub programów zorientowanych dziedzinowo (np. Salford Systems – drzewa decyzji i regresji; MagnumOpus – reguły asocjacyjne; See5 i Cubist – drzewa decyzji i regresji itd.). Dla wielu problemów i obszarów zastosowań najefektywniejsze okazują się jednak rozwiązania dostosowane do specyfiki rozważanego problemu. Często zorientowane problemowo modyfikacje metod analitycznych stanowią interesujące i nietrywialne rozszerzenia metod stosowanych rutynowo.

Opisane w niniejszym rozdziale trzy przykłady zastosowań systemów indukcji reguł obejmują trzy dziedziny: przemysł, medycynę i bioinformatykę. W zastosowaniach tych użyto prezentowanych w poprzednich rozdziałach metod indukcji, oceny i przycinania zbiorów reguł.

Pierwszy przykład opisuje zastosowanie klasyfikatora regułowego do prognozowania zagrożeń naturalnych [256, 258]. Prognozowanie takich zagrożeń, pojawiających się w związku z prowadzoną przez człowieka działalnością wydobywczą, jest nowym obszarem zastosowań systemów uczących się. W opisie rozwiązania przedstawiono metodykę:

przygotowania zbioru przykładów, weryfikacji wiarygodności systemu oraz jego wdrażania do praktycznego działania.

Drugi przykład związany jest z wyjaśnianiem przyczyn powodzenia/niepowodzenia procedur medycznych [268]. W trakcie badań konieczne było zmodyfikowanie algorytmu indukcji w taki sposób, aby umożliwiał on weryfikację hipotez-reguł definiowanych przez eksperta-lekarza. Dla grup pacjentów opisywanych przez najlepsze z wyznaczonych reguł przeprowadzano analizę przeżycia. W ten sposób połączono reguły decyzyjne z krzywymi przeżycia, definiując tzw. wzorce przeżycia.

Trzeci przykład prezentuje zastosowanie systemów regułowych w bioinformatyce. Zastosowanie to wymagało dostosowania algorytmu indukcji do charakteru analizowanych danych oraz określania miar jakości (w tym miar subiektywnych), pozwalających na wybór reguł najbardziej interesujących. Badania nad konstrukcją regułowych opisów grup genów doprowadziły także do przedstawienia metody redukcji atrybutów, w której pod uwagę brana jest semantyka wartości atrybutów.

Poza wymienionymi przykładami autor stosował systemy regułowe w takich obszarach, jak: diagnostyka maszyn (diagnostyka pomp głębinowych w kopalniach [245, 252]), prognozowanie zagrożeń gazowych (prognozowanie stężenia metanu i tlenku węgla w podziemiach kopalń [250, 259, 263]), prognozowanie stężenia dwutlenku węgla w stacjach odwadniania kopalń [247, 263], aproksymacja masy materiału przesuwającego się na taśmociągu [260]. W części z tych badań stosowano nie tylko reguły decyzyjne, ale także reguły regresyjne oraz metody hybrydowe, łączące indukcję reguł z innymi metodami analitycznymi.

## 6.1. Prognozowanie zagrożeń sejsmicznych w kopalniach

Jednym z głównych zadań stacji geofizycznych w kopalniach węgla kamiennego jest ustalenie aktualnego stanu zagrożenia sejsmicznego (utożsamianego z zagrożeniem wystąpienia destrukcyjnego wstrząsu wysokoenergetycznego, zwanego tapnięciem).

Na przestrzeni ostatnich lat do prognozowania zagrożeń sejsmicznych próbowało stosować metody inteligencji obliczeniowej. Do identyfikacji stref o zwiększym zagrożeniu sejsmicznym stosowano techniki grupowania danych [180, 320], a do prognozowania wstrząsów – sztuczne sieci neuronowe [150, 236]. W przeważającej większości metody te jako wynik prognozy prezentują jeden z dwóch stanów interpretowanych jako *jest zagrożenie* oraz *brak zagrożenia*.

W polskim górnictwie węgla kamiennego dla oceny i prognozy stanu zagrożenia sejsmicznego stosowane są rutynowo dwie metody: sejsmiczna i sejsmoakustyczna. Istotą metody sejsmicznej jest rejestracja i analiza wstrząsów górotworu, a fizyczną podstawą jej stosowania jest występowanie związku między wstrząsami a tapniami [13]. Istotą metody

sejsmoakustycznej jest rejestracja i analiza zjawisk sejsmoakustycznych w rejonach prowadzonych robót górniczych, a fizyczną podstawą jej stosowania jest występowanie związku między tymi zjawiskami (emisją energii sejsmoakustycznej) a zagrożeniem sejsmicznym i tapaniami [13].

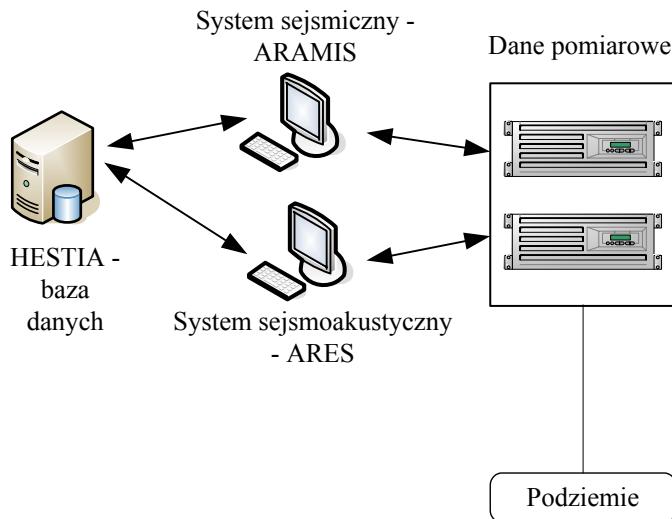
Wynikiem metod sejsmicznej i sejsmoakustycznej jest ocena zagrożenia, wyrażana w skali porządkowej (*brak zagrożenia – ocena a, słabe zagrożenie – b, silne zagrożenie – c, stan niebezpieczny – d*). Horyzont prognozy wynosi osiem godzin i obejmuje najbliższą zmianę wydobywczą. Metoda sejsmoakustyczna pozwala także na ocenę zagrożenia z horyzontem prognozy, wynoszącym jedną godzinę.

Z punktu widzenia dziedziny zastosowań ważną cechą obu metod jest to, że bazują one na danych pomiarowych, gromadzonych rutynowo przez kopalniane stacje geofizyki górniczej. W chwili obecnej w przeważającej większości polskich kopalń węgla kamiennego dane pomiarowe, pobierane z podziemi, transmitowane są na powierzchnię i zapisywane w bazie danych systemu do oceny zagrożeń sejsmicznych Hestia [256]. Hestia jest zaawansowanym systemem, umożliwiającym ocenę (według metod: sejsmicznej i sejsmoakustycznej) i wizualizację zjawisk i zagrożeń sejsmicznych. Pierwsza wersja systemu została opracowana przez autora, kolejne wersje autor rozwija wspólnie z zespołem naukowców i programistów z Instytutu Technik Innowacyjnych EMAG w Katowicach. System Hestia poza Polską z powodzeniem wdrożono w Chinach, Rosji i na Ukrainie.

W pracach [256, 258, 262] autor badał możliwości poprawy prognoz zagrożenia sejsmicznego przez utworzenie regułowego systemu klasyfikacji zagrożeń. Klasyfikator tworzony jest na podstawie analizy przykładów pochodzących z systemu Hestia.

### **6.1.1. Definicja zadania prognozy zagrożenia sejsmicznego**

Ocena zagrożenia sejsmicznego odbywa się oddzielnie dla każdego wyrobiska (np. ściany wydobywczej). Podstawowymi urządzeniami dokonującymi akwizycji danych pomiarowych są geofony i sejsmometry. Dane pochodzące z geofonów przetwarzane są przez sejsmoakustyczny system Ares, a dane pochodzące z sejsmometrów – przez mikrosejsmiczny system Aramis. Dane pochodzące z systemu Ares są w znaczącym stopniu zanieczyszczone, gdyż geofony poza aktywnością górotworu rejestrują także emisję związaną z prowadzeniem prac wydobywczych. Dane pomiarowe oraz ich przetworzona postać (m.in. lokalizacje epicentrów, energie wstrząsów itd.) przekazywane są z systemów Ares i Aramis do systemu Hestia, gdzie na ich podstawie uruchamiane są procedury oceny zagrożenia, zgodnie z metodami sejsmiczną i sejsmoakustyczną. Schemat przepływu danych przedstawiono na rysunku 6.1.



Rys. 6.1. Typowa infrastruktura stacji geofizyki górniczej  
Fig. 6.1. Typical infrastructure of a geophysical station in a coal mine

W zadaniu prognozowania zagrożenia sejsmicznego obowiązują dwa horyzonty prognozy: zmianowy (ośmiogodzinny) i godzinowy. W praktyce górniczej preferowany jest horyzont zmianowy, jako zgodny z cyklem prowadzenia robót wydobywczych. Ponieważ akwizycja danych pomiarowych dokonuje się w sposób ciągły, przed analizą dane poddawane są agregacji. Rozwiązujeając problem prognozowania zagrożeń sejsmicznych, wykorzystano następujące informacje, które w tablicy decyzyjnej pełniły rolę atrybutów warunkowych:

- zmianowe oceny zagrożenia, wyznaczane metodami sejsmiczną i sejsmoakustyczną;
- informację o tym, czy zmiana jest wydobywcza czy niewydobywcza;
- maksymalną sumaryczną energię rejestrowaną w czasie agregacji przez geofony monitorujące wyrobisko (dla ułatwienia dalszego zapisu przyjmijmy, że geofon rejestrujący maksymalną energię oznaczymy przez GMax);
- maksymalną sumaryczną liczbę impulsów, rejestrowaną w czasie agregacji przez GMax;
- odchyłkę sumarycznej energii, rejestrowanej przez GMax, od średniej energii, obliczonej dla ośmiu wcześniejszych okresów agregacji danych;
- odchyłkę liczby impulsów rejestrowanych przez GMax od średniej liczby impulsów, obliczonej dla ośmiu wcześniejszych okresów agregacji danych;
- ocenę zagrożenia, wyznaczoną dla GMax za pomocą metody sejsmoakustycznej;
- liczbę zjawisk sejsmicznych, zarejestrowanych w czasie agregacji;
- sumaryczną energię zarejestrowanych w czasie agregacji zjawisk sejsmicznych;
- maksymalną energię zjawiska sejsmicznego pośród zjawisk zarejestrowanych w czasie agregacji.

Jeśli do wyrobiska przyporządkowany był więcej niż jeden geofon, to w zbiorze zmiennych pojawiały się średnie wartości energii, impulsów oraz odchyłek.

Głównym celem prognozy jest przewidywanie, z dokładnością ustaloną co do miejsca i czasu, podwyższonej aktywności sejsmicznej, która może prowadzić do wystąpienia tapnięcia. Wobec tego zdefiniowano trzy zadania prognozy:

- zadanie 1 – przewidywanie sytuacji, w których suma energii sejsmicznej (zarejestrowanych wstrząsów) i energii rejestrowanej przez geofon o maksymalnej aktywności przekroczy, w ciągu najbliższych ośmiu godzin, wartość  $5 \cdot 10^5$  J;
- zadanie 2 – przewidywanie sytuacji, w których suma energii sejsmicznej (zarejestrowanych wstrząsów) i energii rejestrowanej przez geofon o maksymalnej aktywności przekroczy, w ciągu najbliższej godziny, wartość  $4.8 \cdot 10^4$  J;
- zadanie 3 – przewidywanie, że w ciągu najbliższych ośmiu godzin zostanie zarejestrowany wstrząs sejsmiczny o energii większej od  $10^4$  J.

Łatwo zauważyc, że tak zdefiniowane zadania prognozy zagrożenia można potraktować jako zadania klasyfikacji z dwiema klasami decyzyjnymi: *jest zagrożenie* oraz *brak zagrożenia*. Zdefiniowane zadania prognozy nie koncentrują się na przewidywaniu czasu wystąpienia tapnięcia, gdyż – jak twierdzi wielu badaczy – niemożliwe jest rozwiązanie tego zadania z dobrą dokładnością, aczkolwiek przekroczenie wartości progowych, zdefiniowanych w zadaniach 1, 2, 3, uznawane jest za sytuację niebezpieczną. Przewidywanie stanów podwyższonej aktywności sejsmicznej ma również duże znaczenie praktyczne, gdyż informacja o możliwości zaistnienia sytuacji niebezpiecznej uruchamia odpowiednie działania (np. wykonanie strzałań przeciwstrzałowych), zmierzające do zmniejszenia zagrożenia lub wycofania załogi z zagrożonego rejonu.

Przygotowując dane eksperymentalne, będące podstawą do trenowania klasyfikatorów, agregację danych wykonano w przesuwających się, następujących po sobie oknach czasowych. Przesunięcie (czas agregacji) wynosiło odpowiednio jedną godzinę i osiem godzin. Ponieważ ocenę metodą sejsmiczną wykonuje się raz na osiem godzin, w przypadku agregacji godzinowej informacje o wynikach tej oceny zmieniały się co osiem rekordów. Otrzymane tablice decyzyjne były tablicami temporalnymi. Na ich podstawie możliwe było zdefiniowanie opóźnień oraz atrybutów dynamiki. Jak pokazano w [256], dla rozważanych zbiorów przykładów użycie atrybutów opóźnionych nie poprawiało wyników klasyfikacji.

Podczas badań przeanalizowano dane pochodzące z dwóch ścian wydobywczych, znajdujących się w kopalni węgla kamiennego „Mysłowice-Wesoła”. Charakterystykę zbiorów danych związanych z zadaniami 1 i 3 zamieszczono w tabeli 6.1. Jak widać, wszystkie zbiory danych charakteryzują się nierównomiernym rozkładem liczby przykładów pozytywnych (*jest zagrożenie*) i negatywnych (*brak zagrożenia*). Z nierównomiernym

rozkładem liczby przykładów spotykamy się również podczas prognozowania zagrożeń metanowych [262] oraz zagrożeń pożarem endogenicznym [172].

Tabela 6.1

## Charakterystyka analizowanych zbiorów danych

Wyrobisko	Liczba przykładów	Przykłady pozytywne	Przykłady negatywne
Prognozowanie energii sumarycznej – zadanie 1			
SC503	1097	188	909
SC508	864	97	767
Prognozowanie wstrząsów wysokoenergetycznych – zadanie 3			
SC503	1097	118	979
SC508	864	52	812

Ze względu na różnice warunków geologicznych dla każdego wyrobiska oraz każdego zadania prognozy tworzono oddzielny klasyfikator. W dalszej części zaprezentowano wyniki uzyskane przez klasyfikatory tworzone w celu rozwiązania zadań 1 i 3. Zadania te mają największe znaczenie dla praktyki górniczej.

### 6.1.2. Budowa i weryfikacja jakości klasyfikatora

Do budowy klasyfikatora użyto algorytmu q-ModLEM. Algorytm ten uruchamiano w różnych konfiguracjach miar jakości.

Kryterium jakości klasyfikatora była miara *SSS* (2.24). Nadrzędnym zadaniem było zatem uzyskanie klasyfikatora o jak najwyższych wrażliwości i specyficzności. Klasyfikator o dobrej wrażliwości i dobrej specyficzności będzie dobrze prognozował stany zagrożenia, jednocześnie minimalizując liczbę tzw. fałszywych alarmów.

W trakcie badań najlepsze klasyfikatory otrzymano, stosując algorytm q-ModLEM w wersji EMM (entropia warunkowa do oceny warunków elementarnych, miara jakości w fazach przycinania i klasyfikacji). W tabeli 6.2 zaprezentowano wyniki najlepszego z klasyfikatorów. W celu znalezienia najlepszego klasyfikatora wykonano eksperymenty dla każdej z 9 najbardziej obiecujących miar jakości, zawartych w zbiorze *Max*. Do utworzonych reguł stosowano algorytmy filtracji. Jeśli filtracja poprawiała wartość miary *SSS*, jako wyjściowy wybierano przefiltrowany zbiór reguł. W tabeli 6.2 informacja o filtracji zamieszczona jest w postaci pierwszej litery nazwy użytego algorytmu (np. F – *Forward*). W tabeli 6.2 prezentowane są wyniki średnie, otrzymane po zastosowaniu 5-krotnej warstwowej walidacji krzyżowej, powtózonej 10 razy. W tabeli użyto standardowych, stosowanych w poprzednich rozdziałach, nazw kolumn.

Tabela 6.2  
Wyniki klasyfikacji zagrożeń sejsmicznych – indukcja i filtracja

Zadanie prognozy/ wyrobisko	Miara/ algorytm	Acc [%]	BAcc [%]	Liczba reguł
Zadanie 1 SC503	C2 (F)	90.4	84.4	11.6
	RIPPER	90.6	84.6	4.4
	PART	88.5	79.2	24.7
	CART	86.8	<b>87.1</b>	-
	SSV	90.6	83.9	-
Zadanie 1 Sc508	Corr (B)	73.8	76.5	23.4
	RIPPER	87.4	60.9	3
	PART	86.0	60.5	27
	CART	76.0	74.7	-
	SSV	85.9	60.0	-
Zadanie 3 SC503	RSS	73.2	<b>66.2</b>	18.0
	RIPPER	88.2	52.0	1.8
	PART	86.4	53.2	22.8
	CART	73.4	65.7	-
	SSV	89.4	49.8	-
Zadanie 3 Sc508	RSS	72.1	62.2	16.8
	RIPPER	96.5	50.0	1.0
	PART	95.6	50.1	16.5
	CART	72.0	60.6	-
	SSV	96.5	50.0	-

W tabeli 6.2 zamieszczono także rezultaty dwóch pokryciowych algorytmów indukcji reguł (RIPPER [57] i PART [333]) oraz dwóch algorytmów indukcji drzew decyzyjnych (SSV [103], CART [38]). Do indukcji drzew CART wykorzystano komercyjną wersję oprogramowania CART firmy Salford Systems, a do indukcji drzew SSV użyto komercyjnej wersji oprogramowania GhostMiner. Ponieważ algorytmy RIPPER, PART, CART, SSV można parametryzować (np. włączając lub wyłączając przycinanie, wybierając szerokość wiązki – SSV, zmieniając kryterium oceny podziału węzła – CART itd.), wykonano wiele eksperymentów dla różnych ustawień tych parametrów. Dobierając wartości parametrów starano się uzyskać jak najwyższą wartość miary SSS. W tabeli 6.2 zamieszczono najlepsze z otrzymanych rezultatów. Analizując zawartość tej tabeli, można wyciągnąć wniosek, że RIPPER, PART oraz SSV w trzecim zadaniu prognozowania dokonują przyporządkowania większości przykładów testowych do klasy decyzyjnej, wskazującej na *brak zagrożenia*. Wyniki takie nie są użyteczne dla konstrukcji systemu prognozowania. Najlepszymi algorytmami okazały się q-ModLEM oraz CART, przy czym nieznacznie lepszy jest q-ModLEM. W szczególności dla 3. zadania prognozy zagrożenia i wyrobiska SC508 różnice

pomiędzy wartościami SSS, uzyskanymi przez te algorytmy, są statystycznie istotne (wykorzystano skorygowaną postać testu t, poziom istotności – 0.05). Wyniki klasyfikacji, zamieszczone w tabeli 6.2 różnią się nieznacznie od przedstawionych w pracach [256, 258], co wynika z przyjętej tutaj bardziej restrykcyjnej metodyki testowania 10x10CV.

W dalszej części badań sprawdzono, czy zastosowanie algorytmów agregacji i redefinicji reguł może poprawić dokładność prognoz. Podstawę do agregacji stanowiły najlepsze (spośród utworzonych) zbiory reguł. Biorąc pod uwagę doświadczenia związane ze stosowaniem algorytmu redefinicji oraz rozmiar analizowanych danych, podczas oceny ważności warunków elementarnych użyto podstawowego indeksu Banzhafa (5.2).

Agregacja reguł (algorytm CHIRA) nie pozwoliła na poprawę jakości prognoz, poprawę taką odnotowano po zastosowaniu algorytmu redefinicji, dopuszczającego do pojawiania się w przesłankach reguł zanegowanych warunków elementarnych. Algorytm ten pozwolił na poprawę jakości klasyfikatora dla zadania 1. i wyrobiska SC508 oraz dla zadania 3. i wyrobiska SC503. Po zastosowaniu algorytmu redefinicji otrzymano klasyfikatory o charakterystyce takiej, jak przedstawia to tabela 6.3.

Tabela 6.3

**Wyniki klasyfikacji zagrożeń sejsmicznych –  
indukcja, filtracja i redefinicja reguł**

Zadanie prognozy/ wyrobisko	Miara/ algorytm	Acc [%]	BAcc [%]	Liczba reguł
Zadanie 1 – SC503	C2 (F)	90.4	84.4	11.6
Zadanie 1 – Sc508	Corr – Redefinicja	74.1	77.3	17.2
Zadanie 3 – SC503	RSS – Redefinicja	72.2	68.2	14.2
Zadanie 3 – SC508	RSS	72.1	62.2	16.8

#### **6.1.3. Analiza otrzymanych wyników**

Aby porównać wyniki klasyfikatora i metody rutynowej, przyjęto, że ocena rutynowa *a* odpowiada ocenie klasyfikatora *brak zagrożenia*, natomiast pozostałe wyniki metody rutynowej (*b*, *c*, *d*) odpowiadają ocenie klasyfikatora, wskazującej na zagrożenie (*jest zagrożenie*). Wyniki porównania zaprezentowano w tabeli 6.4. W tabeli tej zamieszczono także informację o podstawowej dokładności klas decyzyjnych (*a priori*), obliczonej na podstawie liczby przykładów reprezentujących poszczególne klasy decyzyjne. Warto zaznaczyć, że porównań pomiędzy metodą rutynową a klasyfikatorem dokonano również dla innych powiązań pomiędzy ocenami metody rutynowej a klasyfikatorem (np. ocena *a* lub *b* = *brak zagrożenia*; oceny *c*, *d* = *jest zagrożenie*). Otrzymane wyniki były dla metody rutynowej jeszcze mniej korzystne niż te prezentowane w tabeli 6.4.

Tabela 6.4

Porównanie dokładności prognoz –  
klasyfikator regułowy vs metoda rutynowa

Zadanie prognozy/ wyrobisko	Metoda	Stan „jest zagrożenie” [%]	Stan „brak zagrożenia” [%]	SSS
Zadanie 1 SC503	klasyfikator	75.2	93.6	<b>0.69</b>
	rutynowa	56.8	57.0	0.14
	<i>a priori</i>	17.2	82.8	0.00
Zadanie 1 SC508	klasyfikator	81.5	73.1	<b>0.55</b>
	rutynowa	77.0	43.0	0.20
	<i>a priori</i>	11.3	88.7	0.00
Zadanie 3 SC503	klasyfikator	63.2	73.2	<b>0.37</b>
	rutynowa	52.5	82.4	0.35
	<i>a priori</i>	10.8	89.2	0.00
Zadanie 3 SC508	klasyfikator	51.4	72.9	<b>0.29</b>
	rutynowa	48.3	47.8	-0.03
	<i>a priori</i>	6.1	93.9	0.00

Wyniki zamieszczone w tabeli 6.4 jednoznacznie pokazują, że dla celu prognozowania zagrożeń sejsmicznych oceny generowane przez klasyfikator są zdecydowanie dokładniejsze od ocen generowanych przez metodę rutynową. Dopuszczenie klasyfikatora jako uzupełniającej metody oceny zagrożenia sejsmicznego wymaga jednak podjęcia wielu działań formalnych. Między innymi konieczne jest przedstawienie modelu, na podstawie którego dokonuje się oceny zagrożenia.

Pokryciowe algorytmy indukcji reguł nie są algorytmami zbyt stabilnymi. W niektórych przypadkach małe zmiany w zbiorze przykładów treningowych mogą powodować duże zmiany modelu prognostycznego. Zauważmy również, że dla każdego wyrobiska górnictwa tworzony jest oddzielny klasyfikator, gdyż utworzenie jednego uniwersalnego modelu dla dowolnego wyrobiska okazało się niemożliwe [256].

Pomimo że nie można poddać weryfikacji jednego uniwersalnego modelu klasyfikacji zagrożeń, można próbować odpowiedzieć na następujące pytania:

- Które atrybuty warunkowe pojawiają się najczęściej w przesłankach wyznaczonych reguł?
- Czy możliwe jest wskazanie warunków elementarnych, mających największy wpływ na prognozę zagrożenia?
- Jaka jest jakość wyznaczonych reguł i czy wśród nich znajdują się reguły sprzeczne z wiedzą dziedzinową?

Zauważmy, że wybór optymalnych parametrów klasyfikatora następuje na podstawie wyników walidacji krzyżowej. Klasyfikator stosowany w praktyce powstaje jednak

na podstawie całego dostępnego zbioru przykładów. Ponieważ w zależności od wyrobiska i rodzaju zadania prognozy optymalne klasyfikatory otrzymano dla różnych miar jakości ( $C_2$ ,  $Corr$ ,  $RSS$ ), przeprowadzono analizę zbiorów reguł, utworzonych za pomocą tych miar.

Najczęściej występującymi atrybutami w przesłankach analizowanych reguł były:

- energia rejestrowana przez geofon GMax;
- liczba impulsów, rejestrowana przez GMax;
- całkowita energia wstrząsów rejestrowanych w okresie ostatnich 8 godzin;
- wartość odchyłki liczby impulsów, rejestrowanych przez GMax, od średniej liczby impulsów, obliczonej dla 8 wcześniejszych okresów agregacji danych;
- wartość odchyłki sumarycznej energii, rejestrowanej przez GMax, od średniej energii, obliczonej dla 8 wcześniejszych okresów agregacji danych;
- maksymalna energia rejestrowanych wstrząsów;
- liczba wstrząsów zarejestrowanych w klasach energetycznych  $10^2 J$ ,  $10^3 J$ ,  $10^4 J$ ;
- informacja o tym, czy zmiana jest wydobywcza czy niewydobywcza.

Zauważmy, że w wyznaczonych regułach nie występują atrybuty informujące o wartościach ocen generowanych przez metody rutynowe (sejsmiczną i sejsmoakustyczną).

Analiza ważności warunków elementarnych była trudna, gdyż w zależności od analizowanego zbioru treningowego oraz użytej miary jakości otrzymywano nieco odmienne wyniki. Z punktu widzenia klasy decyzyjnej *jest zagrożenie* najważniejsze były warunki elementarne  $w_1$  i  $w_2$ , które można zinterpretować jako:  $w_1$  – *energia rejestrowana przez geofon GMax jest duża* ( $> 3.5 \cdot 10^5 J$ ),  $w_2$  – *liczba impulsów rejestrowanych przez geofon GMax jest duża*. Dla klasy decyzyjnej *brak zagrożenia* najważniejsze były warunki elementarne, które można interpretować jako negację  $w_1$  i  $w_2$ .

Dokładność reguł wskazujących na klasę decyzyjną *jest zagrożenie* była średnio o 20% niższa niż dokładność reguł wskazujących na klasę *brak zagrożenia*. Reguły wskazujących na klasę decyzyjną *brak zagrożenia* było też zdecydowanie więcej. Różna liczba reguł opisujących klasy decyzyjne ma najczęściej negatywny wpływ na dokładność klasyfikacji klasy mniejszościowej [124]. W rozważanym przypadku (podobnie zresztą jak podczas klasyfikacji danych benchmarkowych) problem ten częściowo rozwiązywany jest przez dobór odpowiedniej miary jakości, nadzorującej proces indukcji reguł i decydującej o tym, jak rozwiązywane są konflikty klasyfikacji. Zauważmy, że najlepsze klasyfikatory otrzymano dla miar biorących pod uwagę rozmiar klas decyzyjnych.

Poniżej przytoczono przykłady czterech reguł opisujących część z przykładów treningowych, definiujących 1. zadanie prognozy zagrożenia. Reguły te są dosyć stabilne, gdyż pojawiały się w zbiorach indukowanych przez różne miary jakości ( $RSS$ ,  $Corr$ ).

### Zadanie 1 – SC503

R1

**Jeżeli** energia rejestrowana przez GMax w czasie ostatniej zmiany  $\leq 1.2 \cdot 10^5$  J **oraz**  
liczba impulsów rejestrowana przez GMax w czasie ostatniej zmiany  $\in [2986, 6932]$ ,  
**oraz** sumaryczna energia zjawisk sejsmicznych w czasie ostatniej zmiany  $< 5.7 \cdot 10^3$  J,  
**to brak zagrożenia** (precision: 1.00).

R2

**Jeżeli** energia rejestrowana przez GMax w czasie ostatniej zmiany  $\geq 2.8 \cdot 10^5$  J,  
**to jest zagrożenie** (precision: 0.72).

### Zadanie 1 – SC508

R3

**Jeżeli** średnia liczba impulsów rejestrowanych w czasie ostatniej zmiany przez geofony  
przyporządkowane do wyrobiska  $< 1712$  **oraz** sumaryczna energia zjawisk sejsmicznych  
rejestrowanych w czasie ostatniej zmiany  $< 4.6 \cdot 10^3$  J,  
**to brak zagrożenia** (precision: 0.94).

R4

**Jeżeli** średnia energia rejestrowana przez geofony przyporządkowane do wyrobiska  $>$   
 $1.8 \cdot 10^5$  J **oraz** średnia odchyłka energii sejsmicznej rejestrowanej przez geofony  $> 345\%$ ,  
**to jest zagrożenie** (precision: 0.63).

Dokładność ostatniej reguły nie wydaje się zbyt duża. Pamiętajmy jednak, że liczba przykładów pozytywnych dla 1. zadania prognozy i wyrobiska SC508 stanowi niecałe 10% wszystkich przykładów.

#### **6.1.4. Zagadnienia związane z wdrażaniem**

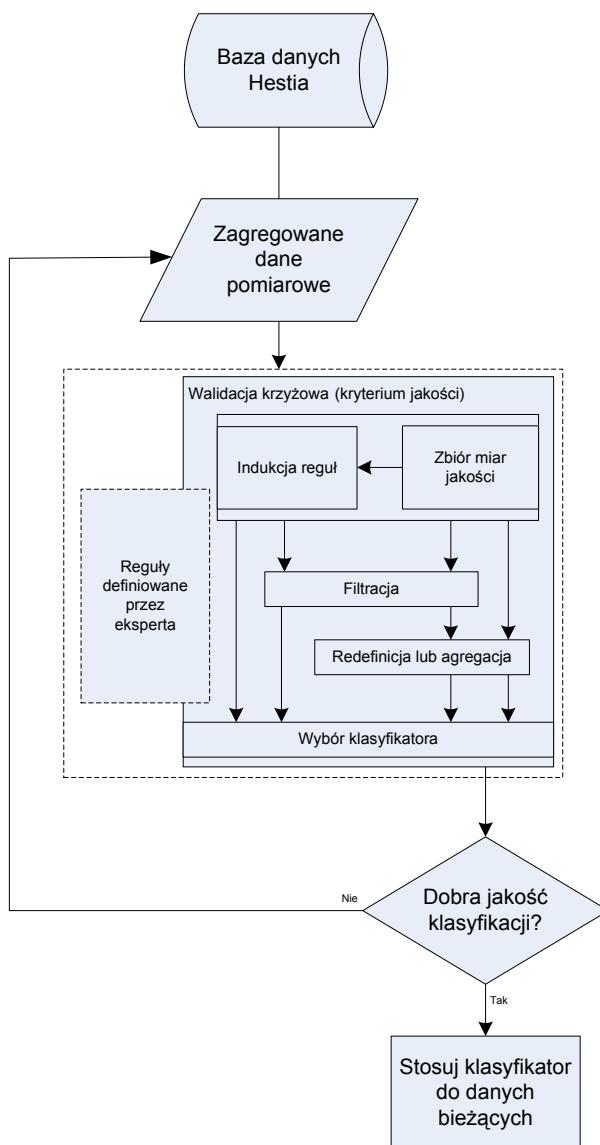
Rozważając wdrożeniowe aspekty systemu prognozującego, należy mieć na uwadze to, że dla każdego wyrobiska górnictwego konieczne jest utworzenie oddzielnego klasyfikatora. Oznacza to, że w początkowym okresie eksploatacji jedynym zadaniem systemu prognostycznego jest gromadzenie przykładów, które będą stanowić podstawę do utworzenia klasyfikatora. Dane treningowe, będące podstawą do przeprowadzenia opisanych tutaj eksperymentów, pochodziły z okresu stanowiącego około 30% całego czasu eksploatacji wyrobiska. Wyniki klasyfikatorów utworzonych na podstawie mniejszych zbiorów treningowych były nieznacznie gorsze [256, 258].

Wykorzystując rezultaty przeprowadzonych eksperymentów, można zaproponować następujący schemat tworzenia i wdrażania klasyfikatora regułowego do prognozowania zagrożeń sejsmicznych (rys. 6.2). Klasyfikator jest podstawowym składnikiem systemu prognostycznego. System dokonuje akwizycji i agregacji danych pomiarowych, które następnie stanowią podstawę do trenowania klasyfikatora. Pomijając kwestie implementacyjne, możemy powiedzieć, że proces trenowania klasyfikatora uruchamiany jest

cyklicznie na zwiększającym się zbiorze przykładów i jest powtarzany dopóty, dopóki otrzymany klasyfikator nie uzyskuje wyników lepszych od zadanych wartości minimalnych. W szczególności prognozy klasyfikatora muszą być lepsze od prognoz generowanych za pomocą metod stosowanych rutynowo. Klasyfikator spełniający wymagania minimalnej jakości stosowany jest następnie do napływających danych pomiarowych, cały czas monitorowana jest także jakość generowanych prognoz (rys. 6.2).

Proces tworzenia klasyfikatora jest wieloetapowy i polega na trenowaniu różnych klasyfikatorów, sprawdzaniu ich efektywności w drodze walidacji krzyżowej i wyborze tego, który osiąga najwyższe wartości ustalonego kryterium jakości. Dokładniej, w trybie walidacji krzyżowej ustalane są najbardziej odpowiednie wartości parametrów klasyfikatora. Parametry te są następnie użyte podczas trenowania klasyfikatora na całym dostępnym zbiorze treningowym.

Opisany schemat postępowania można adaptować do dowolnej dziedziny zastosowań i dowolnej metody konstrukcji klasyfikatora. W zależności od dziedziny zastosowania będziemy mieli do czynienia z różnymi miarami efektywności klasyfikatora oraz z różnymi procedurami kontroli jakości generowanych prognoz. W pracach [250, 259, 263] autor wraz ze współpracownikami zajmował się prognozowaniem stężenia metanu w wyrobiskach górniczych. Zadanie prognozowania stężenia metanu jest problemem z pogranicza analizy regresji i prognozowania szeregów czasowych, dlatego w tworzonych systemach prognostycznych stosowano m.in. reguły regresyjne [239, 333]. Utworzenie systemu prognostycznego wymaga rozwiązania takich samych problemów, jakie występują przy tworzeniu systemu klasyfikacyjnego. Konieczne jest: zebranie odpowiedniego zbioru danych treningowych, opracowanie procedury trenowania i dobór optymalnych wartości parametrów algorytmu prognozowania, określenie minimalnej akceptowalnej jakości prognoz, zdefiniowanie procedury nadzorującej jakość prognoz, wreszcie – zdefiniowanie procedury wyboru przykładów treningowych w przypadku konieczności ponownego trenowania systemu.



Rys. 6.2. Metodyka tworzenia systemu prognozującego zagrożenia sejsmiczne

Fig. 6.2. Methodics of defining a seismic hazard forecasting system

Z koniecznością ponownego trenowania systemu autor zetknął się głównie podczas prognozowania stężenia metanu. W celu wyboru nowego zbioru przykładów treningowych stosowano wtedy prostą procedurę, polegającą na rozszerzeniu istniejącego zbioru treningowego o najnowsze dane pomiarowe. Jeśli rozmiar zbioru przykładów był zbyt duży, to usuwano z niego najstarsze rekordy. Chociaż badań takich nie prowadzono, wydaje się, że w przypadku zagrożeń sejsmicznych i ewentualnej konieczności ponownego trenowania klasyfikatora sensowne byłoby usuwanie jedynie najstarszych przykładów reprezentujących większość klasę decyzyjną *brak zagrożenia*. W dalszych pracach interesujące będzie wykorzystanie inkrementacyjnych wersji algorytmów indukcji reguł [242,316,335] i metod radzących sobie z (nierozważanym w niniejszej monografii) problemem zmieniającego się opisu klasy docelowej (ang. *concept drift*) [311].

Do reguł utworzonych w sposób automatyczny mogą być dołączane reguły definiowane przez eksperta dziedzinowego. Reguły te mogą reprezentować pewne zdroworozsądkowe zależności lub mogą być odzwierciedleniem wiedzy górniczej. Siłę reguł definiowanych przez eksperta można określić podobnie jak siłę reguł utworzonych automatycznie (tzn. na podstawie tablicy kontyngencji i zbioru treningowego). Inny sposób wykorzystania wiedzy eksperta dziedzinowego może polegać na tym, iż będzie on modyfikował reguły wyznaczone automatycznie.

Do rozwiązania zadania prognozowania zagrożeń sejsmicznych stosowano także algorytm RMatrix oraz aparat analityczny zbiorów przybliżonych (wskaźniki umożliwiające ocenę ważności atrybutów, redukty względne tablicy decyzyjnej) [262]. Uzyskane wyniki są spójne z wynikami prezentowanymi w niniejszym rozdziale.

## 6.2. Prognozowanie czasu przeżycia pacjentów po przeszczepie szpiku kostnego

Jednym z ważnych celów analizy danych medycznych jest poszukiwanie w nich prawidłowości (wzorców), które mogą stanowić podstawę do zrozumienia lub ulepszania stosowanych procedur diagnostycznych i terapeutycznych. Szczególnym przypadkiem analizy danych jest analiza przeżycia, polegająca m.in. na identyfikacji czynników mających największy wpływ na czas przeżycia pacjentów po przeprowadzeniu u nich określonych procedur terapeutycznych. W przeważającej większości przypadków w środowisku medycznym do analizy przeżycia stosowane są metody statystyczne (analiza korelacji, estymator Kaplana-Meiera [155], model proporcjonalnego hazardu Coxa [57]). W ostatnich latach analiza przeżycia wspomagana jest również metodami maszynowego uczenia się [347], w tym – metodami indukcji drzew decyzyjnych [34].

W badaniach opisanych w dalszej części rozdziału do wspomagania analizy przeżycia stosowano pokryciowy algorytm indukcji reguł decyzyjnych. Wyznaczone reguły po odpowiedniej selekcji zamieniane są w tzw. wzorce przeżycia. W rozdziale przedstawiono także modyfikację algorytmu indukcji reguł, pozwalającą weryfikować hipotezy badawcze, definiowane przez użytkownika. Hipotezy te reprezentują zależności, których prawdziwość użytkownik (tutaj lekarz) chce zweryfikować w świetle dostępnych danych.

Przedstawione propozycje motywowane są badaniami nad identyfikacją czynników wpływających na czas przeżycia pacjentów po przeszczepie szpiku kostnego. Badania te prowadzone były przez autora wraz ze współpracownikami (w tym lekarzami z Kliniki Transplantacji Szpiku, Onkologii i Hematologii Dziecięcej we Wrocławiu).

Główna zasada działania algorytmów indukcji reguł jest taka, że w sposób automatyczny dokonują one wyboru warunków elementarnych, z jakich złożone są przesłanki reguł. Reguły utworzone w ten sposób nie zawsze są interesujące i użyteczne dla użytkownika.

W szczególności w kontekście omawianej w podrozdziale 3.1.9 miary interwencji dobrze byłoby móc ukierunkować algorytm na indukcję takich reguł, które będą stanowić wskazówki dla podjęcia skutecznych (i możliwych) działań.

Do chwili obecnej w stosunkowo niewielu pracach opisano, w jaki sposób wymusić, aby algorytm indukcji brał pod uwagę preferencje użytkownika, dotyczące budowy reguł. Stefanowski [284] przedstawia interaktywną wersję algorytmu Explore, w której użytkownik może określić wymagania dotyczące atrybutów i/lub ich wartości, pojawiających się w przesłankach reguł. Ponieważ algorytm przetwarza jedynie atrybuty symboliczne (atrybuty ciągle należą poddać dyskretyzacji), umożliwia to również wskazanie warunków elementarnych, jakie mają się znaleźć w przesłankach wyznaczonych reguł. W pracach nad indukcją reguł asocjacyjnych przedstawiono przykłady interaktywnej budowy reguł [231] oraz indukcji tzw. reguł niespodziewanych (ang. *unexpected*). Reguły niespodziewane tworzone są na podstawie analizy definiowanych przez użytkownika szablonów, wskazujących, z jakich atrybutów składają się reguły typowe [218]. Gamberger i Lavrac [94] przedstawiają podobną propozycję dla algorytmu indukcji reguł decyzyjnych, przeznaczonych dla celów opisowych. Algorytmy stosujące paradygmat uczenia się na podstawie argumentacji [200, 206] pozwalają użytkownikowi na umieszczanie obok każdego przykładu wyjaśnienia informującego o tym, dlaczego został on zaklasyfikowany do tej, a nie innej klasy decyzyjnej. Przykłady zastosowań medycznych pokazują, że takie postępowanie może znacząco ograniczyć zbiór generowanych reguł. Metoda uczenia na podstawie argumentacji nie pozwala jednak na weryfikację hipotez reprezentujących zależności, jakie zdaniem użytkownika występują w danych. Częściowo możliwość taką zaprezentowano w pracy [49], gdzie użytkownik definiuje zbiór reguł, które spodziewa się znaleźć w analizowanym zbiorze danych. Następnie uruchamiana jest regułowa wersja algorytmu C4.5 i generowane są trzy typy reguł: zgodne z regułami wprowadzonymi przez użytkownika, niezwiązane z regułami użytkownika, niezgodne z wiedzą użytkownika. Regułę  $r$  uznaje się za zgodną z wiedzą użytkownika, jeśli w zbiorze zdefiniowanych przez niego reguł istnieje co najmniej jedna reguła  $e$ , taka że  $r$  i  $e$  wskazują na identyczną klasę decyzyjną oraz zbiór przykładów pokrywanych przez  $r$  jest podzbiorem przykładów pokrywanych przez  $e$ . Warto zauważyć, że w przedstawianym podejściu reguły zgodne z wiedzą eksperta nie muszą być zbudowane z atrybutów, które są dla niego interesujące. Omawiany algorytm bada jedynie semantyczne podobieństwo reguł, czyli sprawdza, czy reguły pokrywają te same zbiory przykładów.

W praktycznych zastosowaniach medycznych metody indukcji reguł decyzyjnych stosowano najczęściej do rozwiązywania problemów diagnostycznych (np. w systemach diagnozujących choroby głowy [312, 315], encefalopatię mitochondrialną u dzieci [322], choroby wątroby [318], choroby reumatyczne [72] itd.). Indukcja reguł stanowi także

uzupełnienie dla analizy danych medycznych, przeprowadzanej za pomocą metod teorii przybliżonych (np. [291, 278]) lub obliczeń granularnych (np. [173]).

Indukcja reguł stanowi także podstawową metodę analityczną serwisu internetowego, przeznaczonego do analizy danych medycznych [50]. W [60] Daud i Crone dokonali przeglądu efektywności różnych algorytmów indukcji reguł, stosowanych do analizy danych medycznych. Efektywność rozumiana była jako całkowita dokładność klasyfikacji i liczba wyznaczanych reguł. Podobną analizę przeprowadzają Ilczuk i Wakulicz-Deja [138], analizując wpływ redukcji atrybutów na wyniki: algorytmu LEM2 oraz algorytmów indukcji drzew i reguł decyzyjnych, zawartych w pakiecie Weka.

Stosunkowo niewielka liczba publikacji koncentruje się na powiązaniu analizy przeżycia i metod indukcji reguł [18, 219]. W [219] Pattaraintakorn i Cercone zastosowali metody teorii zbiorów przybliżonych do identyfikacji głównych czynników mających wpływ na czas przeżycia pacjentów, przy czym czas ten rozpatrywany jest jako zmienna dyskretna (np. *czas przeżycia* ∈ [56, 73] miesięcy). Bazan i inni [18] do indukcji reguł przeżycia również stosują metody teorii zbiorów przybliżonych. W celu określenia zakresów klas decyzyjnych autorzy ci definiują współczynnik *PI* (ang. *prognostic index coefficient*), którego wartość obliczana jest na podstawie modelu proporcjonalnego hazardu Coxa. Zakres wartości *PI* dzielony jest w taki sposób, aby krzywe przeżycia, wyznaczane dla obserwacji należących do różnych elementów tego podziału, były statystycznie różne. W [167] Kronek i Reddy przedstawili algorytm LASD, umożliwiający indukcję reguł na podstawie danych informujących o przeżyciu. Na podstawie analizy dwóch benchmarkowych zbiorów danych twierdzą oni, że algorytm ten uzyskuje lepsze rezultaty niż algorytm indukcji drzew przeżycia [34].

### **6.2.1. Pokryciowy algorytm indukcji reguł, sterowany hipotezami definiowanymi przez użytkownika**

W standardowym algorytmie indukcji reguł użytkownik nie ma wpływu na postać tworzonych warunków elementarnych. Pokryciowy algorytm indukcji reguł można jednak w stosunku prosty sposób zmodyfikować, tak aby pod uwagę brał on preferencje użytkownika. Preferencje te najczęściej wyrażane są przez:

- wymuszanie (zabranianie) pojawiania się konkretnych atrybutów w przesłankach reguł,
- wymuszanie (zabranianie) pojawiania się konkretnych warunków elementarnych w przesłankach reguł,
- wymuszanie (zabranianie) pojawiania się konkretnych reguł w opisach klas decyzyjnych.

Preferencje dotyczące atrybutów i warunków elementarnych mogą być definiowane oddzielnie dla każdej klasy decyzyjnej. Przedstawiona dalej modyfikacja pokryciowego algorytmu indukcji reguł pozwala na wzięcie pod uwagę wszystkich trzech typów preferencji.

W zależności od tego, jakiego rodzaju preferencje zostaną zdefiniowane, mamy do czynienia z różnymi trybami działania algorytmu.

O1. Użytkownik wprowadza zbiór gotowych reguł decyzyjnych:

- a) algorytm traktuje te reguły jako gotowy klasyfikator; algorytm nie ingeruje w budowę reguł ani nie dodaje do nich nowych reguł; efektywność klasyfikatora może być weryfikowana na całym dostępnym zbiorze przykładów lub w trybie walidacji krzyżowej; poza obliczeniem miar oceny jakości klasyfikatora obliczane są także wybrane miary jakości reguł (np. statystyczna istotność reguł);
- b) algorytm dokonuje redefinicji reguł wprowadzonych przez użytkownika; redefinicję rozumiemy jako dodanie do przesłanek nowych warunków elementarnych (faza wzrostu) i/lub usunięcie warunków istniejących (faza przycinania); redefinicja przeprowadzana jest po to, aby zwiększyć jakość reguł zdefiniowanych przez użytkownika; po zakończeniu redefinicji badane są: efektywność klasyfikatora i jakość reguł;
- c) algorytm dokonuje redefinicji reguł wprowadzonych przez użytkownika w sposób identyczny z opisany w poprzednim punkcie; po zakończeniu redefinicji generowane są kolejne reguły, tak aby pokryć zbiór przykładów treningowych; po zakończeniu indukcji badane są: efektywność klasyfikatora i jakość reguł.

O2. Użytkownik definiuje dla każdej klasy decyzyjnej warunki elementarne, jakie muszą pojawić się w co najmniej jednej z reguł opisujących daną klasę decyzyjną. Dopuszczalne jest definiowanie pojedynczego warunku elementarnego lub ich koniunkcji.

O3. Użytkownik wskazuje atrybuty, które muszą się pojawić w co najmniej jednej regule opisującej daną klasę decyzyjną.

Opcja O1.a pozwala na weryfikację hipotez definiowanych przez użytkownika. Hipotezy te wyrażane są w postaci reguł decyzyjnych. Opcja O1.b pozwala na doprecyzowanie reguł przedstawionych przez użytkownika, tak aby zmaksymalizować ich jakość. W trakcie oceny efektywności klasyfikatora będącego rezultatem zastosowania opcji O1.a lub O1.b pod uwagę brane są jedynie przykłady pokrywane przez reguły. Innymi słowy, jeżeli przykład testowy nie jest pokrywany przez żadną z reguł, wówczas nie podlega on klasyfikacji.

W początkowej fazie działania algorytm stara się wygenerować reguły spełniające wymagania eksperta, a następnie kontynuuje indukcję dla pozostałych, niepokrytych jeszcze przykładów pozytywnych (oczywiście z wyjątkiem opcji O1.a oraz O1.b). Opracowana

implementacja umożliwia łączenie opcji O1, O2, O3. Przykładowo możliwe jest połączenie opcji O1.c z opcją O2 i/lub opcją O3. Algorytm rozpoczyna indukcję od doprecyzowania reguł zadanych przez użytkownika i jeśli po redefinicji nie spełniają one wymagań określonych opcjami O2, O3, to w kolejnym etapie generowane są reguły spełniające te wymagania. W przypadku jednoczesnego stosowania opcji O2 oraz O3 algorytm najpierw stara się wygenerować reguły spełniające warunki zdefiniowane w O2.

Reguły utworzone przez algorytm indukcji mogą być porównywane z regułami definiowanymi przez użytkownika. Założymy, że reguły takie są dla niego najbardziej pożądaną i użyteczną syntetyczną reprezentacją zgromadzonych danych. W nietrywialnym przypadku należy spodziewać się tego, że nie wszystkie reguły zdefiniowane przez użytkownika będą prawdziwe, a ich redefinicja może znaczaco zmieniać ich pierwotną postać. Dobrze byłoby zatem dysponować pewną miarą, pozwalającą oceniać stopień użyteczności hipotez tworzonych w sposób pół- lub całkowicie automatyczny. Miary pozwalające mierzyć użyteczność (ang. *usefulness*) bądź możliwość zastosowania (ang. *actionability*) reguł uznawane są za subiektywne miary oceniające, gdyż oceniają one reguły ze względu na subiektywne wymagania użytkownika [94, 132, 240, 345].

Miary oceniające użyteczność reguł zazwyczaj w jakimś stopniu bazują na pojęciu podobieństwa. Założymy, że dostępne są: zbiór reguł  $RUL$  oraz zbiór  $TEMP$ , zawierający reguły-szablony użyteczne dla użytkownika. Do badania użyteczności reguły  $r \in RUL$  w kontekście reguły-szablonu  $T \in TEMP$  można użyć niesymetrycznej miary podobieństwa  ${}^a sim_{Flex}^{syn}(r, T)$  (3.36), mierzącej syntaktyczne podobieństwo reguły  $r$  do szablonu  $T$ . Możliwe jest również wykorzystanie semantycznego odpowiednika tej miary:  ${}^a sim_{Flex}^{sem}$ . W przypadku badania użyteczności reguł miary podobieństwa traktowane są jako miary korzyści. Im większe podobieństwo reguły do szablonu, tym jest ona bardziej użyteczna. Przez  $u(r, T)$  oznaczmy miarę użyteczności reguły  $r$  w odniesieniu do szablonu  $T$ , przez  $U(r, TEMP)$  oznaczmy użyteczność reguły  $r$  w odniesieniu do zbioru szablonów  $TEMP$ . W zależności od dziedziny zastosowania i preferencji użytkownika definicje  $U(r, TEMP)$  mogą być różne. Przykładowo  $U(r, TEMP)$  można definiować nie tylko jako  $U(r, TEMP) = \min\{u(r, T) : T \in TEMP\}$ , ale także jako  $U(r, TEMP) = \max\{u(r, T) : T \in TEMP\}$ .

Nieco bardziej restrykcyjne podejście do badania użyteczności reguł proponuje Zhu i inni [345]. Autorzy ci zajmują się stosowalnością reguł wyznaczonych w sposób automatyczny w odniesieniu do reguł-szablonów definiowanych przez użytkownika. Jeśli reguła  $r$  zawiera chociaż jeden warunek elementarny  $w$ , taki że  $sim_T^w = 0$ , gdzie  $sim_T^w$  definiowane jest zgodnie ze wzorem (3.34) lub (3.35), to stosowalność reguły  $r$  jest równa 0.

Przyjęte założenie oznacza, że każdy warunek elementarny, zawarty w  $r$ , musi mieć niepuste przecięcie z jakimś warunkiem elementarnym, znajdującym się w szablonie  $T$ .

### 6.2.2. Reguły decyzyjne jako wzorce przeżycia

W indukcji reguł decyzyjnych zbiór przykładów podzielony jest na klasy decyzyjne. W danych zawierających informacje o przeżyciu podział przykładów na klasy decyzyjne nie jest z góry zdefiniowany. W danych tych każdy przykład charakteryzowany jest przez dwie zmienne o charakterze decyzyjnym:  $d$ , określającą status przykładu (najczęściej przykłady reprezentują pacjentów, ale mogą to być również urządzenia, których awaryjność chcemy zbadać – mamy wtedy do czynienia z analizą niezawodności) oraz  $t$ , określającą czas, jaki upłynął od pewnego momentu początkowego (może to być moment wykonania procedury terapeutycznej, remontu lub wymiany komponentu urządzenia itd.). Przed zastosowaniem algorytmu indukcji reguł do danych zawierających informacje o przeżyciu konieczna jest kategoryzacja wszystkich przykładów, czyli przyporządkowanie ich do klas decyzyjnych.

Kategoryzując dane o przeżyciu, każdy przykład przyporządkowujemy do jednej z czterech klas decyzyjnych. Aby to uczynić, niezbędne jest określenie pewnej granicznej wartości zmiennej  $t$ , która będzie stanowić podstawę do kategoryzacji przykładów. W analizie danych medycznych, zwłaszcza danych opisujących pacjentów onkologicznych, granicą taką jest 5 lat od chwili zakończenia leczenia. W przypadku pacjentów po przeszczepie szpiku kostnego jest to 5 lat od chwili wykonania przeszczepu. Oczywiście w zależności od celu analizy i jednostek, w jakich mierzone są wartości zmiennej  $t$ , wartość graniczna może być inna.

Mając określoną wartość graniczną  $th$  zmiennej  $t$ , przykłady przyporządkowujemy do jednej z czterech klas decyzyjnych. W analizie przeżycia, dotyczącej pacjentów i wykonanych u nich zabiegów terapeutycznych, będą to następujące klasy:

- *alive*, wskazująca na pacjentów żyjących, u których od chwili wykonania procedury terapeutycznej upłynęło więcej czasu niż  $th$ ;
- *dead*, wskazująca na pacjentów zmarłych, którzy od chwili wykonania procedury terapeutycznej żyli krócej niż  $th$ ;
- *dead-th*, wskazująca na pacjentów zmarłych, którzy od wykonania procedury terapeutycznej żyli dłużej niż  $th$ ;
- *alive-th*, wskazująca na pacjentów żyjących, u których od chwili wykonania procedury terapeutycznej upłynął czas krótszy niż  $th$ .

Podstawą do indukcji reguł są przykłady należące do klas *alive* i *dead*. Podczas definiowania krzywych przeżycia wykorzystywane są również przykłady z klasy *alive-th*. Przykłady reprezentujące klasę *dead-th* są ignorowane. Czas trwania życia pacjentów, reprezentujący klasę *dead-th*, był dłuższy niż  $th$ , więc teoretycznie można by je

przyporządkować do klasy *alive*, jednak nie możemy całkowicie ignorować informacji, że ci pacjenci nie żyją. Przyporządkowanie pacjentów z klasy *dead-th* do klasy *dead* również nie wydaje się prawidłowe, gdyż czas przeżycia tych pacjentów był dłuższy od wartości granicznej *th*. Jeżeli analityk nie dysponuje informacją o przyczynach śmierci pacjentów reprezentujących klasę *dead-th*, wówczas wyłączenie związań z nimi przykładów ze zbioru treningowego jest rozwiązaniem najrozsądzniejszym. W opisanych w dalszej części badaniach eksperymentalnych klasa *dead-th* była zbiorem pustym.

W literaturze opisano metodyłączenia przykładów z klas *alive-th* i *dead-th* do zbioru przykładów treningowych. Najogólniej można powiedzieć, że metody te przyporządkowują przykłady z klas *alive-th* i *dead-th* do klasy *alive* lub *dead*. Przyporządkowanie to najczęściej odbywa się przez nadanie przykładom wag odzwierciedlających ich szansę (niekoniecznie rozumianą jako prawdopodobieństwo) na znalezienie się w zbiorze *alive* lub *dead* [347, 281]. Jednak dla rozważanego w niniejszym rozdziale zbioru danych próba przyporządkowania przykładom wag zakończyła się niepowodzeniem. Otrzymane wyniki były gorsze niż użycie podczas indukcji jedynie przykładów należących do klas *alive* i *dead*.

Fundamentalną częścią analizy danych zawierających informacje o przeżyciu jest porównywanie krzywych przeżycia pomiędzy dwiema grupami obserwacji. Jedną grupę stanowią obserwacje spełniające pewne warunki, drugą grupę – obserwacje, które warunków tych nie spełniają. W naszym przypadku pierwsza grupa obserwacji utożsamiana jest z przykładami pokrywającymi regułę decyzyjną. Każda z wyznaczonych reguł dzieli analizowany zbiór na dwie grupy przykładów (pokrywanych i niepokrywanych przez regułę). Z punktu widzenia analizy przeżycia interesująca jest odpowiedź na pytanie: czy różnica pomiędzy krzywymi przeżycia, zbudowanymi na podstawie przykładów należących do tych grup, jest statystycznie istotna? Podczas wyznaczania krzywych przeżycia celowe staje się wykorzystanie przykładów należących do klasy *alive-th*. Ponieważ wszystkie przykłady opisane są identycznym wektorem cech, łatwo sprawdzić, które z przykładów reprezentujących klasę *s-alice* pokrywane są przez wyznaczone reguły.

Przyjęta metodyka analizy danych wiąże z każdą regułą (krzywą przeżycia), która definiowana jest na podstawie przykładów pokrywanych przez tę regułę. Ograniczenie liczby krzywych przeżycia równoważne jest z ograniczeniem liczby reguł. W rozważanym przypadku zastosowanie jednego ze standardowych algorytmów filtracji nie jest możliwe, gdyż algorytmy te nie mogą analizować przykładów należących do klasy *alive-th*. W związku z tym zaproponowany został nowy algorytm filtracji, pozwalający na identyfikację tzw. modeli przeżycia.

Zanim przedstawione zostaną założenia algorytmu, omówiona będzie miara *SurvDiff* (ang. *Survival Difference*), która stosowana jest przez algorytm filtracji do identyfikacji reguł najistotniejszych. Miara *SurvDiff* ocenia reguły z punktu widzenia różnicy pomiędzy

krzywymi przeżycia, definiowanymi za pomocą metody Kaplana-Meiera, dla pokrywanych i niepokrywanych przez nią zbiorów przykładów. Wartość  $SurvDiff$  równa jest  $(1-p)$ , gdzie  $p$  jest p-wartością testu log-rank pomiędzy tymi krzywymi. Większe wartości  $SurvDiff$  oznaczają większe różnice pomiędzy krzywymi przeżycia,  $SurvDiff$  jest zatem miarą korzyści. Podczas obliczania wartości miary  $SurvDiff$  pod uwagę brane są przykłady należące do klas *alive*, *alive-th* i *dead*.

Założenia algorytmu filtracji, dokonującego wyboru modelu przeżycia, są następujące:

- C1. Zbiór reguł zawiera reguły o wartości  $SurvDiff$  większej od  $\alpha$ .
- C2. Zbiór reguł pokrywa co najmniej  $n \cdot 100\%$  wszystkich dostępnych przykładów z klas *alive*, *dead*, *alive-th*.
- C3. Usunięcie zprefiltrowanego zbioru dowolnej reguły oznacza, że kryterium C2 nie jest spełnione.

Podzbiór reguł spełniających kryteria C1, C2 oraz C3 będziemy nazywać  $\alpha/n$ -Modelem-Przeżycia (ang.  $\alpha/n$ -Survival-Model), a reguły wchodzące w jego skład – wzorcami przeżycia. W zależności od klasy decyzyjnej, wskazywanej przez regułę, możemy wyróżnić dwa rodzaje wzorców przeżycia: *a-survival patterns* (są to wzorce utworzone przez reguły wskazujące na klasę *alive*) oraz *d-survival patterns* (są to wzorce utworzone przez reguły wskazujące na klasę *dead*).

Wartości parametrów  $\alpha \in [0,1]$  i  $n \in (0,1]$  definiowane są przez użytkownika. Warunek C1 ogranicza przestrzeń poszukiwania wzorców do tych, które charakteryzują się określonym poziomem statystycznej istotności. Warunek C2 pozwala na redukcję liczby wzorców przeżycia kosztem spadku pokrycia przykładów treningowych. Dodanie do C1 i C2 warunku C3 oznacza, że interesują nas modele przeżycia, które są minimalne w sensie liczby tworzących je wzorców.

W związku z przyjętymi założeniami, w zależności od parametrów zadanych przez użytkownika możemy mieć do czynienia z następującymi sytuacjami:

- model przeżycia, spełniający zadane ograniczenia, nie istnieje,
- istnieje dokładnie jeden model przeżycia, spełniający zadane warunki,
- istnieje wiele modeli przeżycia, spełniających zadane warunki.

W pierwszym przypadku użytkownik może jedynie zmniejszyć wymagania dotyczące modelu. W trzecim przypadku wskazówką do ostatecznego wyboru modelu jest wartość kryterium (6.1):

$$\sum_{r \in SM} |Cov(\{r\}, E) - Cov(SM - \{r\}, E)|. \quad (6.1)$$

We wzorze (6.1),  $Cov(SM, E) = \bigcup_{r \in SM} Cov(\{r\}, E)$  oraz  $Cov(\{r\}, E) = [r]_E$ . Wartość  $Cov(\{r\}, E)$

jest po prostu liczbą przykładów ze zbioru przykładów  $E$ , pokrywanych przez regułę-

wzorzec  $r$ . Model przeżycia o największej wartości (6.1) składa się z reguł najbardziej rozłącznych.

Znalezienie optymalnego zbioru spełniającego kryteria C1, C2, C3 oraz maksymalizującego wartość (6.1) jest problemem NP-zupełnym, gdyż sprowadza się do rozwiązania klasycznego NP-zupełnego problemu znalezienia minimalnego pokrycia zbioru [154]. Jeśli zbiory danych treningowych i reguł są niewielkie, najlepszą strategią wyboru modelu przeżycia jest zastosowanie procedury wyczerpującej, polegającej na zbadaniu wszystkich modeli spełniających warunki C1–C3 i wyborze tego, który maksymalizuje (6.1).

Dla większych zbiorów danych i reguł konieczne jest zastosowanie procedury heurystycznej. Do wyboru modelu przeżycia można zastosować algorytm, którego podstawę stanowi odpowiednio zaadaptowany algorytm filtracji *Forward*. Algorytm rozpoczyna pracę od wybrania z wejściowego zbioru tych reguł, które spełniają warunek jakości C1. Reguły te stanowią początkowy model przeżycia. Jeżeli model ten nie spełnia wymogu minimalnego pokrycia, wówczas zwracany jest pusty zbiór reguł. W przeciwnym wypadku wyjściowy zbiór reguł inicjowany jest najlepszymi, według wartości *SurvDiff*, regułami z każdej klasy decyzyjnej. Następnie kolejno do wynikowego zbioru dodawane są reguły powodujące największy wzrost wartości (6.1). Proces ten powtarzany jest dopóty, dopóki wybrany zbiór reguł nie spełni wymogu minimalnego pokrycia C2. W wynikowym zbiorze mogą znajdować się reguły redundantne, pokrywane przez inne reguły. W ostatnim etapie działania algorytm filtracji dokonuje usunięcia reguł redundantnych (następuje spełnienie warunku C3).

Wyniki algorytmu filtracji są silnie uzależnione od rankingu reguł, determinowanego przez miarę *SurvDiff*. W praktyce nie ma przeszkód, żeby ranking ten tworzony był w dowolnie inny, odpowiedni dla użytkownika sposób.

### 6.2.3. Analiza danych

#### 6.2.3.1. Charakterystyka zbioru przykładów

Analizowany zbiór danych opisuje 187 pacjentów (75 dziewcząt i 112 chłopców) w wieku od 0.6 do 20.2 lat. Spośród pacjentów 155 chorowało na różnego rodzaju nowotwory złośliwe, a 32 – na choroby innego typu. Niezależnie od rodzaju choroby u wszystkich pacjentów wymagane było wykonanie procedury terapeutycznej, polegającej na przeszczepie szpiku kostnego.

Każdy z pacjentów charakteryzowany był przez zbiór 39 atrybutów warunkowych. Atrybuty te zawierały podstawowe informacje hematologiczne oraz wybrane informacje o stanie klinicznym pacjenta w pierwszych 100 dniach od chwili wykonania przeszczepu. Znaczenie najważniejszych atrybutów opisano w tabeli 6.5.

Tabela 6.5

## Wybrane atrybuty warunkowe, opisujące pacjentów

Nazwa atrybutu	Znaczenie
Recipient_ABO Recipient_age Recipient_body_mass	Informacje o parametrach krwi, wieku i wadze biorcy
Recipient_CMV Donor_CMV CMV_status	Informacje o występowaniu wirusa cytomegalii u dawcy i biorcy oraz o tym, czy w tej materii występuje zgodność lub niezgodność (np. dawca – tak, biorca – nie)
Donor_ABO ABO_match	Parametry krwi dawcy, informacja o zgodności lub niezgodności parametrów krwi dawcy i biorcy
Gender_match	Informacja o zgodności płci pomiędzy dawcą a biorcą
Donor_age	Wiek dawcy
HLA_match HLA_mismtach	W zbiorze danych znajduje się kilka atrybutów opisujących ew. niezgodność genetyczną pomiędzy dawcą a biorcą. Podstawowy atrybut informuje o zgodności lub niezgodności (atrybut binarny). Inne atrybuty dokładniej określają typ niezgodności (7 typów niezgodności; czy niezgodność dotyczy antygenów czy aleli i jak jest duża)
Stem_cell_source	Sposób pozyskania komórek do transplantacji (szpik, krew)
Relapse	Czy obecna choroba jest wznową wcześniej przebytej choroby?
PLT_recovery	Po jakim czasie (po ilu dniach) od chwili przeszczepu pacjent uzyskuje odnowę płytek krwi (liczba płytek >50000/mm <sup>3</sup> )?
ANT_recovery	Po jakim czasie (po ilu dniach) od chwili przeszczepu pacjent uzyskuje odnowę granulocytów-neutrofili (liczba neutrofili >0.5 x 10 <sup>9</sup> /L)?
T_aGvHD_III_IV	Liczba dni od chwili przeszczepu do momentu stwierdzenia wystąpienia III lub IV stopnia nasilenia choroby <i>przeszczep przeciwko gospodarzowi</i>
aGVHD_III_IV	Czy odnotowano ostrą postać choroby <i>przeszczep przeciwko gospodarzowi</i> w nasileniu III lub IV stopnia?
extcGvHD	Czy odnotowano wystąpienie przewlekłej postaci choroby <i>przeszczep przeciwko gospodarzowi</i> ?
CD34 (10 <sup>6</sup> /kg)	Wielkość dawki komórek CD34 na kg masy biorcy
CD3 (10 <sup>8</sup> /kg)	Wielkość dawki komórek CD34 na kg masy biorcy
CD3/CD34	Stosunek dawki CD3 do CD34

Z danych usunięto wszelkiego rodzaju atrybuty mogące bezpośrednio świadczyć o przeżyciu pacjenta (m.in. informacje o roku, w którym wykonano przeszczep, gdyż po 2004 roku znaczowo poprawiono procedury medyczne).

Głównym celem analizy było utworzenie systemu klasyfikującego, przewidującego 5-letnią przeżywalność pacjentów. Analizowano również, według jakich zasad (za pomocą jakich reguł) przebiega proces klasyfikacji oraz to, czy klasyfikator używa atrybutów i warunków elementarnych, szczególnie interesujących z medycznego punktu widzenia. Dodatkowo sprawdzano także, jak wymuszenie wystąpienia pewnych atrybutów i warunków elementarnych w indukowanych regułach wpływa na jakość klasyfikatora. Motywacją takiego działania była chęć weryfikacji hipotezy łączącej wielkość dawek komórek CD34 i CD3 z czasem przeżycia pacjentów. Kolejnym celem analizy było sprawdzenie, czy zwiększenie dawek CD34 i CD3 nie wpływa na częstsze występowanie choroby *przeszczep przeciwko gospodarzowi*. Łagodne postaci tej choroby są korzystne dla pacjentów onkologicznych, gdyż przyczyniają się do eliminacji komórek rakowych,

pozostających w cieles pacjenta po kursie megachemioterapii. Ostre i przewlekłe postaci choroby *przeszczep przeciwko gospodarzowi* są jednak niepożądane, gdyż często kończą się śmiercią pacjenta. Zgodnie z metodyką opisaną w poprzednim podrozdziale w zbiorze danych wydzielono trzy klasy pacjentów: żyjących, u których od chwili przeszczepu upłynęło ponad 5 lat (*alive*); zmarłych w okresie krótszym niż 5 lat od chwili przeszczepu (*dead*); żyjących, u których od chwili przeszczepu upłynęło mniej niż 5 lat (*alive-th*). Rozkład przykładów liczby pomiędzy klasami decyzyjnymi był następujący: 36 przykładów z klasy *alive*, 85 z klasy *dead* i 66 przykładów klasy *alive-th*. Więcej szczegółów na temat analizowanego zbioru danych można znaleźć w [158].

#### 6.2.3.2. Metodyka

Analizując dane, przyjęto następującą metodykę postępowania:

- Dokonano indukcji reguł, używając do tego celu różnych miar jakości, w szczególności tych znajdujących się w zbiorze miar najbardziej obiecujących *Max*.
- Jako kryterium jakości klasyfikatora stosowano miarę  $\text{AvgSurvDiff} \cdot \text{BAcc}$ ; *AvgSurvDiff* oznacza średnią wartość miary *SurvDiff* reguł tworzących klasyfikator; *BAcc* jest średnią dokładnością klas decyzyjnych.
- Do weryfikacji klasyfikatorów zastosowano 10-krotną warstwową walidację krzyżową, którą powtórzono 10 razy.
- Za pomocą miary prowadzącej do najlepszych wyników klasyfikacji dokonano indukcji reguł na podstawie całego dostępnego zbioru przykładów. Na podstawie utworzonych reguł zidentyfikowano model przeżycia.
- Każdorazowo oceniano również ważność warunków elementarnych, tworzących wyznaczone reguły. Do oceny warunków stosowano standardową postać indeksu Banzhafa (5.2).

#### 6.2.3.3. Rezultaty

Klasyfikator o najlepszej jakości otrzymano po zastosowaniu algorytmu q-ModLEM(MMM) i miary *Corr*. W tabeli 6.6 zamieszczono rezultaty klasyfikatorów utworzonych na podstawie kilku miar jakości. W tabeli 6.7 przedstawiono wyniki klasyfikatorów otrzymanych za pomocą innych pokryciowych algorytmów indukcji reguł i drzew decyzyjnych.

Tabela 6.6  
Charakterystyka klasyfikatorów utworzonych na podstawie różnych miar jakości

Miara	Liczba reguł	Średnia liczba warunków	Średnia dokładność reguł	Średnie pokrycie reguł	Średnia dokładność klasy <i>alive</i> [%]	Średnia dokładność klasy <i>dead</i> [%]	<i>AvgSurvDiff</i> · <i>BAcc</i>
<i>Corr</i>	19.5	4.4	0.903	0.539	61.8	71.8	0.65
<i>CN2</i>	32.0	3.0	0.995	0.266	62.3	68.4	0.63
<i>RSS</i>	22.8	3.9	0.888	0.501	41.7	82.0	0.60
<i>g</i>	39.2	2.4	1.000	0.195	31.7	80.7	0.55
<i>wLap</i>	46.3	1.9	1.000	0.134	46.3	62.5	0.51
<i>CI</i>	49.8	1.7	1.000	0.113	37.7	72.7	0.50

Tabela 6.7  
Porównanie jakości klasyfikatorów utworzonych przez różne algorytmy indukcji reguł i drzew decyzyjnych

Algorytm	Średnia dokładność klasy <i>alive</i> [%]	Średnia dokładność klasy <i>dead</i> [%]	BAcc	Acc
RIPPER	49.8±31.3	82.1± 13.7	66.0±13.0	72.6± 8.9
PART	39±22.7 <sup>(+)</sup>	75.6±13.3	57.3±11.9	64.6± 10.5
J48	30.9±22.8 <sup>(+)</sup>	75.3±14.1	53.1 ±12.6 <sup>(+)</sup>	62.1±11.4
SimpleCart	1.3 (8.5) <sup>(+)</sup>	98.8±6.6 <sup>(+)</sup>	50.1 ±2.8 <sup>(+)</sup>	69.8±5.1
q-ModLEM <i>Corr</i>	61.8±28.7	71.8±15.1	66.8±14.4	68.6±12.0

W tabeli 6.7 oznaczenia (-)/(+) informują o tym, czy dany wynik jest statystycznie gorszy/lepszy od rezultatów uzyskanych przez algorytm q-ModLEM(MMM), stosujący miarę *Corr*. Ze względu na przyjętą metodykę testowania (10x10cv) do porównań wykorzystano skorygowaną postać testu t, a poziom istotności ustalono na 0.05.

W jednym przypadku wynik algorytmu q-ModLEM(MMM) nie jest statystycznie najlepszy. Najwyższą dokładność klasy *dead* pozwala osiągnąć model utworzony przez algorytm SimpleCart. Jednocześnie model ten najgorzej prognozuje przynależność przykładów dla klasy *alive*, konsekwencją czego są: niska wartość *BAcc* oraz niska jakość wzorców przeżycia, opisujących klasę *alive*.

Wariancja wyników jest duża, analiza reguł tworzonych w kolejnych przebiegach walidacji krzyżowej wskazywała jednak, że są one stosunkowo stabilne. Stabilność rozumiemy tutaj w tym sensie, że w kolejnych przebiegach walidacji krzyżowej reguły zbudowane były na podstawie podobnych zbiorów warunków elementarnych.

Analiza ważności warunków elementarnych wykazała, że dla klasy *alive* najważniejsze warunki to: brak przewlekłej postaci choroby *przeszczep przeciwko gospodarzowi*, brak wznowy, odnowa płytek krwi i młody wiek pacjenta. Dla klasy *dead* najważniejsze były warunki określające: wznowę choroby, niskie dawki CD34 oraz bliski pełnoletniości wiek

pacjenta. Warunki te okazały się również najważniejsze dla reguł utworzonych na podstawie analizy całego dostępnego zbioru przykładów.

W zbiorze przykładów treningowych algorytm wyznaczył 18 reguł (7 dla klasy *alive* i 11 dla klasy *dead*). Reguły te poprawnie sklasyfikowały 92.6% przykładów treningowych (dokładność klasy *alive* wynosiła 100%, dokładność klasy *dead* – 89.4%). Zastosowanie algorytmu redefinicji nie poprawiało tego wyniku. Na podstawie reguł wyznaczono następujący model przeżycia, o parametrach 0.999/0.8:

R1

**Jeżeli** extcGvHD=Nie **oraz** Relapse=Nie **oraz** Donor\_age $\in[26.4;46.1)$  **oraz** PLT\_recovery=Tak **oraz** ANC\_recovery<24 **oraz** Recipient\_body\_mass<73.5 **oraz** (T\_aGvHD\_III\_IV>14 lub nie wystąpił), **to alive**  
(precision=0.8621, coverage=0.6944, SurvDiff=1.0, covered=64(4), 5-year mean=4.71).

R2

**Jeżeli** PLT\_recovery $\geq 13.5$  **oraz** Relapse=Nie **oraz** Recipient\_age<17.6 **oraz** Recipient\_body\_mass< 72 **oraz** Donor\_age<45.5 **oraz** (T\_aGvHD\_III\_IV>14 lub nie wystąpił) **oraz** Gender\_match= (inni niż Kobieta do Mężczyzny), **to alive**  
(precision=0.7073, coverage=0.8056, SurvDiff=1.0, #covered=76(12), 5-year mean=4.31).

R3

**Jeżeli** CD34 $\in[1.265 \cdot 10^6; 10.815 \cdot 10^6)$  **oraz** Recipient\_age $\geq 11.6$  **oraz** Donor\_age $\geq 20.5$ , **to dead**  
(precision=0.95, coverage=0.4471, SurvDiff=0.999996, #covered=56(38), 5-year mean=1.93).

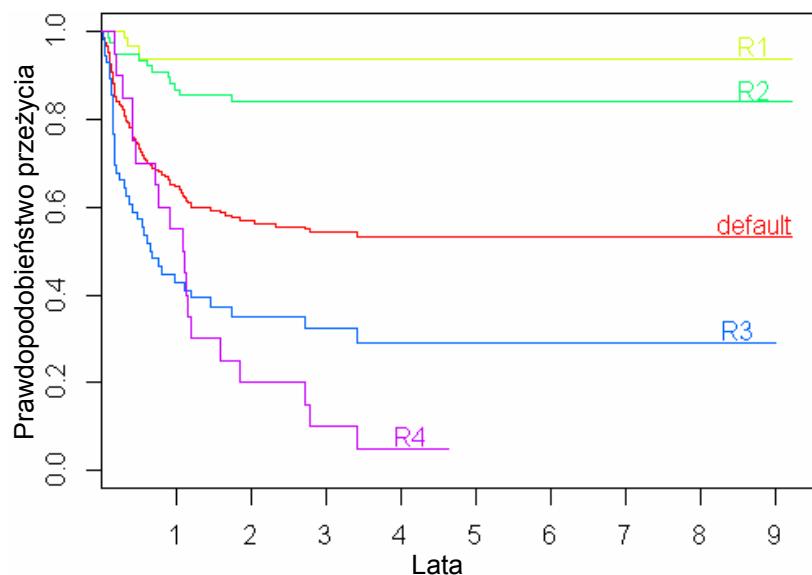
R4

**Jeżeli** Relapse =Tak **oraz** RecipientRh=Rh+, **to dead**  
(precision=1.0, coverage=0.2235, SurvDiff=0.999864, #covered=20(19), 5-year mean=1.36).

Pod każdą z reguł umieszczono informacje o jej dokładności, pokryciu, wartości miary *SurvDiff* oraz ogólnej liczbie pokrywanych przez nią przykładów (w nawiasie podawana jest liczba pokrywanych przykładów, należących do klasy *dead*). Wartość określana jako *5-year mean* oznacza średnią oczekiwana długosć życia pacjentów spełniających warunki danej reguły, przy czym do obliczenia tej średniej przyjęto, że maksymalna długosć przeżycia po przeszczepie wynosi 5 lat (tzn. *5-year mean* nie odzwierciedla oczekiwanej wartości całkowitego przeżycia).

Dla lekarzy szczególnie interesujące było sprawdzenie, czy w regułach pojawiają się warunki zbudowane na podstawie atrybutów CD34 i CD3. Warunek  $CD34 \in [1.265 \cdot 10^6; 10.815 \cdot 10^6)$  znalazł się jedynie w regułach opisujących klasę *dead*. Znaczenie warunku należy interpretować w taki sposób, że niskie dawki CD34 nie sprzyjają wydłużeniu czasu przeżycia pacjentów. Warunek  $CD34 \in [1.265 \cdot 10^6; 10.815 \cdot 10^6)$  był drugi pod względem ważności w klasie *dead*. Najważniejszym warunkiem dla tej klasy był *Relapse =Tak*, oznaczający, że najgorzej rokują pacjenci po nawrocie choroby.

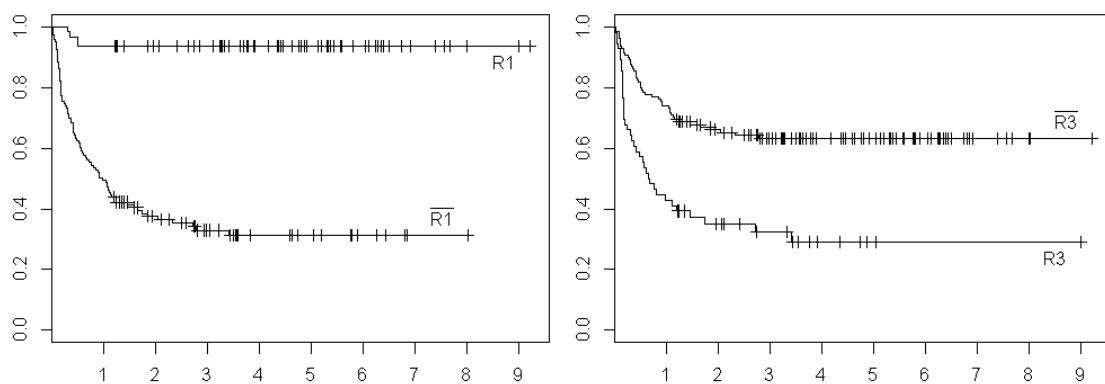
Na rysunku 6.3 przedstawiono krzywe przeżycia, wyznaczone metodą Kaplana-Meiera, dla przykładów pokrywanych przez reguły R1–R4. Jako uzupełnienie przedstawiono również krzywą przeżycia, wyznaczoną dla wszystkich przykładów treningowych. Pacjenci spełniający warunki reguł R1 i R2 charakteryzują się znacznie dłuższym średnim 5-letnim czasem przeżycia niż wynosi średnia w całej badanej grupie pacjentów. U pacjentów spełniających warunki reguł R3 i R4 czas ten jest znacznie krótszy niż średnia w badanej grupie. Potwierdzają to zarówno krzywe przeżycia, jak i wartość *5-year mean*.



Rys. 6.3. Krzywe przeżycia dla przykładów pokrywanych przez reguły R1–R4

Fig. 6.3. Survival curves for examples covered by the rules R1–R4

Wartość miary *SurvDiff* jest bliska jedności dla każdej z reguł R1–R4. Oznacza to, że istnieją statystycznie istotne różnice pomiędzy krzywymi przeżycia, wyznaczanymi dla grup pacjentów spełniających i niespełniających przesłanki tych reguł. Na rysunku 6.4 zaprezentowano krzywe przeżycia dla przykładów pokrywających i niepokrywających reguły R1 (wykres lewy) i reguły R3 (wykres prawy). Przez + oznaczono obserwacje ucięte, a więc albo wskazujące na pacjentów zmarłych, albo żyjących, u których od chwili przeszczepu upłynęło mniej niż 5 lat.

Rys. 6.4. Krzywe przeżycia reguł R1 i R3 oraz ich dopełnień  $\overline{R1}$  i  $\overline{R3}$ Fig. 6.4. Survival curves for rules R1 and R3, and their complements  $\overline{R1}$  and  $\overline{R3}$ 

Podstawą dotychczas prezentowanych rezultatów był algorytm indukcji, tworzący reguły w sposób całkowicie automatyczny. Jak już wspomniano, jednym z celów analizy było sprawdzenie, czy większe dawki preparatów CD34 i CD3 pozytywnie wpływają na czas przeżycia. Skoncentrujmy się na dawkowaniu CD34. W pracy [158] Kaławak i inni stawiają hipotezę, że większe dawki tego preparatu mają pozytywny wpływ na przeżywalność pacjentów, a dawki mniejsze – negatywne. Należy zaznaczyć, że nie zawsze możliwe jest uzyskanie z materiału przeszczepowego wysokich dawek CD34. Możliwe jest jednak podejmowanie pewnych procedur medycznych, mających za zadanie zwiększenie liczby komórek CD34 w podawanym materiale przeszczepowym. Dawkowanie preparatu podzielono na dwie kategorie:  $CD34 \geq 10 \cdot 10^6$  komórek/kg wagi pacjenta oraz  $CD34 \leq 10 \cdot 10^6$ . Granica podziału kategorii przebiega blisko mediany wartości atrybutu CD34. Na podstawie zdefiniowanego podziału wyspecyfikowano dwie reguły (R5 i R6), które stanowiły hipotezy podlegające weryfikacji. Zauważmy, że w regule R3, wygenerowanej w sposób automatyczny, jeden z warunków elementarnych –  $CD34 \in [1.265 \cdot 10^6; 10.815 \cdot 10^6]$  – ma prawie identyczną postać z warunkiem  $CD34 \leq 10 \cdot 10^6$ , postulowanym przez lekarzy.

Na podstawie reguł R5 i R6 przeprowadzono proces klasyfikacji całego dostępnego zbioru przykładów, a także wykonano walidację krzyżową.

R5

**Jeżeli**  $CD34 \geq 10 \cdot 10^6$ , **to alive**

(precision=0.3922, coverage=0.5556, SurvDiff=0.9873, #covered=86(31), 5-year mean=3.43).

R6

**Jeżeli**  $CD34 < 10 \cdot 10^6$ , **to dead**

(precision=0.7714, coverage=0.6353, SurvDiff=0.9873, #covered=101(54), 5-year mean=2.61).

W zbiorze treningowym reguły R5 i R6 klasyfikują przykłady z dokładnością wynoszącą, 55.6% dla klasy *alive* oraz 63.5% dla klasy *dead*. W trybie walidacji krzyżowej klasyfikator składający się z tych reguł odnotował prawie identyczną dokładność (wynika to z faktu

użycia walidacji warstwowej). Zdolności klasyfikacyjne klasyfikatora złożonego z reguł R5 i R6 nie różnią się w sposób statystycznie istotny od klasyfikatora uzyskanego przez q-ModLEM(MMM) i miarę *Corr*. Dokładność reguły R5, wynosząca 0.3922, nie wydaje się zbyt wysoka, zauważmy jednak, że dokładność klasy *alive*, wynikająca z rozkładu liczby przykładów, wynosi 0.297. Aby udokładnić reguły R5 i R6, zainicjowano nimi algorytm indukcji reguł. W ten sposób uzyskano reguły R7 i R8.

R7

**Jeżeli**  $CD34 \geq 10.815$  **oraz**  $PLT\_recovery \geq 14$  **oraz**  $Relapse = \text{Nie}$  **oraz**  $Donor\_age < 45.7$  **oraz**  $T\_aGvHD\_III\_IV = 10000$  ( $T\_aGvHD\_III\_IV = 10000$  oznacza, że wystąpiła ostra postać choroby przeszczep przeciwko gospodarzowi), **to alive**  
(precision=1.0, coverage=0.4167, *SurvDiff*=1.0, #covered=31(0), 5-year mean=5.0).

R8

**Jeżeli**  $CD34 \in [1.265; 10.815)$  **oraz**  $Recipient\_age \geq 11.6$  **oraz**  $Donor\_age \geq 20.5$  **oraz**  $Recipient\_bodymass > 31$ , **to dead**  
(precision=0.9737, coverage=0.4353, *SurvDiff*=1.0, #covered=53(37), 5-year mean=1.84).

Reguły R7 i R8 charakteryzują się znacznie większą dokładnością i nieco mniejszym pokryciem niż reguły R5 i R6. R7 i R8 rozpoznają 43.8% przykładów treningowych, klasyfikując je z dokładnością: 93.8% (klasa *alive*) oraz 100% (klasa *dead*). Podobne rezultaty uzyskano w drodze walidacji krzyżowej. Dokładność klasyfikatora jest znakomita, jednakże pokrycie zbioru przykładów jest zbyt małe, dlatego przeprowadzono analizę ważności warunków elementarnych, tworzących te reguły. Najmniej ważne warunki elementarne w regule R7 to  $Donor\_age < 45.7$ , a także  $T\_aGvHD\_III\_IV = 10000$  (nie wystąpiła ostra postać choroby przeszczep przeciwko gospodarzowi). Najmniej ważnymi warunkami dla reguły R8 są  $Donor\_age \geq 20.5$  oraz  $Recipient\_bodymass > 31$ . Po usunięciu najmniej istotnych warunków elementarnych otrzymano reguły R9 i R10, które ostatecznie zostały zaakceptowane przez lekarzy jako zawierające informację, która doprecyzowuje warunki skutecznego przewidywania czasu przeżycia po zwiększym i zmniejszonym dawkowaniu CD34.

R9

**Jeżeli**  $CD34 \geq 10.815$  **oraz**  $PLT\_recovery \geq 14$  **oraz**  $Relapse = \text{Nie}$ , **to alive**  
(precision=0.8, coverage=0.4444, *SurvDiff*=1.0, #covered=42(4), 5-year mean= 4.59).

R10

**Jeżeli**  $CD34 < 10.815$  **oraz**  $Recipient\_age \geq 11.6$ , **to dead**  
(precision=0.8889, coverage=0.4706, *SurvDiff*= 0.99995, #covered=62(40), 5-year mean= 2.1).

Reguły R9 i R10 pokrywają 56.2% przykładów treningowych, klasyfikując je z dokładnością równą: 75.5% dla klasy *alive* oraz 85.1% dla klasy *dead*.

Otrzymane rezultaty potwierdzają hipotezę, że większe dawki preparatu CD34 przyczyniają się do wydłużenia czasu przeżycia. Powodzenia terapii należy się spodziewać wśród pacjentów z pierwszym rzutem choroby. Zwłaszcza u pacjentów nastoletnich zauważono zależność pomiędzy małymi dawkami CD34 a przeszczepem zakończonym niepowodzeniem. Ważną informacją jest również to, że w zbiorze przykładów pokrywanych przez regułę R9 zagęszczenie pacjentów, u których wystąpiła ostra postać choroby *przeszczep przeciwko gospodarzowi* jest takie samo jak w zbiorze przykładów spełniających jedynie warunek ( $PLT\_recovery \geq 14$  oraz  $Relapse = Nie$ ). Oznacza to, że zastosowanie większych dawek preparatu CD34 nie wpływa na pojawianie się niekorzystnej, ostrej postaci choroby *przeszczep przeciwko gospodarzowi*. Interesującym i silnie potwierdzającym celowość większego dawkowania CD34 spostrzeżeniem jest także to, że w zbiorze spełniającym warunek ( $CD34 \geq 10.815$  oraz  $PLT\_recovery \geq 14$  oraz  $Relapse = Nie$ ) występuje o 50% mniej pacjentów z przewlekłą postacią choroby *przeszczep przeciwko gospodarzowi* niż w zbiorze pacjentów spełniających warunek ( $PLT\_recovery \geq 14$  oraz  $Relapse = Nie$ ).

Przedstawiony w niniejszym rozdziale przykład analizy danych pokazuje, że:

- indukcja reguł decyzyjnych może być podstawą do definiowania wzorców i modeli przeżycia,
- półautomatyczna indukcja reguł stanowi użyteczne narzędzie do weryfikacji hipotez badawczych, umożliwia także interaktywne konstruowanie regułowych modeli danych, składających się z reguł interesujących dla użytkownika.

Badania przeprowadzone nad efektywnością wzorców przeżycia, generowanych na podstawie reguł decyzyjnych, pokazują, że ich efektywność jest zadowalająca. W tabeli 6.8 zaprezentowano wartości wskaźnika Breira (który jest miarą kosztu) [102], otrzymane po analizie opisywanego w niniejszym rozdziale zbioru danych okołoprzeszczepowych. W analizie wykorzystano: metodę Kaplana-Meiera [155], drzewa przeżycia, wyznaczone przez algorytmy zawarte w pakiecie obliczeń statystycznych R [229], oraz wzorce przeżycia, utworzone na podstawie reguł wyznaczonych przez algorytm q-ModLEM(MMM) i miarę *Corr*. Eksperyment wykonano, stosując 10-krotną warstwową walidację krzyżową, powtórzoną 10 razy.

Tabela 6.8

Porównanie wartości wskaźnika Briera dla modeli przeżycia otrzymanych na podstawie algorytmów indukcji drzew i reguł przeżycia

Algorytm	Wartość wskaźnika Briera
Kaplan-Meier	$0.264 \pm 0.010$
RPART	$0.364 \pm 0.104$
CTREE	$0.226 \pm 0.070$
q-ModLEM(MMM) <i>Corr</i>	$0.256 \pm 0.028$

Prowadzone przez autora prace nad metodami definiowania modeli przeżycia koncentrują się obecnie na indukcji reguł, ukierunkowanej bezpośrednio na maksymalizację wartości miar przeznaczonych do oceny krzywych przeżycia (np. *SurvDiff* ).

Badania nad danymi okołoprzeszczepowymi będą kontynuowane. W szczególności planowane jest wzięcie pod uwagę czynnika temporalnego określającego stan pacjenta w kolejnych dniach po przeszczepie, oraz informacji o zastosowanych procedurach terapeutycznych. W badaniach tych pomocne będą metody pomiaru siły interwencji reguł [108] oraz metodyka odkrywania i modelowania, za pomocą reguł, złożonych systemów zmieniających się w czasie [19, 212, 274, 294]. Planowany jest także powrót do prac Mrózka, prowadzonych wspólnie z Plonką i autorem [202, 203]. W publikacjach tych użyto maszyny stanów do opisana zasad, według jakich zmieniają się stany modelowanego procesu.

### 6.3. Opis grup genów za pomocą reguł decyzyjnych

Ostatni z przykładów praktycznych zastosowań algorytmów indukcji reguł związany jest z indukcją reguł decyzyjnych, przeznaczonych jedynie do opisu danych. Celem jest indukcja wszystkich statystycznie istotnych reguł przeznaczonych do funkcjonalnego opisu genów. Z punktu widzenia indukcji reguł problem jest interesujący dlatego, że atrybuty warunkowe w analizowanej tablicy decyzyjnej tworzą strukturę znaną jako ontologia [114]. Pomiędzy atrybutami zachodzą więc pewne zależności, które można wykorzystać do ukierunkowania procesu indukcji. Algorytm indukcji pod uwagę musi brać także to, że interesujące są jedynie reguły zawierające określony rodzaj warunków elementarnych. Wymagania te stały się przyczynkiem do modyfikacji opracowanego przez Stefanowskiego i Vandarpootena [285] algorytmu Explore oraz do przedstawienia nowej metody redukcji atrybutów.

Z punktu widzenia oceny reguł interesującym zagadnieniem, poruszonym w niniejszym rozdziale, jest selekcja reguł za pomocą miary złożonej. Miara ta bierze pod uwagę wartości obiektywnych miar jakości oraz wartość miary subiektywnej, zdefiniowanej po to, aby uwzględnić specyficzne wymagania użytkownika.

#### 6.3.1. Definicja problemu

Podstawą indukcji reguł jest tablica decyzyjna, w której przykłady utożsamiane są z genami (identyfikatorami genów), natomiast atrybuty – z terminami opisującymi geny. Terminy te tworzą bazę Ontologii Genowych (ang. *Gene Ontology*) [9]. Baza ta składa się z trzech ontologii zawierających terminy opisujące różne funkcje genów. Wyróżniane są następujące ontologie: proces biologiczny (ang. *biological process BP*), funkcja molekularna (ang. *molecular function MF*), komponent komórkowy (ang. *cellular component CC*). W literaturze przedmiotu termin zawarty w którejkolwiek z tych ontologii określa się mianem terminu GO (ang. *GO term*).

Tablica decyzyjna podzielona jest na klasy decyzyjne, geny przyporządkowywane są do klas decyzyjnych na podstawie analizy wyniku eksperymentu z wykorzystaniem mikromacierzy DNA [11]. Nie wnikając w szczegóły biologiczne, które nie mają tutaj zasadniczego znaczenia, można powiedzieć, że klasy decyzyjne złożone są z genów zachowujących się w podobny sposób podczas eksperymentu. Końcowym i podstawowym celem eksperymentu mikromacierzowego jest biologiczna interpretacja otrzymanych wyników. W szczególności celem analizy jest znalezienie pewnych wspólnych cech charakteryzujących grupy genów. Taką wspólną cechą mogą być pojedyncze terminy GO, opisujące geny należące do danej grupy i nieopisujące genów z innych grup. Bardziej zaawansowana analiza polega na znalezieniu zbioru (zbiorów) terminów, które jednocześnie opisują daną grupę genów lub pewne jej podzbiory. Łatwo zauważyc, że tak zdefiniowany problem opisu można przedstawić jako zadanie indukcji reguł. Podczas indukcji grupa genów, będąca przedmiotem szczególnego zainteresowania, stanowi zbiór przykładów pozytywnych, a wszystkie pozostałe geny stanowią zbiór przykładów negatywnych. Zadanie ma charakter czysto opisowy, gdyż wyznaczone reguły nie zostaną użyte do klasyfikacji nieznanych przykładów. Zakładamy, że rozważamy wszystkie możliwe geny i wszystkie terminy ontologiczne lub że ich liczba jest celowo ograniczona przez eksperymentatora.

Tablicę decyzyjną, zawierającą dostępny zbiór przykładów, zdefiniujemy jako  $\mathbf{DT} = (G, T \cup \{d\})$ , gdzie  $G$  jest zbiorem genów podzielonych na  $n$  grup  $G_1, G_2, \dots, G_n$ ,  $T$  jest zbiorem terminów GO, a atrybut decyzyjny  $d$  wskazuje na przynależność danego genu  $g$  do jednej z grup  $G_1, G_2, \dots, G_n$ . Zakładamy ponadto, że zbiór  $T$  zawiera terminy z jednej lub kilku ontologii ( $BP$ ,  $MF$ ,  $CC$ ).

Ontologia genowa to acykliczny graf skierowany, w którym każdy węzeł identyfikowany jest przez termin GO. Do każdego terminu przypisane są opisywane przez niego geny. Formalnie każdą z ontologii genowych możemy rozważać jako zbiór częściowo uporządkowany  $GO = (T, \leq)$ , gdzie  $\leq$  jest dwuargumentową relacją, określoną w  $T^2$  w następujący sposób:

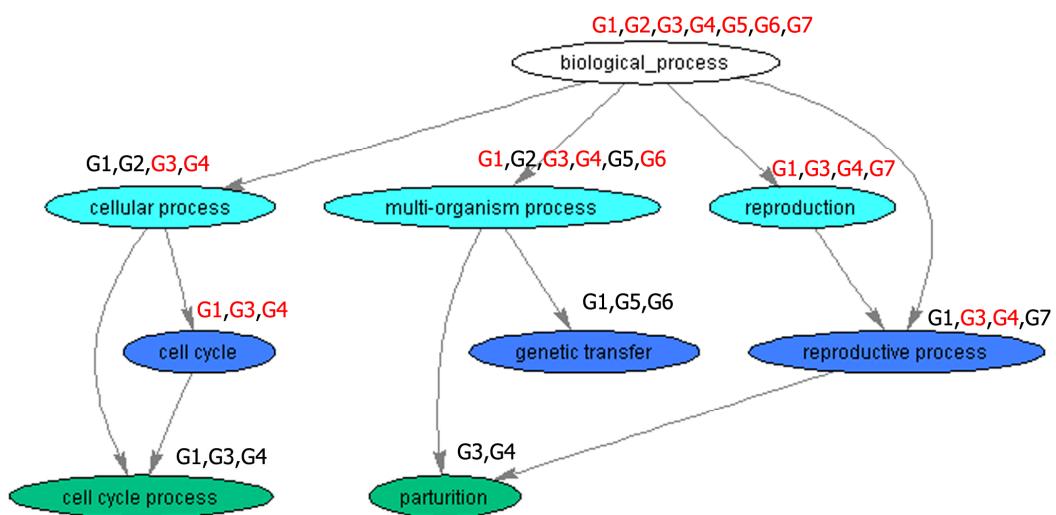
$$\forall t_i, t_j \in T \quad t_j \leq t_i \Leftrightarrow G^{t_j} \subseteq G^{t_i}, \quad (6.2)$$

gdzie  $G^t$  oznacza zbiór genów opisanych terminem  $t$ .

Relacja  $\leq$  jest relacją słabego częściowego porządku (zwrotna, antysymetryczna i przechodnia). Zauważmy, że  $t_j \leq t_i$  wtedy i tylko wtedy, gdy istnieje ciąg indeksów (numerów) terminów GO ( $i = l_1, l_2, \dots, l_k = j$ ), identyfikujący w grafie GO ścieżkę  $(t_{i=l_1}, t_{l_2}, \dots, t_{l_{k-1}}, t_{l_k=j})$ , taką że  $t_{l_{m+1}} \leq t_{l_m}$  dla każdego  $m = 1, 2, \dots, k-1$ .

Na rysunku 6.5 przedstawiono fragment ontologii genowej. W bazie ontologii każdy termin może opisywać funkcje wielu genów. Na rysunku 6.5 geny oznaczone kolorem

czarnym przypisano w bazie GO do konkretnych terminów GO na podstawie wiedzy eksperckiej. Każde takie przypisanie weryfikowane jest przez konsorcjum GO [9]. Część genów przypisana jest do terminów GO na podstawie zależności wynikających z hierarchicznego charakteru ontologii (są to geny oznaczone na rysunku 6.5 ciemniejszą czcionką).



Rys. 6.5. Fragment Ontologii Genowej  
Fig. 6.5. The part of Gene Ontology graph

Zauważmy, że na podstawie ontologii możliwe jest zdefiniowanie systemu informacyjnego. W systemie tym każdy termin GO utożsamiany jest z binarnym atrybutem warunkowym. Dla każdego  $g \in G$ ,  $t(g)=1$  oznacza, że gen  $g$  jest opisywany przez termin  $t$ , a zapis  $t(g)=0$  oznacza, że  $g$  nie jest opisywany przez  $t$ . Jeśli uwzględnimy hierarchię pomiędzy terminami GO, to otrzymamy inny system informacyjny niż system bazujący jedynie na bezpośrednich przypisaniach genów do terminów GO. Różnicę tę ilustruje tabela 6.9. W tabeli tej podkreślone wartości 1 w systemie niebiorącym pod uwagę przypisań wynikających z hierarchii pomiędzy terminami zostały zastąpione przez 0. Tabela 6.9 powstała na podstawie analizy ontologii prezentowanej na rysunku 6.5, nazwy atrybutów są skrótami nazw terminów widocznych na tym rysunku.

Tabela 6.9

System informacyjny, utworzony na podstawie analizy struktury ontologii genowej, przedstawionej na rysunku 6.5

Gen\termin	bp	cp	mp	r	cc	gt	rp	ccp	p
g1	1	1	1	1	1	1	1	1	0
g1	1	1	1	0	0	0	0	0	0
g3	1	1	1	1	1	0	1	1	1
g4	1	1	1	1	1	1	1	1	1
g5	1	0	1	0	0	1	0	0	0
g6	1	0	1	0	0	0	0	0	0
g7	1	0	0	1	0	0	1	0	0

Dodając do otrzymanego w ten sposób systemu informacyjnego kolumnę informującą o przyporządkowaniu genów do grup, uzyskamy tablicę decyzyjną, która stanowi treningowy zbiór przykładów.

W tablicy tej interesuje nas indukcja reguł o następującej semantycie:

*jeśli gen opisany jest przez terminy GO, tworzące przesłankę reguły, to należy on do grupy wskazywanej w konkluzji reguły.*

Cel indukcji reguł jest czysto opisowy, dlatego też zadaniem algorytmu jest wyznaczenie wszystkich możliwych reguł charakteryzujących się istotnością statystyczną lepszą od zadanej przez użytkownika wartości progowej.

Zadanie opisu grupy genów za pomocą terminów ontologicznych było realizowane przez kilku autorów [44, 115, 137, 195, 196]. W szczególności Carmona-Saez i inni stosowali do tego celu odmianę algorytmu indukcji reguł asocjacyjnych. Opracowana metoda była podstawą do implementacji serwisu internetowego Genecodis [45]. Serwis ten nie uwzględnia faktu, że pomiędzy terminami istnieje hierarchia wynikająca z ontologii, stąd w przesłankach tworzonych reguł znajdują się koniunkcje terminów GO, które mogą być zastąpione przez jeden (najbardziej specyficzny) termin. Hvidsten, Komorowski i inni [137], a także Midelfart [195, 196] generują reguły w nieco odmienny sposób. Algorytm opracowany przez Hvidstena i Komorowskiego wraz ze współpracownikami buduje reguły, których przesłanki zawierają warunki pozwalające na identyfikację grupy genów zachowujących się w podobny sposób podczas eksperymentu mikromacierzowego, natomiast konkluzja wskazuje na termin GO, opisujący te geny. Podejście takie pozwala uniezależnić się od procesu grupowania danych, ale natręcza innych problemów, m.in. mamy do czynienia z dużą liczbą reguł pokrywających podobne zbiory genów. Interpretacja biologiczna utworzonych w ten sposób reguł może być utrudniona i czasochłonna. Midelfart prezentuje podobny algorytm indukcji, stosując dodatkowo metody teorii zbiorów przybliżonych. W konkluzjach wyznaczanych reguł znajdują się terminy GO, leżące jak najniżej w hierarchii terminów, a same reguły muszą spełniać zadane przez użytkownika

wymogi, dotyczące minimalnej dokładności i minimalnej ogólności. Obie metody indukcji nie pozwalają na opis genów za pomocą koniunkcji terminów GO.

Gruca w pracy [115] rozważa problem indukcji reguł generowanych na podstawie reduktów względnych, nie biorąc pod uwagę hierarchii terminów GO. Takie podejście prowadzi do indukcji niewielkiej liczby reguł, które pokrywają niewielki procent genów tworzących poszczególne grupy. Wyznaczone reguły charakteryzują się niskim pokryciem. Ponadto w przywoływanej pracy odróżnialność, interpretowana jako *termin opisuje geny z danej grupy i nie opisuje genów z pozostałych grup*, traktowana jest na równi z odróżnialnością interpretowaną jako *termin nie opisuje genów z danej grupy i opisuje geny należące do pozostałych grup*. Oczywiście jest, że aby dokonać indukcji interesujących nas reguł decyzyjnych, interesujący jest jedynie pierwszy z wymienionych rodzajów odróżnialności.

Proponowane w dalszej części metody: indukcji, oceny i redukcji liczby reguł, a także metoda redukcji terminów GO są wynikiem prac prowadzonych przez autora wspólnie z Grucą oraz Polańskim [116, 255, 257]. Pozwalają one na uniknięcie niedogodności, jakimi obarczone są metody omówione powyżej.

### 6.3.2. Dostosowanie algorytmu indukcji reguł do specyfiki problemu

Zgodnie z przedstawionymi założeniami, dotyczącymi semantyki reguł, każdy zawarty w nich warunek elementarny musi być prostym warunkiem o postaci  $t=1$ . Warunek taki spełniają wszystkie geny opisywane przez termin  $t$ . Możemy powiedzieć, że interesuje nas indukcja wszystkich reguł, postaci:

**Jeżeli  $t_{i1}$  oraz  $t_{i2}$  oraz ... oraz  $t_{ik}$ , to  $d = l$ ,**

gdzie:  $\{t_{i1}, t_{i2}, \dots, t_{ik}\} \in T$ , a  $G_i$  jest grupą genów. Ponadto reguły muszą być statystycznie istotne. Użytkownik wskazuje grupy genów, dla których należy dokonać indukcji. W szczególności może to być jedna grupa, wówczas pozostałe geny tworzą grupę przykładów negatywnych.

Do zadania indukcji reguł użyty został, opracowany przez Stefanowskiego i Vanderpootena, algorytm Explore [285]. Umożliwia on znalezienie wszystkich koniunkcji warunków elementarnych, spełniających zdefiniowane przez użytkownika wymagania minimalnej jakości (standardowo wymagania te dotyczą minimalnej ogólności i minimalnej dokładności reguł). Explore rozpoczyna proces indukcji od reguły zawierającej jeden warunek elementarny. Następnie realizowana jest faza wzrostu, w której przesłanka rozszerzana jest o kolejne warunki elementarne, przy czym algorytm działa zgodnie ze strategią przeszukiwania wszerz, ponieważ wymagania minimalnej jakości może spełnić każda z tworzonych reguł. Po dodaniu kolejnego warunku elementarnego sprawdzane jest, czy aktualna postać reguły spełnia wymagania minimalnej jakości. Jeśli tak jest, reguła

dodawana jest do zbioru wynikowego. Oznacza to, że dla zbioru warunków elementarnych, zawartych w tej regule, oraz dla każdego jego nadzbioru poszukiwanie reguł zostało zakończone. Jeśli aktualna postać reguły nie spełnia wymagań minimalnej jakości, to algorytm sprawdza, czy jest szansa na to, aby w przyszłości reguła spełniła te wymagania (np. czy reguła pokrywa wystarczającą liczbę przykładów pozytywnych). Jeśli takiej szansy nie ma, to dla aktualnie rozważanej koniunkcji warunków elementarnych proces przeszukiwania reguł również jest przerwany. Podczas rozrostu reguły opisującej klasę decyzyjną  $X$  rozważane są jedynie warunki elementarne, pokrywające co najmniej jeden przykład z tej klasy.

Przed zastosowaniem algorytmu Explore do opisu grup genów za pomocą terminów GO konieczne było wprowadzenie do niego dwóch modyfikacji:

- jeżeli aktualnie tworzona reguła spełni wymóg statystycznej istotności, to jest ona dodawana do zbioru wynikowego, jednakże w przeciwieństwie do standardowej wersji algorytmu koniunkcja warunków elementarnych, tworzących przesłankę reguły, jest dalej rozszerzana; motywacją takiego działania jest chęć znalezienia jeszcze innych, bardziej specyficznych reguł, również spełniających wymóg statystycznej istotności; do oceny statystycznej istotności reguł użyto miary  $p_{val}$  (3.14);  $p_{val}$  jest miarą kosztu i zgodnie z ideą testu dokładnego Fishera charakteryzuje się własnościami  $M_p$ ,  $M_n$  (odpowiednio dostosowanymi do miar kosztu –  $M_p$  oznacza funkcję malejącą,  $M_n$  – funkcję rosnącą); dla ustalonego zbioru przykładów można określić minimalną liczbę przykładów pozytywnych, jaką musi pokrywać reguła, aby jej  $p_{val}$  była mniejsza od progu zadanego przez użytkownika; to spostrzeżenie stanowi warunek zatrzymania indukcji dla reguł, które nie mają szans na osiągnięcie zadanej wartości  $p_{val}$ ;
- dodanie warunku elementarnego do przesłanki jest równoważne z dodaniem do niej terminu GO; w zmodyfikowanej wersji algorytmu Explore po dodaniu terminu  $t$  do przesłanki  $\varphi$  żaden inny termin, znajdujący się na dowolnej ścieżce, zawierającej termin  $t$  i prowadzącej od korzenia do dowolnego liścia w ontologii, nie jest rozpatrywany jako kandydat do rozszerzenia przesłanki  $\varphi$ ; zgodnie z definicją relacji  $\leq$  dla dowolnych terminów  $t_1, t_2 \in T$ , takich że  $t_1 \leq t \leq t_2$ , zachodzi bowiem następujące zależności: koniunkcja  $\varphi \wedge t_2 \wedge t$  sprowadza się do wyrażenia  $\varphi \wedge t$ , gdyż  $t$  znajduje się niżej w ontologii niż  $t_2$ , zatem  $G^t \subseteq G^{t_2}$ ; oznacza to, że przesłanki te pokrywają identyczną liczbę przykładów, a warunek (termin)  $t_2$  jest redundantny; podobną argumentację można przeprowadzić dla formuły  $\varphi \wedge t \wedge t_1$ , która sprowadza się do koniunkcji  $t_1 \wedge \varphi$ ; koniunkcja ta będzie rozpatrywana, gdy algorytm rozpoczęcie budowę przesłanki od terminu  $t_1$ .

Druga z przedstawionych modyfikacji znacznie skraca czas działania algorytmu oraz, przede wszystkim, nie dopuszcza do indukcji reguł, których przesłanki zawierają wzajemnie powiązane terminy GO. Jak pokazano w pracy [257], takie rozwiązanie pozwala na uniknięcie niedogodności, jakimi charakteryzują się metody indukcji, niebiorące pod uwagę hierarchii terminów ontologicznych [44,45,115].

### 6.3.3. Kryteria i sposób oceny reguł

Podstawowym kryterium oceny wyznaczanych reguł jest ich statystyczna istotność. Po wyznaczeniu wszystkich reguł ich p-wartości korygowane są za pomocą procedury redukcji fałszywych odkryć (patrz: podrozdział 3.1.8).

Dla standardowo stosowanych poziomów istotności statystycznej – 0.05 i 0.01 – liczba wyznaczanych reguł jest bardzo duża, w związku z tym konieczne było opracowanie pewnych procedur i miar pozwalających na wybór reguł najbardziej interesujących. Poza miarami obiektywnymi do oceny reguł użyto także dwóch miar o charakterze subiektywnym.

Pierwszą z miar subiektywnych jest znormalizowana długość reguły. W rozważanym przykładzie miara oceniająca długość reguły nie jest jednak traktowana jako miara kosztu, ale jako miara korzyści. Wynika to z założenia, że im więcej terminów jednocześnie opisuje daną grupę genów, tym opis ten jest bogatszy i bardziej użyteczny dla eksperymentatora. Oczywiście w praktycznych zastosowaniach użytkownik przed uruchomieniem procesu indukcji określa maksymalną liczbę terminów, jaka może się pojawić w przesłance reguły. Ponieważ ogólna liczba terminów tworzących tablicę decyzyjną jest duża (może wynosić nawet kilka tysięcy terminów), wartości użytej miary są normalizowane przez liczbę terminów, jaka znajduje się w przesłance najdłuższej z reguł opisujących daną grupę genów (6.3):

$$nL^{GO}(r) = \frac{Ec(r)}{\max\{Ec(r'): r' \in RUL_r\}}. \quad (6.3)$$

We wzorze (6.3)  $Ec(r)$  oznacza liczbę terminów GO, zawartych w przesłance  $r$ ,  $RUL_r$  oznacza zbiór wszystkich reguł opisujących tę samą grupę genów co reguła  $r$ .

Druga miara subiektywna pod uwagę bierze poziom, na jakim w ontologii znajdują się terminy zawarte w przesłance reguły. Im jest to poziom niższy, tym regułę należy ocenić wyżej, jako zawierającą bardziej specyficzną i dokładną wiedzę biologiczną. Poziom w ontologii definiowany jest za pomocą relacji  $\leq$ . W grafie ontologii  $i$ -ty poziom stanowią wszystkie terminy  $t \in T$ , dla których istnieje ścieżka  $path = (root, t_1, \dots, t_{i-1}, t_i = t)$ , taka że  $t_1 \leq root$ ,  $t_m \leq t_{m-1}$ ,  $m = 2, \dots, i-1, i$ . Dodatkowo ścieżka  $path$  jest najkrótszą z tych, jakie prowadzą od korzenia ontologii (ozn.  $root$ ) do terminu  $t$ . Oznacza to, że korzeń grafu

ontologii znajduje się na zerowym poziomie. Miarę biorącą pod uwagę położenie terminów w grafie ontologii nazwano *depth* (6.4):

$$\text{depth}(r) = \frac{\sum_{t \in Ec(r)} \text{level}(t)}{\sum_{t \in Ec(r)} \text{max\_path}(t)}. \quad (6.4)$$

We wzorze (6.4) *level(t)* to poziom w ontologii, na jakim znajduje się termin *t*, natomiast *max\_path(t)* to długość najdłuższej ścieżki, prowadzącej od korzenia ontologii, poprzez termin *t* aż do liścia. Miara *depth* jest miarą korzyści.

Zauważmy, że obie przedstawione miary przyjmują wartości w zbiorze  $[0,1]$  oraz są monotoniczne. Dodanie terminu do przesłanki powoduje wzrost wartości miary  $nL^{GO}$ . Obniżenie poziomu któregokolwiek z terminów  $t \in Ec(r)$ , przy niezmienionych poziomach pozostałych terminów należących do  $Ec(r)$ , powoduje wzrost wartości miary  $\text{depth}(r)$ .

Zdefiniowane miary subiektywne stanowią dwa składniki z trzech składników złożonej miary jakości (6.5), oceniającej wyznaczone reguły:

$$q_{\text{Compl}}^{GO}(r) = q(r) \cdot nL^{GO}(r) \cdot \text{depth}(r). \quad (6.5)$$

Miara *q* może być dowolną miarą z obiektywnych miar jakości, definiowanych na podstawie tablicy kontyngencji. Pożądane jest, aby *q* charakteryzowała się własnościami  $M_p$ ,  $M_n$ ,  $D_{in}$ ,  $D_>$ ,  $D_<$  oraz aby jej wartości zawierały się w przedziale  $[0,1]$  lub  $[-1,1]$ .

Miara złożona jest podstawą do utworzenia rankingu reguł. Ranking ten wykorzystany jest przez odmianę algorytmu filtracji *Coverage*. Jeśli zbiór przykładów pozytywnych, pokrywanych przez regułę *rr*, jest podzbiorem zbioru przykładów pokrywanych przez regułę *r*, to reguła *rr* traktowana jest jako zbędna (linie 8 i 10). Jeżeli jednak reguła *rr* jest niepodobna do reguły *r*, wówczas pozostaje ona w opisie (linia 9). Przedstawiona idea filtracji opiera się na założeniu, że reguły pokrywające podobne zbiory przykładów pozytywnych mogą opisywać te zbiory w różny sposób (za pomocą różnych koniunkcji terminów GO).

Algorytm filtracji zbioru reguł opisujących grupy genów

Wejście:  $G$  – zbiór genów,  $q_{\text{Complex}}$  – złożona miara jakości reguł  
 $\text{RUL}_G$  – zbiór reguł opisujących grupę genów  $G$   
 $\varepsilon$  – graniczny poziom podobieństwa reguł

Wyjście:  $\text{filtrRUL}$  – przefiltrowany zbiór reguł

```

1. Begin
2.    $\text{filtrRUL} := \emptyset;$ 
3.   While ( $G \neq \emptyset$  and  $\text{RUL}_G \neq \emptyset$ ) do
4.     wybierz z  $\text{RUL}_G$  dowolną regułę  $r$  o najwyższej wartości miary  $q_{\text{Complex}}$ 
5.      $\text{RUL}_G := \text{RUL}_G - \{r\};$ 
6.      $\text{filtrRUL} := \text{filtrRUL} \cup \{r\};$ 

    // usuwanie reguł „pokrywanych” przez regułę  $r$ 
7.     Foreach  $rr \in \text{RUL}_G$  do
8.       If  $[rr] \subseteq [r]$  then

        // nie są usuwane reguły, które są wystarczająco niepodobne do  $r$ 
9.         If  $\text{sim}(rr, r) < \varepsilon$  then  $\text{filtrRUL} := \text{filtrRUL} \cup \{rr\};$ 
10.         $\text{RUL}_G := \text{RUL}_G - \{rr\};$ 
11.      end foreach
12.       $G := G - [r];$ 
13.    end while
14.  end.
```

Do obliczenia podobieństwa reguł użyto miary (3.32), która jest odmianą miary Jaccarda (3.29). Dla rozważanego zastosowania najważniejszy jest sposób obliczenia wartości  $a, b, c, d$  w tablicy kontyngencji, służącej do obliczenia podobieństwa reguł. Miarę podobieństwa definiuje wzór (6.6), który po niewielkich przekształceniach można sprowadzić do (3.32):

$$\text{sim}_{GO}(r_i, r_j) = 1 - \frac{uGOterms(r_i, r_j) + uGOterms(r_j, r_i)}{Ec(r_i) + Ec(r_j)}. \quad (6.6)$$

We wzorze (6.6) przez  $uGOterms(r_i, r_j)$  oznaczono liczbę terminów zawartych jedynie w  $r_i$  i niezawartych w  $r_j$ , przy czym termin  $t \in Ec(r_i)$  występuje unikalnie w  $r_i$ , jeśli nie tworzy bezpośrednio przesłanki  $r_j$  oraz w grafie ontologii nie istnieje żadna ścieżka prowadząca od korzenia do dowolnego liścia, zawierająca jednocześnie termin  $t$  i jakikolwiek z terminów należących do  $Ec(r_j)$ .

Przedstawiony sposób indukcji, oceny i selekcji reguł tworzy jądro obliczeniowe serwisu internetowego RuleGO ([www.rulego.polsl.pl](http://www.rulego.polsl.pl)) [116]. Głównym wykonawcą i administratorem serwisu jest dr inż. Aleksandra Gruca z Instytutu Informatyki Politechniki Śląskiej. Serwis cały czas ewoluje i rozszerza swoją funkcjonalność.

Jak już wspominano w podrozdziale 3.2.5, w chwili obecnej autor wspólnie z Grucą prowadzi prace nad zastosowaniem metody UTA do identyfikacji złożonej miary ocenяjącej jakość reguł. Zadaniem miary złożonej jest dokładniejsze, niż czyni to miara (6.5) odzwierciedlenie specyficznych preferencji użytkownika. Pierwsze wyniki badań opublikowano w [117]. Obecnie przygotowywana jest większa publikacja, zawierająca

m.in. wytyczne dotyczące sposobu wyboru referencyjnego zbioru reguł, na podstawie którego następuje identyfikacja miary złożonej.

#### **6.3.4. Metoda redukcji atrybutów, biorąca pod uwagę semantykę ich wartości**

Wyznaczone reguły opisują pewne podgrupy interesującej nas grupy genów  $G$ . Reguły, reprezentują lokalne zależności pomiędzy terminami GO a genami z grupy  $G$ . Chcąc popatrzeć globalnie na problem opisu  $G$  za pomocą terminów GO, powinniśmy wskazać te terminy, które pełnią jakąkolwiek rolę w opisie genów należących do tej grupy. W szczególności interesujące jest znalezienie minimalnego zbioru  $M \subseteq T$ , złożonego z terminów GO, pozwalających z taką samą dokładnością jak cały zbiór  $T$  opisać geny należące do  $G$ . Problem ten jest podobny do problemu poszukiwania minimalnego reduktu względnego w tablicy decyzyjnej [220, 270].

Załóżmy, że dany jest zbiór genów  $G$ , podzielony na dwie rozłączne grupy:  $G_1$  i  $G_2$ . Przyjęte założenia pozwalają na zdefiniowanie tablicy decyzyjnej  $\mathbf{DT} = (G, T \cup \{d\})$ , w której atrybut decyzyjny  $d$  wskazuje na przynależność genów do grupy  $G_1$  bądź  $G_2$ . W iloczynie kartezjańskim  $G^2$  relacja nieroróżnialności  $IND$ , wyznaczana przez zbiór terminów  $B$ , zdefiniowana jest jako  $IND(B) = \{(g, g') \in G^2 : \forall t \in T t(g) = t(g')\}$ . Ponieważ  $IND(B)$  jest relacją równoważności, można dla niej wyznaczyć klasy abstrakcji  $[g]_{IND(B)}$  i zbiór ilorazowy  $G / IND(B)$ , stanowiący podział zbioru  $G$ . Zbiór  $Disc_{G_1} = \bigcup_{g \in G_1} [g]_{IND(T)}$  zawiera wszystkie geny ze zbioru  $G$ , należące do klas abstrakcji, których reprezentantami są geny należące do  $G_1$ . Jeśli tablica  $DT$  jest niesprzeczna, co oznacza, że  $\forall g \in G_1 [g]_{IND(T)} \subseteq G_1$ , to  $Disc_{G_1} = G_1$ . Jeśli tablica jest sprzeczna, to  $G_1 \subset Disc_{G_1}$ . W tej drugiej sytuacji można wyznaczyć  $T$ -dolne przybliżenie zbioru  $G_1$ , definiowane jako  $\underline{TG}_1 = \{g \in G_1 : [g]_{IND(T)} \subseteq G_1\}$ . Do zbioru  $\underline{TG}_1$  należą te geny, które w sposób jednoznaczny można przyporządkować do zbioru  $G_1$  na podstawie informacji zawartej w zbiorze terminów  $T$ . W teorii zbiorów przybliżonych interesujące jest znalezienie minimalnego zbioru terminów  $M \subseteq T$ , spełniającego warunek  $TG_1 = MG_1$ . Problem ten rozwiązywany jest przez wyznaczenie najkrótszego reduktu względnego [220, 270]. Ze względu na złożoność obliczeniową, związaną z poszukiwaniem najkrótszego reduktu, w praktyce poszukuje się reduktu quasi-minimalnego [208]. Dla rozważanego przez nas zastosowania konieczne jest wprowadzenie pewnych modyfikacji do procedury poszukiwania quasi-minimalnego zbioru terminów. Zakładając, że w danej chwili poszukujemy zbioru terminów  $M$  dla grupy  $G_1$ , najbardziej interesujące są te terminy GO, które opisują geny z  $G_1$  i nie opisują genów z  $G_2$ . Sytuacja odwrotna – termin opisuje geny

z  $G_2$  i nie opisuje genów z  $G_1$ , nie jest interesująca, pomimo że zapewnia ona odróżnialność genów z  $G_1$  i z  $G_2$ .

Jeżeli przez  $T_{\{1\}G_1}$  oznaczymy zbiór terminów, w którym  $\forall t \in T, \exists g_i \in G_1, \exists g_j \in G_2, (t(g_i) = 1 \text{ oraz } t(g_j) = 0)$ , to zgodnie z przyjętymi wcześniej założeniami poszukujemy najmniejszego zbioru  $M_{\{1\}G_1}$ , spełniającego warunek  $M_{\{1\}G_1}G_1 = T_{\{1\}G_1}G_1$ . Można wykazać, że  $\forall B \subseteq T, B_{\{1\}G_1}G_1 \subseteq BG_1$ . Znalezienie zbioru  $M_{\{1\}G_1}$  może być zrealizowane według następującego schematu, który jest modyfikacją algorytmu poszukiwania quasi-minimalnego reduktu względnego [208]:

1. Dane są zbiory  $G_1, G_2$  oraz  $T_{\{1\}G_1}G_1, X := T_{\{1\}G_1}, G_1^* := G_1, M := \emptyset$ .
2. Dla każdego  $t \in X$  :
  - 2.1. wyznacz  $W_{G_1 \{1\}G_2 \{0\}}^t$  liczbę wszystkich par genów  $\langle g_i, g_j \rangle$ , takich że  $g_i \in G_1, g_j \in G_2$  oraz  $t(g_i) = 1, t(g_j) = 0$ ,
  - 2.2. wyznacz  $W_{G_1 \{1\}G_2 \{1\}}^t$  liczbę wszystkich par genów  $\langle g_i, g_j \rangle$ , takich że  $g_i \in G_1, g_j \in G_2$  oraz  $t(g_i) = 1, t(g_j) = 1$ .
3. Znajdź  $t^{max} \in X$ , który maksymalizuje wartość  $W^t = W_{G_1 \{1\}G_2 \{0\}}^t - W_{G_1 \{1\}G_2 \{1\}}^t$ .
4.  $\forall g_i \in G_1$ , jeśli  $t^{max}(g_i) = 1$ , to  $G_1 := G_1 - \{g_i\}$ .
5.  $\forall g_j \in G_2, \forall t \in X$ , jeśli  $t^{max}(g_j) = 0$ , to podstaw  $t(g_j) = 0$ .
6.  $M := M \cup \{t^{max}\}, X := X - \{t^{max}\}$ .
7. Jeśli  $M_{\{1\}G_1}G_1^* \neq T_{\{1\}G_1}G_1^*$ , idź do punktu 2.
8. Koniec.

W punkcie 3 wybierany jest termin aktualnie najlepiej odróżniający geny z grupy  $G_1$  od genów z  $G_2$ , przy czym odróżnialność zdefiniowana jest w interesujący nas sposób ( $t(g_i) = 1, t(g_j) = 0$ ). W punkcie 5 następuje uaktualnienie tablicy decyzyjnej, tak aby w kolejnym kroku algorytmu wybrać następny termin GO, najlepiej odróżniający nieodróżnione jeszcze geny.

Jeśli w analizowanym zbiorze przykładów wyróżnionych jest kilka grup genów, to przedstawiony sposób analizy należy wykonać oddziennie dla każdej grupy. Grupa  $G_1$  jest grupą referencyjną, natomiast  $G_2$  jest sumą mnogościową pozostałych grup. Indukcja reguł w zredukowanej tablicy decyzyjnej przebiega znacznie szybciej niż indukcja reguł w tablicy niezredukowanej. Wyniki badań nad efektywnością połączenia obu technik (redukcja terminów i indukcji reguł w zredukowanej tablicy) przedstawiono w [257].

### 6.3.5. Analiza danych

Do badania efektywności przedstawionych propozycji wybrano dwa zbiory, funkcjonujące w środowisku bioinformatycznym jako zbiory benchmarkowe. Zbiór *yeast* zawiera wyniki ekspresji drożdży piekarskich (*Saccharomyces cerevisiae*) [74], a zbiór *human* – wyniki ekspresji komórek fibroblastów ludzkich [140]. Do eksperymentów wybrano przetworzone i odpowiednio przygotowane [74, 140] reprezentacje tych zbiorów. Z każdym z nich związano dwie tablice decyzyjne: jedna zawierała jedynie terminy zawarte w ontologii *BP*, druga – terminy ze wszystkich ontologii (ozn. *ALL*). W przeprowadzonych badaniach tablica decyzyjna *yeastBP* składała się z 274 genów opisanych 249 terminami GO; tablica *yeastALL* składała się z 274 genów opisanych 418 terminami GO; tablica *humanBP* zawierała 309 genów opisanych 390 terminami GO; wreszcie tablica *humanALL* zawierała 364 geny opisane 588 terminami GO.

W pracach [255,257] przytoczono wyniki uzyskane oddziennie dla każdej z 10 grup, na jakie podzielono zbiory genów. Poniżej przytoczone zostaną jedynie wyniki uśrednione. W tabeli 6.10 zaprezentowano, jak zastosowanie filtracji wpływa na reducję liczby reguł oraz pokrycie klas decyzyjnych. Algorytm indukcji skonfigurowano w taki sposób, żeby wyznaczane były reguły, dla których  $p_{val} \leq 0.01$ , a w przesłance mogło znajdować się maksymalnie 5 terminów GO. W indukcji używano terminów znajdujących się na co najmniej drugim poziomie ontologii. Graniczny poziom podobieństwa reguł  $\varepsilon$  ustalono na 0.5.

Tabela 6.10

Wyniki indukcji reguł dla problemu opisu grup genów za pomocą reguł decyzyjnych (prezentowane są zaokrąglone do jedności wartości średnie)

Zbiór danych	Średnia liczba reguł bez filtracji	Średnia liczba reguł z filtracją	Średnie pokrycie grup genów [%]
<i>yeastBP</i>	13 547	10	94
<i>yeastALL</i>	634 195	23	99
<i>humanBP</i>	6 218	15	47
<i>yeastALL</i>	41649	36	75

Jak widać, algorytm filtracji pozwolił na znaczne ograniczenie liczby reguł. Miary jakości i podobieństwa zastosowane w tym algorytmie zapewniają, że w wynikowym zbiorze znalazły się reguły interesujące dla użytkownika. Analiza wiedzy biologicznej, zawartej w wyznaczonych regułach, wychodzi poza zakres niniejszej monografii; analizę taką można znaleźć w [255, 257].

W dalszej części eksperymentów sprawdzono, jaka jest efektywność metody redukcji terminów GO. Metodę zastosowano do tablic *yeastALL* oraz *humanALL*.

Po redukcji każdą grupę genów w zbiorze *yeastALL* opisywało średnio 38 terminów (w tablicy niezredukowanej było to 418 terminów). Dolne przybliżenie każdej z grup genów

za pomocą zredukowanego zbioru terminów stanowiło średnio 95% wszystkich genów należących do danej grupy. Jest to spadek o 5% w stosunku do całego rozważanego zbioru terminów i wynika on ze sposobu wyznaczania zbiorów  $M_{\{1\}G_i} G_i$  oraz z zależności

$$M_{\{1\}G_i} G_i \subseteq MG_i.$$

Po redukcji zbioru *humanALL* każda z grup genów opisywana była średnio przez 221 terminów (w tablicy niezredukowanej było to 588 terminów). Dolne przybliżenie każdej z grup genów za pomocą zredukowanego zbioru terminów stanowiło średnio 85% wszystkich genów należących do danej grupy.

W dalszej części eksperymentu przeprowadzono indukcję reguł w zredukowanych tablicach decyzyjnych. Po indukcji w zbiorze *yeastALL* utworzono dla każdej grupy średnio 8 reguł pokrywających średnio 96% wszystkich genów w każdej grupie. Dla zbioru *humanALL* utworzono średnio 16 reguł i pokrycie 66%. Porównując te wyniki z wynikami zawartymi w tabeli 6.10, można zauważać, dalszą znaczną redukcję liczby reguł przy jednoczesnym umiarkowanym spadku pokrycia, wynoszącym dla zbioru *hmunaALL* 9%.

Przesłanki reguł wyznaczonych na podstawie zredukowanych tablic zawierają najbardziej informatywne terminy GO z punktu widzenia opisu danej grupy. Analiza zredukowanego zbioru terminów bez wyznaczania reguł może stanowić uzupełniającą informację na temat funkcji genów tworzących daną grupę. W analizie tej użyteczna może być również informacja o kolejności dodawania terminów do zbioru  $M$ . Terminy dodawane do  $M$  na początku najlepiej odróżniają geny należące do grupy referencyjnej od genów z pozostałych grup.

Kolejny eksperyment polegał na porównaniu otrzymanych reguł z regułami wyznaczonymi przez serwis internetowy Genecodis. Do serwisu przekazano zbiór danych *yeast*, a proces indukcji reguł w obu algorytmach skonfigurowano tak, aby był on maksymalnie podobny (oba algorytmy charakteryzują się nieco innymi parametrami) [257]. Rezultaty eksperymentu przedstawiono w tabeli 6.11.

Tabela 6.11  
Porównanie efektywności reguł utworzonych przez serwis Genecodis  
i modyfikowaną wersję algorytmu Explore, uzupełnioną o algorytm filtracji

Zbiór danych	Algorytm	Średnia liczba reguł z filtracją	Średnie pokrycie grup genów [%]
<i>yeastBP</i>	<i>Genecodis</i>	6	57
<i>yeastBP</i>	<i>ModExplore+filtracja</i>	10	94
<i>yeastALL</i>	<i>Genecodis</i>	30	83
<i>yeastALL</i>	<i>ModExplore+filtracja</i>	23	99

Przefiltrowane zbiory reguł, wygenerowane za pomocą zmodyfikowanej wersji algorytmu Explore, charakteryzują się cechami znacznie lepszymi niż zbiory reguł

generowane przez Genecodis. Ponadto, jak pokazano w [257], wśród reguł wyznaczonych przez Genecodis znajdują się takie, które zawierają powiązane (hierarchia w ontologii) terminy. Reguły generowane przez Explore nie zawierają takich terminów.

Ostatni etap analizy danych polegał na zastosowaniu do reguł, będących rezultatem filtracji, algorytmu redefinicji, opisanego w podrozdziale 5.2. Do oceny ważności terminów użyto zmodyfikowanej postaci indeksu Banzhafa (5.5). Jako miarę jakości wykorzystano RSS. W algorytmie redefinicji można dopuścić do wystąpienia w przesłankach reguł zanegowanych warunków elementarnych. Warunek taki przyjmuje postać  $t = 0$  i oznacza, że termin nie opisuje spełniających go genów. W regułach opisujących grupę  $G$  zanegowane warunki elementarne powstaną na podstawie terminów GO najmniej informatywnych dla tej grupy i najbardziej informatywnych dla pozostałych grup. Zauważmy, że jeśli w regule znajduje się warunek  $t = 0$ , to grupy tej nie opisuje również żaden z terminów  $t'$ , takich że  $t' \leq t$  (znajdujących się na niższych poziomach ontologii). Informacja taka może być w szczególnych przypadkach interesująca dla biologa. Dopuszczenie zanegowanych warunków elementarnych pozwoliło na zwiększenie pokrycia zbioru *human* (tabela 6.12), jednakże (jak przedstawiono to w pracy [255]) interpretacja biologiczna utworzonych w ten sposób reguł jest trudna.

Zastosowanie algorytmu redefinicji pozwala na znaczne ograniczenie liczby reguł oraz na wzrost pokrycia klas decyzyjnych (tabela 6.12). Mając na uwadze, że indukcję przeprowadzano dla celów opisowych, trzeba pamiętać, że redukcja liczby reguł nie zawsze jest efektem pożądanym. W związku z tym, tworząc opisy grup genów za pomocą terminów GO, można rekomendować następujący schemat postępowania: wyznacz reguły za pomocą zmodyfikowanej wersji algorytmu Explore; dokonaj redukcji reguł za pomocą algorytmu filtracji; jeśli liczba reguł po filtracji jest w dalszym ciągu zbyt duża i/lub pokrycie grupy genów jest niewystarczające, zastosuj algorytm redefinicji bez zanegowanych warunków elementarnych; jeśli pokrycie jest w dalszym ciągu niezadowalające, zastosuj algorytm redefinicji z zanegowanymi warunkami elementarnymi.

Tabela 6.12

Wyniki redefinicji reguł na podstawie informacji o ważności terminów GO, znajdujących się w przesłankach reguł  
(podawane są zaokrąglone do jedności wartości średnie dla 10 grup)

Zbiór danych	mExplore+filtracja		Redefinicja bez warunków negatywnych		Redefinicja z warunkami negatywnymi	
	Liczba reguł	Pokrycie [%]	Liczba reguł	Pokrycie [%]	Liczba reguł	Pokrycie [%]
<i>yeastBP</i>	10	94	3	97	2	96
<i>humanBP</i>	15	47	5	48	6	88

## **7. PODSUMOWANIE**

W monografii omówiono problematykę oceny jakości reguł decyzyjnych, przedstawiono także wybrane zagadnienia związane z przycinaniem zbiorów reguł i praktycznymi zastosowaniami algorytmów indukcji reguł decyzyjnych.

Główne cele pracy, związane z miarami jakości, to: przedstawienie szerokiego spectrum obiektywnych miar, umożliwiających ocenę reguł decyzyjnych; usystematyzowanie informacji na temat miar definiowanych na podstawie tablicy kontyngencji; określenie zależności pomiędzy miarami; wybór minimalnego zbioru własności, jakimi miary te powinny się charakteryzować; przedstawienie propozycji złożonej oceny reguł. Celem badań empirycznych było wskazanie zbioru miar najbardziej efektywnych oraz zidentyfikowanie grup miar podobnych ze względu na porządek reguł. Nadrzędnym celem badań nad metodami przycinania było zaproponowanie metod umożliwiających redukcję liczby reguł tworzących klasyfikator. Efekt ten starano się uzyskać przez wprowadzenie do przesłanek złożonych warunków elementarnych oraz eliminację reguł zbędnych.

Przykłady praktycznych zastosowań ilustrują możliwość przeniesienia uzyskanych wyników na realne problemy, związane z indukcją reguł decyzyjnych dla celów klasyfikacji i opisu. Przykłady te pokazują również, że rozwiązywanie praktycznych zadań wymaga definiowania dziedzinowo zorientowanych modyfikacji znanych metod indukcji i oceny reguł.

Rozdział 2 w większości miał charakter przeglądowy; omówiono w nim: metody indukcji reguł decyzyjnych, definiowanych dla celów klasyfikacyjnych i opisowych, oraz najważniejsze kierunki badań i problemy związane z tymi zagadnieniami.

W rozdziale 3 przedstawiono miary oceny jakości reguł decyzyjnych. W sposób szczególny skoncentrowano się na własnościach miar obiektywnych, definiowanych na podstawie tablicy kontyngencji. Zdefiniowano także trzy rodzaje równoważności miar oraz przedstawiono założenia umożliwiające badanie podobieństwa miar ze względu na porządek reguł. W rozdziale 3 zaprezentowano także możliwości złożonej oceny reguł, w której pod uwagę branych jest jednocześnie kilka kryteriów składowych. W dodatku A umieszczono dwu- i trójwymiarowe wykresy wybranych miar.

Rozdział 4 prezentuje wyniki badań eksperymentalnych i teoretycznych nad efektywnością i własnościami miar definiowanych na podstawie tablicy kontyngencji. W pierwszej części rozdziału badano, jaki wpływ na zdolności klasyfikacyjne i złożoność klasyfikatora regułowego ma miara jakości, nadzorująca proces indukcji reguł. Badania te stały się podstawą do zidentyfikowania zbioru miar najbardziej efektywnych oraz przedstawienia adaptacyjnej metody dobru miary w pokryciowym algorytmie indukcji reguł. W ramach badań teoretycznych przeanalizowano 34 miary jakości w kierunku posiadanych przez nie własności oraz ich wzajemnej równoważności.

Rozdział 5 zawiera opis kilku algorytmów przycinania zbiorów reguł. Przycinanie realizowane jest na trzy sposoby – przez: agregację (łączenie), redefinicję oraz filtrację reguł. Zadaniem wszystkich trzech metod jest ograniczenie liczby reguł opisujących dane i biorących udział w klasyfikacji. Pierwsze dwie metody dokonują zmiany reprezentacji warunków elementarnych. Dzięki temu zabiegowi wynikowe reguły mogą opisywać bardziej złożone zależności niż reguły zawierające jedynie proste warunki elementarne.

W rozdziale 6 przedstawiono przykłady praktycznych zastosowań algorytmów indukcji reguł decyzyjnych. Przedstawiono trzy nowe obszary zastosowań. Dwa z nich związane są z budową klasyfikatorów regułowych (prognozowanie zagrożeń naturalnych, analiza danych okołoprzeszczepowych), a trzeci związany jest z indukcją reguł jedynie dla celów opisowych (funkcjonalny opis grup genów). W rozdziale tym zaprezentowano dwie modyfikacje algorytmów indukcji reguł. Pierwsza umożliwia wykorzystanie zbioru rekomendacji

i hipotez, definiowanych przez użytkownika do nadzorowania procesu indukcji reguł. Druga jest ukierunkowana na konkretną dziedzinę zastosowań i w trakcie indukcji reguł wykorzystuje informacje na temat hierarchicznej struktury atrybutów warunkowych.

Najważniejsze rezultaty uzyskane w ramach prowadzonych badań zostały streszczone w następujących punktach.

1. Analiza własności miar definiowanych na podstawie tablicy kontyngencji wykazała, że dla miar przeznaczonych do oceny reguł decyzyjnych pożądane są dwie grupy własności. Pierwszą grupę tworzą własności  $M_p$ ,  $M_n$ , które zapewniają, że miara nadzorująca proces indukcji lub selekcji reguł preferuje reguły pokrywające jak najwięcej przykładów pozytywnych i jak najmniej przykładów negatywnych. Drugą grupę tworzą własności przeznaczone dla miar oceniających zdolności opisowe reguł. Podstawowymi własnościami należącymi do tej grupy są  $D_{in}$ ,  $D_>$ ,  $D_<$ . Jako własności uzupełniające rekomendować można  $D_{eq}$ ,  $D_{ES}$ ,  $D_{HS}$ ,  $D_{WL}$  oraz  $D_{mwEx1}$ .
2. Najsilniejszym rodzajem równoważności miar jest równoważność ze względu na sposób rozstrzygania konfliktów klasyfikacji przez głosowanie. Miary równoważne w ten sposób identycznie porządkują zbiór reguł (bezwzględna równoważność) oraz identycznie

rozstrzygają konflikty klasyfikacji. Słabszym typem równoważności jest równoważność ze względu na sposób rozstrzygania konfliktów zgodnie z zasadą największego zaufania. Miary równoważne w ten sposób identycznie porządkują zbiór reguł (bezwzględna równoważność) oraz w identyczny sposób rozstrzygają konflikty klasyfikacji w schemacie największego zaufania. Najsłabszym rodzajem równoważności jest równoważność ze względu na porządek definiowany w zbiorze reguł wskazujących na identyczną klasę decyzyjną (w skrócie – porządek reguł).

3. Wprowadzone w rozdziale 3 stwierdzenia i uwagi dają podstawy do badania równoważności miar, pokazując powiązania pomiędzy różnymi typami równoważności, pozwalając także wnioskować o własnościach jednej mocy na podstawie własności miar do niej równoważnych. Umożliwia to stosunkowo prostą analizę i kategoryzację kolejnych miar jakości.
4. Badania podobieństwa miar ze względu na porządek reguł pozwoliły na identyfikację grup miar podobnych. W zależności od progu minimalnej korelacji wewnętrzgrupowej zidentyfikowano od 4 do 8 grup miar podobnych. Do grup tych należą mocy przywiązuje podobną wagę do oceny dokładności i pokrycia reguł. Przeprowadzona analiza porządków reguł pokazała, że żadna para z rozważanych miar jakości nie ocenia reguł w sposób całkowicie antagonistyczny.
5. Badania efektywności miar ze względu na trzy kryteria jakości (*Acc*, *BAcc*, *AvC*) pozwoliły na zidentyfikowanie grupy *Max*, złożonej z 9 najbardziej efektywnych miar. Grupę tę stanowią mocy: *C1*, *C2*, *Corr*, *g*, *LS*, *MS*, *RSS*, *s*, *wLap*. Badania efektywności quasi-minimalnych zbiorów miar nieprzegrywających ze względu na kryteria *Acc* i *BAcc* sugerują, aby zbiór *Max* rozszerzyć o moc *CN2*. Ze względu na całkowitą dokładność klasyfikacji najefektywniejsze są mocy *C1* i *C2*, a ze względu na średnią dokładność klas decyzyjnych (która jest powiązana z wrażliwością i specyficznością klasyfikatora) najefektywniejsza jest moc *wLap* oraz ponownie mocy *C1* i *C2*. Ze względu na wartość kryterium *AvC* najefektywniejsze są mocy *RSS* i *Corr*. Klasyfikatory zbudowane na podstawie *RSS* i *Corr* charakteryzują się statystycznie mniejszą liczbą reguł od klasyfikatorów otrzymanych na podstawie *C1* i *C2*. Równocześnie ich dokładność klasyfikacji jest statystycznie gorsza – przy porównaniu miar parami lub analizie liczby ich zwycięstw i porażek – od klasyfikatorów będących rezultatem zastosowania miar *C1* i *C2*. Mocy *Corr*, *RSS* i *s* mają największą liczbę własności pożądanych dla miar ocenujących zdolności opisowe reguł. Należy pamiętać o tym, aby do nadzorowania procesu indukcji reguł stosować zmodyfikowane postaci *Corr*, *CN2*, *LS*. Modyfikacje umożliwiają ocenę reguł dokładnych oraz pokrywających wszystkie przykłady treningowe.

6. Poza nielicznymi wyjątkami miary zawarte w zbiorze *Max* pozwalają na otrzymanie zwycięskiego (lub co najmniej nieprzegrywającego) klasyfikatora regułowego ze względu na każde z kryteriów jakości: *Acc*, *BAcc*, *AvC*. Potwierdzają to wyniki badania efektywności adaptacyjnej metody doboru miary, przeprowadzone na 34 treningowych i 14 testowych zbiorach danych o różnorodnej charakterystyce. Klasyfikatory uzyskane na podstawie miar złożonych, tworzonych w celu maksymalizacji wartości *Acc*, *BAcc*, *AvC*, nie są statystycznie lepsze od wyników najlepszych miar. Zdaniem autora dalsze badania nad określeniem miary złożonej, nadzorującej proces indukcji reguł, powinny być ukierunkowane na tworzenie miary złożonej, budowanej na podstawie miar zawartych w zbiorze *Max*, przy czym wybór konkretnych miar oraz sposób ich połączenia powinny być każdorazowo dopasowane do konkretnego, analizowanego zbioru przykładów.
7. Spośród miar niestosowanych bezpośrednio do nadzorowania indukcji reguł na szczególną uwagę zasługują:
  - miara  $p_{val}$ , dokonująca oceny statystycznej istotności reguły. Jest to miara kosztu, dla której podstawą jest prawostronny dokładny test Fishera. Obliczenie wartości  $p_{val}$  wiąże się z koniecznością obliczenia wartości dwumianów Newtona, w chwili obecnej nie stanowi to jednak większego problemu obliczeniowego. Wartość miary  $p_{val}$  w połączeniu z procedurą kontroli poziomu fałszywych odkryć stanowi użyteczną informację na temat jakości reguł definiowanych dla celów opisowych. Zgodnie z założeniami procedury kontroli fałszywych odkryć nie powinno się korygować wartości  $p_{val}$  w zbiorach reguł, otrzymanych za pomocą algorytmów pokryciowych;
  - miary  $E^\varphi$ ,  $E^\psi$ , umożliwiające pomiar siły interwencji, reprezentowanej przez regułę. Miara  $E^\varphi$  bliższa jest idei traktowania reguły jako quasi-równoważności. W ocenie wykorzystuje ona informacje zawarte we wszystkich komórkach tablicy kontyngencji. Wskaźnikiem uzupełniającym ocenę siły interwencji jest miara *Corr*, którą można stosować do oceny reguł interpretowanych jako quasi-równoważności. Oceniając regułę  $\varphi \rightarrow \psi$ , miara *Corr* bierze pod uwagę jej dokładność i pokrycie oraz dokładność i pokrycie reguły  $\neg\varphi \rightarrow \neg\psi$ . Miary  $E^\varphi$ ,  $E^\psi$ , *Corr* nie są równoważne. W sytuacji, w której  $E^\varphi$  i  $E^\psi$  oceniają reguły w sposób antagonistyczny, wartość *Corr* powinna decydować o wyborze tej interwencji, która jest potencjalnie najskuteczniejsza;
  - miary umożliwiające złożoną ocenę reguł. Stosowanie miar złożonych wymaga normalizacji tworzących je miar składowych. Szczególnego rodzaju miarą złożoną jest funkcja użyteczności. Jej identyfikacji dokonuje się na podstawie wartości miar składowych oraz porządku reguł, definiowanego przez użytkownika. Składnikami

funkcji użyteczności mogą być dowolne, również subiektywne, miary oceniające jakość reguł. W szczególności mogą to być miary będące reprezentantami grup miar podobnych, zidentyfikowanych w podrozdziale 4.3;

- zdaniem autora użytecznych informacji może dostarczyć także miara informująca o równomierności rozkładu przykładów pozytywnych, pokrywanych przez regułę. Zastosowanie jej do modyfikacji siły reguł w trakcie rozstrzygania konfliktów klasyfikacji będzie przedmiotem przyszłych badań.
8. Algorytmy agregacji ograniczają liczbę reguł biorących udział w klasyfikacji. W zależności od wartości parametrów algorytmu CHIRA poziom redukcji liczby reguł wynosi 15–20%. W zbiorach danych, w których przykłady reprezentujące klasy decyzyjne rozdzielane są przez hiperpłaszczyznę nierównoległe do osi atrybutów, redukcja liczby reguł jest znacznie większa. Wadą algorytmu CHIRA jest złożoność obliczeniowa, związana z koniecznością wyznaczenia otoczki wypukłej, zawierającej agregowane reguły. Dla zbiorów złożonych z setek reguł zawierających więcej niż 5 warunków elementarnych stosowanie algorytmu może być nieefektywne ze względu na czas obliczeń. W związku z czasem działania algorytmu CHIRA przed agregacją celowe może być ograniczenie liczby reguł za pomocą jednego z algorytmów filtracji, prezentowanych w podrozdziale 5.3.
9. Algorytm redefinicji reguł na podstawie informacji o ważności warunków elementarnych jest bardzo efektywny, gdyż zmniejsza liczbę reguł o około 40%. O około 36% maleje także liczba warunków elementarnych, tworzących reguły. Tak wysoki poziom redukcji udało się osiągnąć m.in. przez wprowadzenie do przesłanek reguł zanegowanych warunków elementarnych. Ograniczenie liczby reguł nie odbywa się kosztem dokładności klasyfikacji, dla niektórych parametrów algorytmu odnotowano nawet statystyczną poprawę dokładności klasyfikacji. Uproszczona metoda oceny warunków elementarnych pozwala na analizę dużo liczniejszych zbiorów reguł niż metoda standardowa. Analiza podobieństwa rankingów warunków elementarnych, tworzonych przez indeks Banzhafa uproszczony i podstawowy, wykazała, że rankingi te są bardzo podobne (współczynnik korelacji  $\tau$  Kendalla wynosi średnio 0.95).
10. Zaprezentowane algorytmy filtracji eliminują reguły, które są zbędne z punktu widzenia opisu i klasyfikacji przykładów. Najniższy poziom redukcji liczby reguł obserwujemy po zastosowaniu algorytmu *Inclusion*. Jeśli miarę jakości stosowano na każdym etapie budowy reguł oraz podczas klasyfikacji (metoda MMM), to średni poziom redukcji wynosi 18%. Jeśli miarę wykorzystano jedynie w etapach przycinania i klasyfikacji (metoda EMM), to redukcja wynosi 7%. Zastosowanie algorytmu *Inclusion* nie wpływa w sposób istotny na zdolności klasyfikacyjne wynikowego zbioru reguł. Średnia redukcja liczby reguł po zastosowaniu algorytmu *Coverage* wynosi 23% (MMM) i 13% (EMM).

Dla tego rodzaju filtracji oraz reguł wyznaczanych zgodnie z metodą MMM obserwujemy statystycznie istotny spadek dokładności klasyfikacji dla miar preferujących reguły o wysokiej dokładności ( $C1$ ,  $C2$ ,  $LS$ ,  $wLap$ ). Średnia redukcja liczby reguł po zastosowaniu algorytmu *Forward* wynosi 51% (MMM) i 28% (EMM). Zastosowanie tego algorytmu do zbiorów reguł wyznaczonych na podstawie miar *Corr*, *MS*, *RSS* powoduje podniesienie dokładności klasyfikacji klasyfikatorów utworzonych na podstawie tych reguł. Jeśli reguły wyznaczano zgodnie z ideą użycia miary, oznaczaną jako EMM, poprawa dokładności klasyfikacji jest statystycznie istotna. Redukcja liczby reguł po zastosowaniu algorytmu *Backwards* jest podobna do tej, jaką uzyskuje algorytm *Forward*. Ze względu na odnotowane spadki dokładności klasyfikacji oraz największe obniżenie zdolności rozpoznawania przykładów testowych stosowanie algorytmu *Backwards* nie wydaje się celowe. Zastosowanie algorytmów *Inclusion*, *Coverage* i *Forward* nie wpływa negatywnie na możliwość rozpoznania przykładów testowych. Po filtracji wzrasta średnie unikalne pokrycie reguł. Wzrost średniego poziomu statystycznej istotności reguł sugeruje także, że na etapie filtracji usuwane są reguły statystycznie najmniej istotne;

11. przykłady praktycznych zastosowań algorytmów indukcji reguł decyzyjnych pokazują, że przedstawione w monografii metody indukcji, oceny i przycinania przyczyniają się do poprawy zdolności opisowych i klasyfikacyjnych wyznaczanych reguł. Metoda indukcji reguł, sterowana hipotezami definiowanymi przez użytkownika, oraz metoda dostosowująca proces indukcji do hierarchicznej struktury danych pozwalają na wyznaczanie reguł interesujących i użytecznych dla użytkownika. Metoda redukcji atrybutów, sterowana semantyką ich wartości, pozwala na ukierunkowanie redukcji w taki sposób, aby klasa decyzyjna opisana była przez wartości, które są najbardziej interesujące dla użytkownika. Analiza danych okołoprzeszczepowych wykazała, że na podstawie reguł decyzyjnych możliwe jest definiowanie wzorców przeżycia o jakości zbliżonej do wzorców będących rezultatem indukcji drzew przeżycia.

Rezultaty badań przedstawionych w monografii zostały opublikowane lub oczekują na publikację w materiałach konferencyjnych (w większości wydawanych w ramach serii Lecture Notes in Computer Sciences) oraz w czasopismach o zasięgu krajowym i międzynarodowym (m.in. *Expert Systems with Applications*, *Fundamenta Informaticae*, *Fundations of Computing and Decision Sciences*, *International Journal of Applied Mathematics and Computer Sciences*, *International Journal of General Systems*, *Journal of Mining Sciences*, *Nucleic Acid Research*, *Pattern Recognition Letters*, *Transactions on Rough Sets*).

Autor wraz ze współpracownikami zamierza udostępnić szerszemu gronu użytkowników oprogramowanie, w którym zawarta będzie część z opisanych w niniejszej pracy metod

i algorytmów. Projektowane środowisko jest zintegrowane z pakietem obliczeń statystycznych R, gdyż umożliwia to m.in. definiowanie własnych miar jakości oraz korzystanie z wielu gotowych rozwiązań. Rozwój projektu można śledzić na stronie <http://crules.r-forge.r-project.org/>. Złożona ocena reguł na podstawie funkcji użyteczności włączona zostanie do przywoływanego w podrozdziale 6.3 serwisu RuleGO (<http://rulego.polsl.pl>).

Dalsze prace związane z miarami jakości i metodami przycinania reguł można podzielić według trudności zadań. Prostsze zadania będą polegać na przeniesieniu metodyki badania efektywności miar oraz ich adaptacyjnego doboru na pokryciowe algorytmy indukcji reguł regresyjnych i wzorców przeżycia. Jak pokazano w dodatku B, dzięki odpowiedniemu zdefiniowaniu pojęcia przykładu pozytywnego i pojęcia przykładu negatywnego możliwe jest zastosowanie omawianych w monografii miar jakości do nadzorowania indukcji reguł regresyjnych. Wykonane badania eksperimentalne uzasadniają celowość dalszego prowadzenia podjętych prac. Stosunkowo proste powinno być również przeniesienie na grunt reguł regresyjnych i wzorców przeżycia metod oceny ważności warunków elementarnych i filtracji zbiorów reguł. Reguły „regresyjne” z nieznacznie zmodyfikowanymi konkluzjami wydają się również obiecującym narzędziem do opisu funkcji genów. Jednym z istotnych problemów w tej analizie jest konieczność dysponowania prawidłowo zidentyfikowanymi grupami genów o podobnej ekspresji. Zamiast tworzyć takie grupy, można próbować wyznaczyć reguły, w konkluzjach których umieszczona będzie wartość korelacji pomiędzy szeregiem czasowym, określającym ekspresję genów. Idea identyfikowania przykładów pozytywnych i przykładów negatywnych dla takich reguł będzie podobna do idei przedstawionej dla reguł regresyjnych.

Dużo bardziej ambitnym kierunkiem badań jest zastosowanie algorytmów pokryciowych do indukcji reguł temporalnych (zwanych też regułami sekwencji [81]). W indukcji reguł temporalnych informacja na temat każdego przykładu treningowego reprezentowana jest przez sekwencję zdarzeń. W ramach każdego zdarzenia przykład opisany jest przez wektor atrybutów. Wystąpienie zdarzenia może wiązać się z upływem określonego czasu lub ze zmianą wartości opisujących go atrybutów. Każde zdarzenie może być przyporządkowane również do określonego stanu, a stany mogą być powiązane i tworzyć uporządkowaną strukturę (np. ontologię). Odkrywaniem i modelowaniem zachowania systemów złożonych, w których struktura stanów definiowana jest przez eksperta, zajmują się m.in. Bazan, Skowron i Stepaniuk [19, 212, 274, 275]; w szczególności zagadnieniu temu poświęcona jest rozprawa habilitacyjna Bazana [19].

Reguły temporalne można wykorzystać do odkrycia powiązań pomiędzy: wartościami atrybutów w następujących po sobie zdarzeniach lub stanach; wartościami atrybutów a stanem, w jakim znajduje się przykład. Połączenie indukcji reguł, nadzorowanej przez

użytkownika, z oceną ważności i siły interwencji warunków elementarnych powinno przyczynić się do wyznaczania zbiorów reguł, będących podstawą dla efektywnych (w tym możliwych do realizacji) strategii interwencji. Zadaniem strategii interwencji będzie osiągnięcie zamierzonego stanu lub niedopuszczenie do osiągnięcia stanów niepożądanych. Przedmiotem badań może być również analiza osiągalności pewnych stanów przy zadanych ograniczeniach i warunkach początkowych.

Możliwe obszary zastosowań to m.in. medycyna (nadzorowanie przebiegu terapii) oraz przemysł (monitorowanie zagrożeń naturalnych, monitorowanie stanu maszyn).

## BIBLIOGRAFIA

1. Abe H., Tsumoto S., Oshaki M., Yamaguchi T.: Evaluation Learning Algorithms to Construct Rule Evaluation Models Based on Objective Rule Evaluation Indices. IEEE – Conference on Cognitive Informatics 2007, s. 212÷221.
2. Abe H., Tsumoto S.: Analyzing Behavior of Objective Rule Evaluation Indices Based on Pearson Product-Moment Correlation Coefficient. Lecture Notes in Artificial Intelligence 4994, 2008, s. 84÷89.
3. Abe H., Tsumoto S.: Comparing accuracies of rule evaluation models to determine human criteria on evaluated rule sets. IEEE International Conference on Data Mining Workshops, 2008, s.1÷7.
4. Abe H., Tsumoto S.: A comparison of composed objective rule evaluation indices using PCA and single indices. Lecture Notes in Computer Science 5589, 2009, s. 160÷167.
5. Ågotnes T., Komorowski J., Loken T.: Taming large rule models in rough set approaches. Lecture Notes in Artificial Intelligence 1704, 1999, s. 193÷203.
6. Agrawal R., Srikant R.: Fast Algorithms for Mining Association Rules. Proc. of the 20<sup>th</sup> VLDB Conference, Santiago, Chile 2004.
7. Agresti A.: Categorical data analysis. Wiley Interscience, New Jersey 2002.
8. An A., Cercone N.: Rule quality measures for rule induction systems – description and evaluation. Computational Intelligence 17, 2001, s. 409÷424.
9. Ashburner M., Ball C.A., Blake J.A., Botstein D., Butler H., Cherry J.M. et al.: Gene Ontology: Tool for the Unification of Biology. The Gene Ontology Consortium. Nature genetics 25, 2005, s. 25÷95.
10. Bairagi R., Suchindran C.M.: An estimation of the cutoff point maximizing sum of sensitivity and specificity. Sankhya, Series B, Indian Journal of Statistics 51, 1989, s. 263÷269.
11. Baldi P., Hatfield G.W.: DNA Microarrays and Gene Expression. Cambridge University Press, Cambridge 2002.
12. Banzhaf J.F.: Weighted voting doesn't work: A mathematical analysis. Rutgers Law Review 19, 1965, s. 317÷343.

13. Barański A., Drzewiecki J., Kabiesz J., Konopko W., Kornowski J., Krzyżowski A., Mutke G.: Zasady stosowania metody kompleksowej i metod szczegółowych oceny stanu zagrożenia tapaniami w kopalniach węgla kamiennego. Główny Instytut Górnictwa, Instrukcje, 20, 2007.
14. Barber C.B., Dobkin D.P., Huh H.: The Quickhull algorithm for convex hulls. ACM Transactions on Mathematical Software, 22(4), 1996, s. 469÷483.
15. Bayardo R.J., Agrawal R.: Mining the most interesting rules. Proc. of the Fifth Int. Conf. on Knowledge Discovery and Data Mining, 1999, s. 145÷154.
16. Bazan J., Skowron A., Synak P.: Dynamic reducts as a tool for extracting laws from decision tables. Lecture Notes in Artificial Intelligence 869, 1994, s. 346÷355.
17. Bazan, J., Szczyka, M., Wróblewski, J.: A new version of rough set exploration system. Lecture Notes on Computer Science, 2475, 2002, s. 397÷404.
18. Bazan J., A., Skowron A., Ślęzak D., Wróblewski J.: Searching for the Complex Decision Reducts: The Case Study of the Survival Analysis. Lecture Notes in Artificial Intelligence 2871, 2003, s. 160÷168.
19. Bazan J.: Hierarchical Classifiers for Complex Spatio-temporal Concepts. Transactions on Rough Sets IX. LNCS 5390, 2008, s. 474÷750.
20. Benjamini Y., Hochberg Y.: Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of Royal Statistical Society B57(1), 1995, s. 289÷300.
21. Bennett K.P., Blue J.A.: A support vector machine approach to decision trees. Proc. of the IJCNN'97, 1997, s. 2396÷2401.
22. Bergadano F., Matwin S., Michalski R.S., Zhang J.: A general criterion for measuring quality of concept descriptions. Artificial Intelligence Center. Georg Mason Univ. Technical Report, 1988, s. 13÷88.
23. Berrado A., Runger G.C.: Using meta-rules to organize and group discovered association rules. Data Mining and Knowledge Discovery 14, 2007, s. 409÷431.
24. Berry M.J.A., Linoff G.S.: Mastering Data Mining. Wiley, New York 2000.
25. Bishop Y.M.M., Fienberg S.E., Holland P.W.: Discrete multivariate analysis: Theory and practice, The MIT Press, California 1991.
26. Blachnik M., Duch W.: LVQ algorithm with instance weighting for generation of prototype-based rules. Neural Networks 24(8), 2011, s. 824÷830.
27. Blanchard J., Guillet F., Kuntz P.: Semantic-based classification of rule interestingness measures. Post-Mining of Association Rules: Techniques for Effective Knowledge Extraction, 2009, s. 56÷79.

28. Bloedorn E., Michalski R.S.: Data-Driven Constructive Induction. *IEEE Intelligent Systems* 13(2), 1998, s. 30÷37.
29. Blum A.L., Langley P.: Selection of relevant features and examples in machine learning. *Artificial Intelligence* 97, 1997, s. 245÷271.
30. Błaszczyński J., Słowiński R., Szeląg M.: Sequential covering rule induction algorithm for variable consistency rough set approaches. *Information Sciences* 181(5), 2011, s. 987÷1002.
31. Błaszczyński J., Słowiński R., Susmaga R.: Rule-based estimation of attribute relevance. *Lecture Notes in Computer Science* 6954, 2011, s. 36÷44.
32. Błaszczyński J., Greco S., Słowiński R.: Inductive discovery of laws using monotonic rules. *Engineering Applications of Artificial Intelligence* 24, 2012, s. 284÷294.
33. Boser B.E., Guyon I.M., Vapnik V.: A training algorithm for optimal margin classifiers. *Proc. of the 5th Annual ACM Workshop on Computational Learning Theory*, 1992, s. 144÷152.
34. Bou-Hamad I., Larocque D., Ben-Ameur H.: A review of survival trees. *Statistics Surveys* 5, 2011, s. 44÷71.
35. Box G.E., Jenkins G.M.: Time series analysis: forecasting and control. Prentice Hall, New Jersey 1994.
36. Brans J.P., Mareschal B.: PROMETHEE Methods. *Multiple Criteria Decision Analysis: State of the Art Surveys*. Springer, New York 2005, s. 163÷195.
37. Brazdil P.B., Torgo L.: Knowledge acquisition via knowledge interaction. *Current Trends in Knowledge Acquisition*. IOS Press, Amsterdam 1990.
38. Breiman L., Friedman J., Olshen R., Stone R.: *Classification and Regression Trees*. Wadsworth, Statistic/Probability series, California 1984.
39. Breiman L.: Bagging predictors. *Machine Learning* 24(2), 1996, s. 123÷140.
40. Bruha I.: Quality of Decision Rules: Definitions and Classification Schemes for Multiple Rules. *Machine Learning and Statistics, The Interface*. John Wiley and Sons, Chichester 1997.
41. Bruha I.: From machine learning to knowledge discovery: Survey of preprocessing and postprocessing. *Intelligent Data Analysis* 4, 2000, s. 363÷374.
42. Bruha I., Tkadlec J.: Rule quality for multiple-rule classifier: Empirical expertise and theoretical methodology. *Intelligent Data Analysis* 7, 2003, s. 99÷124.
43. Brzezińska I., Greco S., Słowiński R.: Mining Pareto-optimal rules with respect to support and anti-support. *Engineering Applications of Artificial Intelligence*, 20(5), 2007, s. 587÷600.

44. Carmona-Saez P., Chagoyen M., Rodriguez A., Trelles O., Carazo J.M., Pascual-Montano A.: Integrated analysis of gene expression by association rules discovery. *BMC Bioinformatics* 7(54), 2006.
45. Carmona-Saez P., Chagoyen M., Tirado F., Carazo J.M., Pascual-Montano A.: Genecodis: a web based tool for finding significant concurrent annotations in gene list. *Genome Biology* 8, 2007.
46. Carnap R.: *Logical Foundations of Probability*. University of Chicago Press, Chicago 1962.
47. Carvalho D.R., Freitas A.A., Ebecken N.: Evaluating the correlation between objective rule interestingness measures and real human interest. *Lecture Notes in Computer Science* 3721, 2005, s. 453÷461.
48. Chawla N., Bowyer K., Hall L., Kegelmeyer W.: SMOTE Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* 16, 2002, s. 341÷378.
49. Chen S., Liu B.: Generating Classification Rules According to User's Existing Knowledge. Proc. of the SIAM International Conference on Data Mining (SDM-01), 2001.
50. Chorbev I., Mihajlov D., Jolevski I.: Web Based Medical Expert System with a Self Training Heuristic Rule Induction Algorithm. First International Conference on Advances in Databases, Knowledge, and Data Applications, 2009, s. 143÷148.
51. Christensen D.: Measuring confirmation. *Journal of Philosophy* XCVI, 1999, s. 437÷461.
52. Cichosz P.: *Systemy uczące się*. WNT, Warszawa 2000.
53. Clark P., Niblett T.: The CN2 induction algorithm. *Machine Learning* 3, 1989, s. 261÷283.
54. Clark P., Boswell R.: Rule induction with CN2: Some recent improvements. *Machine Learning – Proc. of the Fifth European Conference*, Springer-Verlag, Berlin 1991, s. 151÷163.
55. Clarke E., Waclawiw M.A.: Probabilistic rule induction from a medical research study database. *Computers and biomedical research* 29, 1996, s. 271÷283.
56. Cohen J.: A coefficient of agreement for nominal scales. *Educational and Psych. Meas.* 20, 1960, s. 37÷46.
57. Cohen W.W.: Fast effective rule induction. Proc. of the twelfth International Conference on Machine Learning, 1995, s. 115÷123.
58. Cox D.R., Oakes D.: *Analysis of survival data*. Chapman & Hall, London 1984.
59. Crupi V., Tentori K., Gonzalez M.: On Bayesian measures of evidential support: Theoretical and empirical issues. *Philosophy of Science* 74(2), 2007, s. 229÷252.
60. Daud N.R., Corne D.W.: Human readable rule induction in medical data mining. *Lecture Notes in Electrical Engineering* 27(7), 2009, s. 787÷798.

61. Delimata P., Moshkov M., Skowron A., Suraj Z.: Inhibitory Rules in Data Analysis. *Studies in Computational Intelligence* 163, Springer, Berlin–Heidelberg 2009.
62. Dembczyński K., Kotowski W. Słowiński R.: ENDER: a statistical framework for boosting decision rules. *Data Mining and Knowledge Discovery* 21, 2010, s. 52÷90.
63. Demsar J.: Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research* 7, 2006, s. 1÷30.
64. Diederich J.: Rule extraction from support vector machines. *Studies in Computational Intelligence* 80, Springer, Berlin–Heidelberg 2008.
65. Drwal G., Sikora M.: Fuzzy decision support system with rough set based rule generation method. *Lecture Notes in Artificial Intelligence* 3066, 2004, s. 727÷733.
66. Duch W., Adamczak R., Grabczewski K.: Extraction of logical rules from backpropagation networks. *Neural Processing Letters* 7, 1998, s. 211÷219.
67. Duch W., Adamczak R., Grabczewski K.: A new methodology of extraction, optimization and application of crisp and fuzzy logical rules. *IEEE Transaction on Neural Networks*, 11(2), 2001, s. 1÷31.
68. Duch W., Grudziński K.: Meta-learning via search combined with parameter optimization. *Intelligent Information Systems 2002, Advances in Soft Computing*, Physica-Verlag, 2002, s. 13÷22.
69. Duch W., Mandziuk J.: Chellenges for Computational Intelligence. Springer, Heidelberg 2007.
70. Duda R.O., Gasching J, Hart P.E.: Model design in the Prospector consultant system for mineral exploration. *Readings in Artificial Intelligence*, 1981, s. 334÷348.
71. Dudoit S., Shafer J.P., Boldrick J.: Multiple Hypothesis Testing in Microarray Experiments. *Statistical Science*, 18, 2003, s. 71÷103.
72. Dzeroski S., Lavrac N.: Rule induction and instance-based learning applied in medical diagnosis. *Technol Health Care* 4(2), 1996, s. 203÷221.
73. Eells E., Fitelson B.: Symmetries and asymmetries in evidential support. *Philosophical Studies* 107(2), 2002, s. 129÷142.
74. Eisen M.B., Spellman P.T., Brown P.O., Botstein D.: Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, 95, 1998, s. 14863÷14868.
75. Fawcett T.: An introduction to ROC analysis. *Pattern Recognition Letters* 27, 2006, s. 861÷874.
76. Fawcett T.: PRIE a system for generating rulelist to maximize ROC performance. *Data Mining and Knowledge Discovery* 17, 2008, s. 207÷224.

77. Fayad U.M., Piatetsky-Shapiro G., Smyth P., Uthurusamy R.: From data mining to knowledge discovery. *Advances in knowledge discovery and data mining*. MIT-Press, Cambridge 1996, s. 37÷58.
78. Ferguson D.E.: Fibonaccian searching. *Communications of the ACM* 3(12), 1960, s. 648.
79. Ferreira C.: Gene Expression Programming: Mathematical Modeling by an Artificial Intelligence. *Studies in Computational Intelligence* 26, 2006.
80. Fisher R.A.: On the interpretation of  $\chi^2$  from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society* 85(1), 1922, s. 87÷94.
81. Fourier-Viger P., Faghihi U., Nkambou R., Nguiwo E.M.: CMRules Mining sequential rules common to several sequences. *Knowledge-Based Systems* 25, 2012, s. 63÷76.
82. Frank A., Asuncion A.: UCI Machine Learning Repository (<http://archive.ics.uci.edu/ml>). Irvine, CA: University of California, School of Information and Computer Science, California 2010.
83. Freitas A.A.: On rule interestingness measures. *Knowledge-Based Systems* 12, 1999, s. 309÷315.
84. Friedman J.H., Popescu B.E.: Predictive learning via rule ensembles. *Annals of Applied Statistics* 2(3), 2008, s. 916÷954.
85. Fürnkranz J., Widmer G.: Incremental reduced error pruning. Proc. of the Tenth International Conference on Machine Learning. Morgan Kaufman, New Brunswick 1994, s. 70÷77.
86. Fürnkranz J.: Pruning algorithms for rule learning. *Machine Learning* 27, 1997, s. 139÷172.
87. Fürnkranz J.: Separate and conquer rule learning. *Artificial Intelligence Review* 13, 1999, s. 3÷54.
88. Fürnkranz J.: Modeling Rule Precision. *Lernen – Wissensentdeckung – Adaptivität*. Humboldt-Universität, Berlin 2004, s. 147÷154.
89. Fürnkranz J.: Workshop on Advances in Inductive Rule Learning. 15th European Conference on Machine Learning, 2004, s. 20÷24 (<http://www.ke.tu-darmstadt.de/events/ECML-PKDD-04-WS/opening-remarks.pdf>).
90. Fürnkranz J., Flach P.A.: ROC ‘n’ Rule Learning – Towards a Better Understanding of Covering Algorithms. *Machine Learning* 58, 2005, s. 39÷77.
91. Gago P., Beneto C.: A Metric for Selection of the Most Promising Rules. *Lecture Notes in Computer Science* 1510, 1998, s. 19-27.
92. Gains B.: Transforming rules and trees into comprehensible knowledge structures. *Advances in knowledge discovery and data mining*. MIT-Press, Cambridge 1996, s. 205÷228.

93. Gamberger D., Lavrac N.: Confirmation rule sets. Lecture Notes in Artificial Intelligence 1910, 2000, s. 34÷43.
94. Gamberger D., Lavrac N.: Expert-guided subgroup discovery: methodology and application. Journal of Artificial Intelligence Research 17, 2002, s. 501÷527.
95. Garcia S., Herrera F.: An Extension on Statistical Comparisons of Classifier over Multiple Data Sets for all Pairwise Comparisons. Journal of Machine Learning Research 9, 2008, s. 2677÷2694.
96. Geng L., Hamilton H.J.: Choosing the Right Lens: Finding What is Interesting in Data Mining. W [128].
97. Geng L., Hamilton H.J.: Interestingness measures for data mining: A survey. ACM Computing Surveys 39(3), 2006.
98. Gomolińska A.: On Certain Rough Inclusion Functions. Transactions on Rough Sets IX. LNCS 5490, 2008, s. 35÷55.
99. Goodman R.M., Smyth P.: The induction of probabilistic rule sets – the ITRULE algorithm. Proc. of the sixth international workshop on machine learning. Morgan Kaufmann, San Mateo 1989, s. 129÷132.
100. Gordon A., Glazko G., Qiu X., Yakovlev A.: Control of the mean number of false discoveries, bonferroni and stability of multiple testing. The Annals of Applied Statistics 1, 2007, s. 179÷190.
101. Góra G., Wojna A.: RIONA: a new classification system combining rule induction and instance based learning. Fundamenta Informaticae 51(4), 2002, s. 369÷390.
102. Graf E., Schmoor C., Sauerbrei W., Schumacher M.: Assessment and comparison of prognostic classification schemes for survival data. Statistics in Medicine 18(17-18), 1999, s. 2529÷2545.
103. Grąbczewski K.: Zastosowanie kryterium separowalności do generowania reguł klasyfikacji na podstawie baz danych. Rozprawa doktorska. Instytut Badań Systemowych PAN, Warszawa 2003.
104. Greco S., Matarazzo B., Słowiński R.: Rough sets theory for multicriteria decision analysis. European Journal of Operational Research 129, 2001, s. 1÷47.
105. Greco S., Matarazzo B., Słowiński R., Stefanowski J.: Importance and Interaction of Conditions in Decision Rules. Lecture Notes in Computer Science 2475, 2002, s. 255÷262.
106. Greco S., Pawlak Z., Słowiński R.: Can Bayesian confirmation measures be useful for rough set decision rules? Engineering Applications of Artificial Intelligence 17, 2004, s. 345÷361.

107. Greco S., Matarazzo B., Słowiński R., Stefanowski J.: Rough Membership and Bayesian Confirmation Measures for Parameterized Rough Sets. Lecture Notes in Artificial Intelligence 3641, 2005, s. 314÷324.
108. Greco S., Matarazzo B., Pappalard N., Słowiński R.: Measuring expected effects of interventions based on decision rules. Journal of Experimental and Theoretical Artificial Intelligence 17(1-2), 2005, s. 103÷118.
109. Greco S., Matarazzo B., Słowiński R.: Customer Satisfaction Analysis based on Rough Set Approach. Zeitschrift für Betriebswirtschaft 77(3), 2007, s. 325÷339.
110. Greco S., Słowiński R., Stefanowski J.: Evaluating importance of conditions in the set of discovered rules. Lecture Notes in Artificial Intelligence 4482, 2007, s. 314÷321.
111. Greco S., Matarazzo B., Słowiński R.: Parameterized rough set model using rough membership and Bayesian confirmation measures. International Journal of Approximate Reasoning 49, 2008, s. 285÷300.
112. Greco S., Słowiński R., Szczęch I.: Alternative normalization schemas for Bayesian confirmation measures. Lecture Notes in Computer Science 6178, 2010, s. 230÷239.
113. Greco S., Słowiński R., Szczęch I.: Properties of rule interestingness measures and alternative approaches to normalization of measures. Information Sciences 216, 2012, s. 1÷16.
114. Gruber T.R.: Toward principles for the design of ontologies used for knowledge sharing. International Journal of Human-Computer Studies 43(4-5), 1995, s. 907÷928.
115. Gruca A.: Analysis of GO composition of gene clusters by using multiattribute decision rules. Biocybernetics and Biomedical Engineering 28(4), 2008, s. 21÷31.
116. Gruca A., Sikora M., Polański A.: RuleGO: A logical rule-based tool for description of gene groups by means of Gene Ontology. Nucleic Acids Research 39, 2011, W293–W301.
117. Gruca A., Sikora M.: Identification of the compound subjective rule interestingness measure for rule-based functional description of genes. Lecture Notes in Computer Sciences 7557, 2012, s. 125÷134.
118. Grudziński K., Grochowski M., Duch W.: Pruning Classification Rules with Reference Vector Selection Methods. Lecture Notes in Artificial Intelligence 6113, 2010, s. 347÷354.
119. Grzymała-Busse J.: LERS – a system for learning from examples based on rough sets. W [276], s. 3÷18.
120. Grzymała-Busse J., Wang C. P.: Classification Methods in Rule Induction. Intelligent Information Systems. Proc. of the Workshop IIS, 1996, s. 120÷126.

121. Grzymała-Busse J., Goodwin L., Grzymała-Busse W., Zheng X.: An approach to imbalanced data sets based on changing rule strength. Proc. of learning from imbalanced data sets, AAAI workshop at the 17th conference on AI, 2000, s. 69÷74.
122. Grzymała-Busse J., Goodwin L., Zhang X.: Increasing sensitivity of preterm birth by changing rule strengths. Pattern Recognition Letters 24(6), 2003, s. 903÷910.
123. Grzymała-Busse J., Ziarko J.W.: Data mining based on rough sets. Data Mining Opportunities and Challenges. IGI Publishing, Hershey 2003, s. 142÷173.
124. Grzymała-Busse J., Stefanowski J., Wilk S.: A comparison of two approaches to data mining from imbalanced data. Journal of Intelligent Manufacturing 16, 2005, s. 565÷573.
125. Grzymała-Busse, J.W.: Characteristic relations for incomplete data: A generalization of the indiscernibility relation. Transactions on rough sets IV. LNCS 3700, 2005, s. 58÷68.
126. Grzymała-Busse J.: MLEM2 Rule Induction Algorithms: With and Without Merging Intervals. Data Mining: Foundations and Practice. Studies in Computational Intelligence 118, Springer, Berlin–Heidelberg 2008, s. 153÷164.
127. Gudýś A., Sikora M.: An algorithm for decision rules aggregation. Proc. of the Int. Conference on Knowledge Discovery and Information Retrieval, Valencia, Spain 2010, s. 216÷225.
128. Guillet F., Hamilton H.J.: Quality measures in data mining. Studies in Computational Intelligence 43, Springer-Verlag, New York 2007.
129. Gupta K.G., Strehl A., Ghosh J.: Distance based clustering of association rules. Proc. of ANNIE, Intelligent Engineering Systems Through Artificial Neural Networks, New York 1999, s. 759÷764.
130. Hand D.J.: Measuring classifier performance: a coherent alternative to the area under the ROC curve. Machine Learning 77, 2009, s. 103÷123.
131. Hastie T., Tibshirani R., Friedman J.H.: Elements of statistical learning: data mining, inference, and prediction. Springer, New York 2003.
132. Hilderman R.J., Hamilton H.J.: Knowledge Discovery and Measures of Interest. Kluwer Academic, Boston, MA 2001.
133. Hoeffding W.: Probability inequalities for sums of bounded random variables. Journal of the American Statistical Association 58, 1963, s. 13÷30.
134. Holm S.: A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistic 6, 1979, s. 65÷70.
135. Hühn J., Hüllermeier E.: FURIA: an algorithm for unordered fuzzy rule induction. Data Mining and Knowledge Discovery 19, 2009, s. 293÷319.

136. Huynh X.H., Guillet F., Briand H.: A data analysis approach for evaluating the behavior of interestingness measures. Proc. of the Discovery Science Conference, 2005, s. 330÷337.
137. Hvidsten T.R., Legreid A., Komorowski J.: Learning rule-based models of biological process form gene ontology expression time profiles using Gene Ontology. Bioinformatics 19(9), 2003, s. 1116÷1123.
138. Ilczuk G., Wakulicz-Deja A.: Attribute Selection and Rule Generation Techniques for Medical Diagnosis Systems. Lecture Notes in Artificial Intelligence 3642, 2005, s. 352÷361.
139. Ishibuchi H., Yamamoto T.: Effect of Three-Objective Genetic Rule Selection on the Generalization Ability of Fuzzy Rule-based Systems. Lecture Notes in Computer Science 2632, 2003, s. 608÷622.
140. Iyer V.R., Eisen M.B., Ross D.T., Schuler-Moore G. T., Lee J.C., Trent J.M., Staudt L.M., Hudson J., Boguski M.S., Lashkari D., Shalon D., Botstein D., Brown P.O.: The transcriptional program in the response of human fibroblasts to serum. Science 283, 1999, s. 83÷87.
141. Jacquet-Lagreze E., Siskos J.: Assessing a set of additive utility functions for multicriteria decision-making, the UTA method. European Journal of Operational Research 10, 1982, s. 151÷164.
142. Jankowski N., Duch W., Grąbczewski K.: Meta-Learning in Computational Intelligence, Springer, Berlin–Heidelberg 2011.
143. Janssen F., Fürnkranz J.: On the quest for optimal rule learning heuristics. Machine Learning 78, 2010, s. 343÷379.
144. Janssen F., Fürnkranz J.: Rule-Based Regression via Dynamic Reduction to Classification. Proc. of the 22nd International Joint Conference on Artificial Intelligence, IJCAI/AAAI, 2011, s. 1330÷1335.
145. Janusz A.: Discovering rules-based similarity in microarray data. Lecture Notes in Artificial Intelligence 6178, 2010, s. 49÷58.
146. Jaworski W.: Hybridization of Rough Sets and Statistica Learning Theory. Transactions on Rough Sets XIII. LNCS 6499, 2011, s. 39÷55.
147. Jerald F.: Statistical Models and Methods for Lifetime Data. 2nd edition. John Wiley and Sons, Hoboken 2003.
148. Jovanowski V., Lavrac N.: Classification Rule Learning with APRIORI-C. Lecture Notes in Computer Science 2258, 2001, s. 44÷51.
149. Joyce J.: The Foundations of Causal Decision Theory. Cambridge University Press, Cambridge 1999.

150. Kabiesz J.: Effect of the form of data on the quality of mine tremors hazard forecasting using neural networks. *Geotechnical and Geological Engineering*, 24(5), 2005, s. 1131÷1147.
151. Kamber M., Shinghal R.: Evaluating the interestingness of characteristic rules. Proc. of the Second International Conference on Knowledge Discovery and Data Mining, AAAI, 1996, s. 263÷266.
152. Kaneiwa K.: A rough set approach to mining connections from information systems. Proc. of the 2010 ACM Symposium on Applied Computing. ACM, 2010, s. 990÷996.
153. Kannan S., Bhaskaran R.: Association Rule Pruning based on Interestingness Measures with Clustering. *International Journal of Computer Science* 6(1), 2009, s. 35÷43.
154. Karp R.M.: Reducibility Among Combinatorial Problems. *Complexity of Computer Computations*. Plenum, New York 1972, s. 85÷103.
155. Kaplan E.L., Meier P.: Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.* 53, 1958, s. 457÷481.
156. Kaufman K.A., Michalski R.S.: Learning in Inconsistent World, Rule Selection in STAR/AQ18. Machine Learning and Inference Laboratory Report P99-2, February 1999.
157. Kavsek B., Lavrač N.: APRIORI-SD: Adapting association rule learning to subgroup discovery. *Applied Artificial Intelligence* 20, 2006, s. 543÷583.
158. Kawłak K., Porwolik J., Mielcarek M., Gorczyńska E. et al.: Higher CD34<sup>+</sup> and CD3<sup>+</sup> cell doses in the graft promote long-term survival, and have no impact on the incidence of serve acute or chronic Graft-versus-host disease. American Society for Blood and Marrow Transplantation. *Biology of Blood Marrow Transplantation* 16, 2010, s. 1388÷1401.
159. Kerber R.: Chimerge: Discretization of numeric attributes. Proc. of the Tenth National Conference on Artificial Intelligence. MIT Press, San Jose 1992, s. 123÷128.
160. Khatri P., Draghici S.: Ontological analysis of gene expression data: current tools, limitations and open problems. *Bioinformatics* 21(18), 2005, s. 3587÷3595.
161. Kim H., Loh W.Y.: Classification trees with bivariate linear discriminant node models. *Journal of Computational and Graphical Statistics* 12, 2003, s. 512÷530.
162. Klösgen W.: Explora: A Multipattern and Multistrategy Discovery Assistant. *Advances in Knowledge Discovery and Data Mining*. AAA/MIT Press, Menlo Park 1996, s. 249÷271.
163. Kohavi R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. Proc. of the 14th Int. Joint conference on Artificial intelligence. Morgan Kaufmann, Philadelphia 1995, s. 1137÷1143.
164. Kohavi R., George H.J.: Wrappers for feature subset selection. *Artificial Intelligence* 91, 1997, s. 273÷324.

165. Kononenko I., Bratko I.: Information-based evaluation criterion for classifier's performance. *Machine Learning* 6, 1991, s. 67÷80.
166. Krętowski M., Stepaniuk J.: Selection of Objects and Attributes, a Tolerance Rough Set Approach. Proc. of Ninth International Symposium on Methodologies for Intelligent Systems, 1996, s. 169÷180.
167. Kronek L.P., Reddy A.: Logical analysis of survival data: prognostic survival models by detecting high-degree interactions in right-censored data. *Bioinformatics* 24, 2008, s. 248÷253.
168. Kryszkiewicz M., Rybinski H.: Computation of reducts of composed information systems. *Fundamenta Informaticae*, 27, 1996, s. 183÷195.
169. Kryszkiewicz M.: Representative association rules. *Lecture Notes in Computer Science* 1394, 1998, s. 198÷209.
170. Kryszkiewicz M.: Rules in incomplete information systems. *Information Sciences* 113 (3–4), 1999, s. 271÷292.
171. Kryszkiewicz M.: Non-Derivable Item Set and Non-Derivable Literal Set Representations of Patterns Admitting Negation. *Lecture Notes in Computer Science* 5691, 2009, s. 138÷150.
172. Krzystanek Z., Sikora M., Trenczek S.: System rozmyty wspomagajacy identyfikacje źródeł emisji tlenku węgla w wyrobiskach kopalnianych. *Mechanizacja i Automatyzacja Górnictwa* 2, 2012.
173. Kużelewska U., Stepaniuk J.: Information Granulation: A Medical Case Study. *Transactions on Rough Sets IX. LNCS* 5390, 2008, s. 96÷113.
174. Latkowski R., Mikołajczyk M.: Data decomposition and decision rule joining for classification of data with missing values. *Transactions on Rough Sets I. LNCS* 3100, 2004, s. 299÷320.
175. Laurikkala J.: Improving identification of difficult small classes by balancing class distribution. *Lecture Notes in Artificial Intelligence* 2101, 2001, s. 63÷66.
176. Lavrac N., Flach P.A., Zupan B.: Rule Evaluation Measures: A Unifying View. *Lecture Notes in Artificial Intelligence* 1634, 1999, s. 174÷185.
177. Lavrac N., Kavsek B., Flach P.: Subgroup discovery with CN2-SD. *Journal of Machine Learning Research* 5, 2004, s. 153÷188.
178. Lenca P., Meyer P., Vaillant B., Lallich S.: A multicriteria decision aid for interestingness measure selection. Tech. Rep. LUSSI-TR-2004-01-EN, LUSSI Department, GET/ENST, Bretagne, France 2004.
179. Lenca P., Vaillant B., Meyer P., Lallich S.: Association Rule Interestingness Measures: Experimental and Theoretical Studies. W [276].

180. Leśniak A., Isakow Z.: Space-time clustering of seismic events and hazard assessment in the Zabrze-Bielszowice coal mine, Poland. *Int. Journal of Rock Mechanics and Mining Sciences* 46, 2009, s. 918÷928.
181. Li J.: On optimal rule discovery. *IEEE Transactions on Knowledge and Data Engineering* 18(4), 2006, s. 1÷12.
182. Li J., Cercone N.: A Method of Discovering Important Rules using Rules as Attributes. *International Journal of Intelligent Systems* 25, 2010, s. 180÷206.
183. Ligęza A.: Logical Foundations for Rule-Based Systems. *Studies in Computational Intelligence* 11, Springer-Verlag, Berlin–Heidelberg 2006.
184. Lindgren T., Boström H.: Resolving rule conflicts with double induction. *Intelligent Data Analysis* 8(5), 2004, s. 457÷468.
185. Liu H., Liu L., Zhang H.: A fast pruning redundant rule method using Galois connection. *Applied Soft Computing* 11, 2011, s. 130÷137.
186. Łęski J.: Sieci neuronowo-rozmyte. WNT, Warszawa 2008.
187. McGarry K.: A survey of interestingness measures for knowledge discovery. *The Knowledge Engineering Review* 20(1), Cambridge University Press, 2005, s. 39÷61.
188. Martens D., Baesens B., Van Gestel T., Vanthienen J.: Comprehensible credit scoring models using rule extraction from support vector machines. *European Journal of Operational Research* 183(3), 2007, s. 1466÷1476.
189. Michalski R.S.: Discovering Classification Rules Using variable-Valued Logic System VL\_1. Proc. of the 3rd International Joint Conference on Artificial Intelligence, 1973, s. 162÷172.
190. Michalski R.S., Carbonell J.G., Mitchel T.M.: Machine Learning vol. I. Morgan Kaufmann, Los Altos 1983.
191. Michalski R.S.: O naturze uczenia się – problemy i kierunki badawcze. *Informatyka* No. 2, 3, 1988.
192. Michalski R.S., Bratko I., Kubar M.: Machine learning and data mining. John Wiley and Sons, London 1998.
193. Michalski R.S., Wojtusiak J.: Generalizing Data in Natural Language. *Lecture Notes in Computer Sciecnies* 4585, 2007, s. 29÷39.
194. Michie D., Spiegelhalter D.J., Taylor C.C.: Machine Learning, neural and statistical classification. Ellis Horwood Limited, England 1994.
195. Midelfart H.: Supervised Learning in Gene Ontology Part I: A Rough Set Framework. *Transactions on Rough Sets IV. LNCS* 3700, 2005, s. 69÷97.
196. Midelfart H.: Supervised Learning in Gene Ontology Part II: A Bottom-Up Algorithm. *Transactions on Rough Sets IV. LNCS* 3700, 2005, s. 98÷124.

197. Mikołajczyk M.: Reducing number of decision rules by joining. Lecture Notes in Artificial Intelligence 2475, 2002, s. 425÷432.
198. Moshkov M., Piliszcuk M., Zielosko B.: On construction of partial reducts and irreducible partial decision rules. Fundamenta Informaticae 75(1-4), 2007, s. 357÷374.
199. Moshkov M., Piliszcuk M., Zielosko B: Partial Covers, Reducts and Decision Rules in Rough Sets. Studies in Computational Intelligence 145, Springer, Berlin–Heidelberg 2008.
200. Mozina M., Zabkar J., Bratko I.: Argument based machine learning. Artificial Intelligence 171, 2007, s. 922÷937.
201. Mrózek A.: Rough sets in computer implementation of rule-based control of industrial processes. W [276], s. 19÷33.
202. Mrózek A., Płonka L.: Knowledge Representation in Fuzzy and Rough Controllers. Fundamenta Informaticae 30(3-4), 1997, s. 299÷311.
203. Mrózek A., Sikora M.: Synteza automatu modelującego zachowanie się człowieka w procesie podejmowania decyzji. Zeszyty Naukowe Politechniki Śląskiej s. Informatyka z. 1377, 1997, s. 233÷250.
204. Muggleton S., De Raedt L.: Inductive logic programming: Theory and methods. Journal of Logic Programming 12, 1994, s. 1÷80.
205. Murthy S.K., Kasif S., Salzberg S.: A system for induction of oblique decision trees. Journal of Artificial Intelligence Research 2, 1994, s. 1÷32.
206. Napierała K., Stefanowski J.: Argument Based Generalization of MODLEM Rule Induction Algorithm. Lecture Notes in Computer Science 6086, 2010, s. 138÷147.
207. Napierała K., Stefanowski J.: BRACID: a comprehensive approach to learning rules form imbalanced data. Journal of Intelligent Systems 39(2), 2012, s. 335÷373.
208. Nguyen H.S., Nguyen S.H.: Some efficient algorithms for rough set methods. Proc. of the Sixth International Conference, Information Processing and Management of Uncertainty in Knowledge-Based Systems, 1996, s. 1451÷1456.
209. Nguyen S.H., Nguyen H.S.: Discretization Methods in Data Mining. Rough Sets in Knowledge Discovery. Physica-Verlag, Heidelberg, 1998, s. 451÷482.
210. Nguyen S.H.: Regularity analysis and its applications in Data Mining. Rozprawa doktorska. Uniwersytet Warszawski Wydział Matematyki Informatyki i Mechaniki, 1999.
211. Nguyen H.S.: Approximate Boolean Reasoning: Foundations and Applications in Data Mining. Transactions on Rough Sets V. LNCS 4100, 2006, s. 334÷506.

212. Nguyen H.S., Jankowski A., Peters J.F., Skowron A., Stepaniuk J., Szczuka M.: Discovery of Process Models from Data and Domain Knowledge: A Rough-Granular Approach. Novel Developments in Granular Computing. Information Sciences Reference IGI Global, Hershey–New York 2010.
213. Novak P.K., Lavrac N.: Supervised Descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining. *Journal of Machine Learning Research* 10, 2009, s. 377÷403.
214. Nowak A., Wakulicz-Deja A.: The concept of the hierarchical clustering algorithms for rules based systems. *Advances in Soft Computing*. Springer, Berlin–Heideleber–New York 2005, s. 565÷570.
215. Ohsaki M., Abe H., Tsumoto S., Yokoi H., Yamaguchi T.: Evaluation of rule interestingness measures in medical knowledge discovery in databases. *Artificial Intelligence in Medicine* 41, 2007, s. 177÷196.
216. Orlau C., Wehenkel L.: A complete fuzzy decision tree technique. *Fuzzy Sets and Systems* 138, 2003, s. 221÷254.
217. Orłowska E.: Dynamic Information Systems. Fundamenta. *Informaticae* 5, 1983, s. 101÷118.
218. Padmanabhan B., Tuzhilin A.: A belief-driven method for discovering unexpected patterns. Proc. of the Fourth Int. Conference on Knowledge Discovery and Data Mining, 1998, s. 94÷100.
219. Pattaraintakorn P., Cercone N.: A foundation of rough sets theoretical and computational hybrid intelligent system for survival analysis. *Computers and Mathematics with Applications* 56, 2008, s. 1699÷1708.
220. Pawlak Z.: Rough sets: Theoretical aspects of reasoning about data. Kluwer Academic Publishers Norwell, MA, USA, 1992.
221. Pednault E.: Minimal-Length Encoding and Inductive Inference. *Knowledge Discovery in Databases*. MIT Press, Cambridge 1990.
222. Pham D.T., Afify A.A. Three New MDL-Based Pruning Techniques for Robust Rule Induction. Proc. of the Inst. of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, 2006, s. 553÷564.
223. Piatetsky-Shapiro G.: Discovery, analysis and presentation of strong rules. *Knowledge Discovery in Databases*. MIT Press, Cambridge, MA 1991, s. 229÷248.
224. Pindur R., Sasmuga R., Stefanowski J.: Hyperplane Aggregation of Dominance Decision Rules. *Fundamenta Informaticae* 61(2), 2004, s. 117÷137.
225. Polański A., Kimmel M.: Bioinformatics. Springer, Berlin–Heidelberg–New York 2007.

226. Polkowski L., Skowron A., Źytkow J.: Tolerance based rough sets. Soft computing: Rough sets, fuzzy logic, neural networks, uncertainty management. Simulation Councils Inc., San Diego, USA 1995, s. 55÷58.
227. Polkowski L., Skowron A.: Rough mereology: A new paradigm for approximate reasoning. International Journal of Approximated Reasoning 15, 1996, s. 333÷365.
228. Provost F., Fawcett T., Kohavi R.: The case against accuracy estimation for comparing induction algorithms. Proc. of the Fifteenth Int. Conference on Machine Learning. Morgan Kaufman, San Francisco 1997, s. 445÷453.
229. R Development Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria 2011, <http://www.R-project.org/>.
230. Radspiel-Troger M., Rabenstein T., Schneider H.T., Lausen B.: Comparison of tree-based methods for prognostic stratification of survival data. Artificial Intelligence in Medicine 28, 2003, s. 323÷341.
231. Rafea A.A., Shaflik S.S., Shaalan K.F.: An interactive system for association rule discovery for life assurance. Proc. of the Int. Conference on Computer, Communication and Control Technologies CCCT, 2004, s. 32÷37.
232. Ramsey P.H.: Multiple Comparison of Independent Means. Applied analysis of variance in behavioral science. Chapman & Hall/CRC, Florida 1993, s. 25÷62.
233. Raś Z.W., Daradzińska A., Liu X.: System ADReD for discovering rules based on hyperplanes. Engineering Applications of Artificial Intelligence 17(4), 2004, s. 401÷406.
234. Riddle P., Segal R., Etzioni O.: Representation design and brute-force induction in a Boeing manufacturing domain. Journal of Applied Artificial Intelligence 8, 1994, s. 125÷147.
235. Rioult F., Zanuttini B., Cremilleux B.: Nonredundant Generalized Rules and Their Impact in Classification. Advances in Intelligent Information Systems. Studies in Computational Intelligence 265, 2010, s. 3÷25.
236. Rudajev V., Číž R.: Estimation of Mining Tremor Occurrence by Using Neural Networks. Pure and Applied Geophysics 154(1), 1999, s. 57÷72.
237. Quinlan J.R.: Learning with continuous classes, Proc. of the International Conference on Artificial Intelligence (AI'92), 1992.
238. Quinlan J.R.: C4.5: Programs for Machine Learning. Morgan Kaufman Publishers Inc., USA, 1993.
239. Quinlan J.R.: Learning First-Order Definitions of Functions. Journal of Artificial Intelligence Research 5, 1996, s. 139÷161.
240. Sahar S.: Interestingness measures – On determining what is interesting. Data Mining and Knowledge Discovery Handbook. Part 5. Springer Verlag, Singapore 2010, s. 603÷612.

241. Schumacher M., Holländer N., Schwarzer G., Sauerbrei W.: Prognostic factor studies. *Handbook of Statistics in Clinical Oncology*. Taylor & Francis Group, New York 2006.
242. Shan N., Ziarko W.: Data-based acquisition and incremental modification of classification rules. *Computational Intelligence* 11, 1995, s. 357÷370.
243. Shapley L.S.: A value for n-person games. *Contributions to the Theory of Games II*, Princeton University Press, Princeton 1953, s. 307÷317.
244. Sikora M.: Filtracja zbioru reguł decyzyjnych wykorzystująca funkcje oceny jakości reguł. *Studia Informatica* 22(4), 2001, s. 57÷72.
245. Sikora M., Widera D.: Identification of diagnostics states for dewater pumps working in abyssal mining pump stations. *Proc. of the XV International conference on System Sciences*, 2004, s. 394÷402.
246. Sikora M.: Approximate decision rule induction algorithm using rough sets and rule-related quality measures. *Archives of Theoretical and Applied Informatics* 4, 2004, s. 1÷19.
247. Sikora M., Krzykawski D.: Zastosowanie metod eksploracji danych do analizy wydzielania się dwutlenku węgla w pomieszczeniach stacji odwadniania kopalń węgla kamiennego. *Mechanizacja i Automatyzacja Górnictwa*, 413(6), 2005.
248. Sikora M.: An algorithm for generalization of decision rules by joining. *Foundations of Computing and Decision Sciences* 30(3), 2005, s. 227÷239.
249. Sikora M.: Fuzzy rules generation method for classification problems using rough sets and genetic algorithms. *Lecture Notes in Artificial Intelligence* 3641, 2005, s. 383÷391.
250. Sikora M., Sikora B.: Application of Machine Learning for prediction a methane concentration in a coal-mine. *Archives of Mining Sciences* 51(4), 2006, s. 475÷492.
251. Sikora M.: Rule quality measures in creation and reduction of data rule models. *Lecture Notes in Artificial Intelligence* 4259, 2006, s. 716÷725.
252. Sikora M.: Application of machine learning and soft computing techniques in monitoring systems' data analysis by example of dewater pumps monitoring system. *Archives of Control Sciences* 17(4), 2007, s. 369÷391.
253. Sikora M.: Adaptacyjne stosowanie miar oceniających w algorytmach indukcji reguł. *Bazy danych – Nowe technologie – Architektura, metody formalne i zaawansowana analiza danych*. Wydawnictwa Komunikacji i Łączności, Warszawa 2007, s. 295÷305.
254. Sikora M.: Decision rule-based data models using TRS and NetTRS – methods and algorithms. *Transactions on Rough Sets XI. LNCS* 5946, 2010, s. 130÷160.

255. Sikora M., Gruca A.: Quality improvement of rule based gene groups descriptions using information about GO terms importance occurring in premises of determined rules. *International Journal of Applied Mathematics and Computer Science* 20(3), 2010, s. 555÷570.
256. Sikora M., Wróbel Ł.: Application of rule induction algorithms for analysis of data collected by seismic hazard monitoring systems in coal mines. *Archives of Mining Sciences* 55, 2010, s. 91÷114.
257. Sikora M., Gruca A.: Induction and selection of the most interesting Gene Ontology based multiattribute rules for descriptions of gene groups. *Pattern Recognition Letters* 32(2), 2011, s. 258÷269.
258. Sikora M.: Induction and pruning of classification rules for prediction of microseismic hazards in coal mines. *Expert Systems with Applications* 38(6), 2011, s. 6748÷6758.
259. Sikora M., Krzystanek Z., Bojko B., Śpiechowicz K.: Application of hybrid machine learning method for description and on-line assessment of methane hazards in a mine excavation. *Journal of Mining Sciences* 47(4), 2011, s. 493÷505.
260. Sikora M., Michalak M., Sikora B.: Approximation of a coal mass by an ultrasonic sensor using regression rules. *Lecture Notes on Computer Science* 6743, 2011, s. 345÷350.
261. Sikora M., Wróbel Ł.: Data-driven Adaptive Selection of Rule Quality Measures for Improving the Rule Induction Algorithm. *Lecture Notes in Artificial Intelligence* 6743, 2011, s. 279÷287.
262. Sikora M., Sikora B.: Rough natural hazard monitoring, [in:] Peters G., Lingras P., Ślezak D., Yao Y. (eds.), *Rough Sets: Selected Methods and Applications in Management and Engineering*. Springer, Berlin 2012.
263. Sikora M., Sikora B.: Combination of regression rule induction, k-nearest neighbors and time series forecasting methods to improve prediction models applied in systems monitoring natural hazards and machinery. *International Journal of Applied Mathematics and Computer Science* 22(2), 2012, s. 477÷491 .
264. Sikora M., Wróbel Ł.: Data-driven Adaptive Selection of Rule Quality Measures for Improving Rule Induction and Filtration Algorithms. *International Journal of General Systems* (w druku, ukaże się w numerze 42(4), 2013).
265. Sikora M., Gudyś A.: CHIRA – Convex Hull Based Iterative Algorithm of Rules Aggregation. *Fundamenta Informaticae* (w druku, ukaże się w numerze 123, 2013).
266. Sikora M.: Redefinition of classification rules by evaluation of elementary conditions occurring in the rule premises. *Fundamenta Informaticae* (w druku, ukaże się numerze 123, 2013).
267. Sikora M., A. Skowron, Wróbel Ł.: Rule Quality Measure-Based Induction of Unordered Sets of Regression Rules. *Lecture Notes in Computer Sciences* 7557, 2012, s. 162÷171.

268. Sikora M., Wróbel Ł., Mielcarek M., Kawłak K.: Decision rule-based analysis of survival data: methodology and application in discovering survival factors of patients after bone marrow transplantation. *Medical Informatics and Technologies* (przyjęte do druku).
269. Siskos Y., Grigoroudis E., Matsatsinis N.F.: UTA Methods. State of the Art in Multiple Criteria Decision Analysis, Springer, New York 2005, s. 297÷343.
270. Skowron A., Rauszer C.: The Discernibility Matrices and Functions in Information systems. W [276], s. 331÷362.
271. Skowron A., Stepaniuk J.: Tolerance Approximation Spaces. *Fundamenta Informaticae* 27(2-3), 1996, s. 245÷253.
272. Skowron A., Wang H., Wojna A., Bazan J.G.: A Hierarchical Approach to Multimodal Classification. *Lecture Notes in Computer Science* 3642, 2005, s. 119÷127.
273. Skowron A., Szczuka M.S.: Toward Interactive Computations: A Rough-Granular Approach. Advances in Machine Learning II. Dedicated to the Memory of Professor Ryszard S. Michalski. *Studies in Computational Intelligence* 263, 2010, s. 23÷42.
274. Skowron A.: Discovery of Processes and Their Interactions form Data and Domain Knowledge. *Lecture Notes in Artificial Intelligence* 6070, 2010, s. 12÷21.
275. Skowron A., Stepaniuk J., Świniarski R.: Modeling rough granular computing based on approximation spaces. *Information Sciences* 184, 2012, s. 20÷43.
276. Słowiński R.: Intelligent decision support: Handbook of Applications and Advances of Rough Sets Theory. Kluwer Academic Publishers, Dordrecht–Boston–London 1992.
277. Słowiński R., Greco S.: Measuring attractiveness of rules from the viewpoint of knowledge representation, prediction and efficiency of intervention. *Lecture Notes in Artificial Intelligence* 3528, 2005, s. 11÷22.
278. Słowiński K., Stefanowski J., Siwiński D.: Application of rule induction and rough sets to verification of megnetic resonance diagnosis. *Fundamenta Informaticae* 53, 2002, s. 345÷363.
279. Smyth P., Goodman R.M.: Rule induction using information theory. Proc. of Knowledge Discovery in Databases. MIT Press, Boston 1991, s. 159÷176.
280. Srinivasan A: Note on the location of optimal classifiers in n-dimensional ROC space. Technical Report PRG-TR-2-99, Oxford University Computing, Oxford 1999.
281. Stajduhar I., Dalbelo-Basic B.: Uncensoring censored data for machine learning: A likelihood-based approach. *Expert Systems with Applications* 2012.
282. Stapor K.: Automatyczna klasyfikacja obiektów. Problemy Współczesnej Nauki. Teoria i Zastosowania. Akademicka Oficyna Wydawnicza EXIT, Warszawa 2005.
283. Stefanowski J.: Rough set based rule induction techniques for classification problems. Proc. Of the 6<sup>th</sup> European Congerss of Intelligent Techniques and Soft Computing, 1998, s.107÷119.

284. Stefanowski J.: Algorytmy indukcji reguł decyzyjnych w odkrywaniu wiedzy. Wydawnictwo Politechniki Poznańskiej, s. Rozprawy, nr 361, 2001.
285. Stefanowski J., Vanderpooten D.: Induction of Decision Rules in Classification and Discovery-Oriented Perspectives. International Journal of Intelligent Systems 16, 2001, s. 13÷27.
286. Stefanowski J., Wilk S.: Evaluating business credit risk by means of approach integrating decision rules and case based learning. International Journal of Intelligent Systems in Accounting, Finance and Management 10, 2001, s. 97÷114.
287. Stefanowski J., Tsoukiàs A.: Incomplete information tables and rough classification. Computational Intelligence 17(3), 2001, s. 545÷566.
288. Stefanowski J., Wilk S.: Combining rough sets and rule based classifiers for handling imbalanced data. Fundamenta Informaticae 72(1-3), 2006, s. 379÷391.
289. Stefanowski J.: On Combined Classifiers, Rule Induction and Rough Sets. Transactions on Rough Sets VI. LNCS 4374, 2007, s. 329÷350.
290. Stepaniuk J.: Optimizations of Rough Set Model. Fundamenta Informaticae 25, 1998, s. 1÷19.
291. Stepaniuk J.: Rough Set Data Mining of Diabetes Data. Lecture Notes in Computer Science 1609, 1999, s. 457÷465.
292. Stepaniuk J.: Knowledge Discovery by Application of Rough Set Models. Instytut Podstaw Informatyki PAN, Raport 887, Warszawa 1999.
293. Stepaniuk J., Honko P.: Learning First-Order Rules: A Rough Set Approach Fundamenta Informaticae 61(2), 2004, s. 139÷157.
294. Stepaniuk J., Bazan J.G., Skowron A.: Modeling Complex Patterns by Information Systems. Fundamenta Informaticae 67, 2005, s. 203÷217.
295. Stepaniuk J.: Rough-granular computing in knowledge discovery and data mining. Studies in Computational Intelligence 152. Springer, Berlin–Heidelberg 2008.
296. Sulzman J.N., Fürnkranz J.: An Empirical Comparison of Probability Estimation Techniques for Probabilistic Rules. Lecture Notes in Computer Science 5808, 2009, s. 317÷331.
297. Suzuki E.: Negative Encoding Length as a Subjective Interestingness Measure for Groups of Rules. Proc. Of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, 2009, s. 220÷231.
298. Suzuki E.: Compresion-based Measures for Mining Interesting Rules. Lecture Notes in Artificial Intelligence 5579, 2009, s. 741÷746.
299. Synak P.: Temporal Templates and Analysis of Time Related Data. Lecture Notes in artificial Intelligence 2005, 2001, s. 420÷427.

300. Szczęch I.: Multicriteria Attractiveness Evaluation of Decision and Association Rules. *Transactions on Rough Sets X.* LNCS 5656, 2009, s. 197÷274.
301. Ślęzak D.: Approximate reducts in decision tables. Proc. of the Sixth International Conference, Information Processing and Management of Uncertainty in Knowledge-Based Systems, 1996, s. 1159÷1164.
302. Ślęzak D., Wróblewski J.: Classification Algorithms Based on Linear Combinations of Features. *Lecture Notes in Computer Science 1704*, 1999, s. 548÷553.
303. Ślęzak D.: Appriximate Entropy Reducts. *Fundamenta Informaticae* 54(3-4), 2002, s. 365÷390.
304. Ślęzak D.: Rough Sets and Bayes Factor. *Transactions on Rough Sets III.* LNCS 3400, 2005, s. 202÷229.
305. Ślęzak D., Widz S.: Is It Important Which Rough-Set-Based Classifier Extraction and Voting Criteria Are Applied Together? *Lecture Notes in Artificial Intelligence* 6086, 2010, s. 187÷196.
306. Ślęzak D., Widz S.: Rough-Set-Inspired Feature Subset Selection, Classifier Construction, and Rule Aggregation. *Lecture Notes in Computer Science* 6954, 2011, s. 81÷88.
307. Tan P., Kumar V., Srivastava J.: Selecting the right interestingness measure for association analysis. *Information Sciences* 29, 2004, s. 293÷313.
308. Tickle A.B., Andrews R., Golea M., Diederich J.: The truth will come to light: Directions and challenges in extracting the knowledge embedded within trained artificial neural networks. *IEEE Transactions on Neural Networks* 9, 1998, s. 1057÷1068.
309. Todorovski L., Flach P., Lavrac N.: Predictive performance of weighted relative accuracy. *Lecture Notes in Computer Science* 2258, 2000, s. 255÷264.
310. Towell G.G., Shavlik J.W.: Extracting Refined Rules from Knowledge-Based Neural Networks. *Machine Learning* 13, 1993, s. 71÷101.
311. Tsai C.J., Lee C.I., Yang W.P.: Mining decision rules on data streams in the presence of concept drifts. *Expert Systems with Applications* 36, 2009, s. 1164÷1178.
312. Tsumoto S., Tanaka H.: Automatem Discovery of Medical Expert System Rules form Clinical Databases based on Rough Sets. *KDD-96 Proc. of the AAAI 1996*, s. 63÷69.
313. Tsumoto S.: Automated Discovery of Positive and Negative Knowledge in Medical Databases. *IEEE Engineering in Medicine and Biology* 19(4), 2002, s. 56÷62.
314. Tsumoto S., Hirano S.: Visualization of Rules's Similarity using Multidimensional Scaling. *Proc. of the 3<sup>rd</sup> IEEE Int. Conference on Data Mining.* IEEE, 2003, s. 339÷346.
315. Tsumoto S.: Mining diagnostic rules from clinical databases using rough sets and medical diagnostic model. *Information Sciences* 162, 2004, s. 65÷80.

316. Tsumoto S.: A new framework for incremental rule induction based on rough sets. IEEE International Conference on Granular Computing, IEEE Press, California 2011, s. 681÷686.
317. Tzacheva A., Raś Z.: Action rules mining, International Journal of Intelligent Systems 20(7), 2005, s. 719÷736.
318. Ulutasdemir N., Dagli O.: Evaluation of risk of death in hepatitis by rule induction algorithms. Scientific Research and Essays 5(20), 2010, s. 3059÷3062.
319. Vaillant B., Lenca P., Lallich S.: A clustering of interestingness measures. Lecture notes in Artificial Intelligence 3245, 2004, s. 290÷297.
320. van Aswegen G.: Routine seismic hazard assessment in some South African mines. Proc. of the sixth international symposium on rockburst and seismicity in mines. Australian Centre for Geomechanics, Nedlands 2005, s. 437÷44.
321. Vapnik V.N.: Statistical Learning Theory. John Wiley and Sons, New York 1998.
322. Wakulicz-Deja A., Paszek P.: Applying Rough Set Theory to Multi Stage Medical Diagnosing. Fundamenta Informaticae 54(4), 2003 str.387-408.
323. Wakulicz-Deja A., Nowak A.: From Information Systems to Decision Support Systems. Lecture Notes in Computer Science 5390, 2008, s. 379÷404.
324. Wakulicz-Deja A., Przybyła-Kasperek M.: Multi-Agent Decision Taking System. Fundamenta Informaticae 101(1-2), 2010, s. 125÷141.
325. Wakulicz-Deja A., Przybyła-Kasperek M.: Application of the Method of Editing and Condensing in the Process of Global Decision-making. Fundamenta Informaticae 106(1), 2011, s. 93÷117.
326. Wakulicz-Deja A., Brzezińska-Nowak A., Jach T.: Inference Processes in Decision Support Systems with Incomplete Knowledge. Lecture Notes in Computer Science 6954, 2011, s. 616÷625.
327. Webb G.I.: Further experimental evidence against the utility of Occam's razor. Journal of Artificial Intelligence Research 4, 1996, s. 397÷417.
328. Webb G.I., Zhang S.: K-optimal rule discovery. Data Mining and Knowledge Discovery 10, 2005, s. 39÷79.
329. Webb G.I.: Discovering significant patterns. Machine Learning 68(1), 2007, s. 1÷33.
330. Weiss S.M., Indurkhya N.: Rule-based machine learning methods for functional prediction. Journal of Artificial Intelligence Research 3, 1995, s. 383÷403.
331. Winiarko E., Roddick J.F.: ARMADA – An algorithm for discovering richer relative temporal association rules from interval-based data. Journal of Data and Knowledge Engineering 63(1), 2007, s. 76÷90.
332. Winston P.H.: Artificial Intelligence. Addison-Wesley, Cambridge 1992.

333. Witten I.H., Frank E.: Data mining: practical machine learning tools and techniques. Morgan Kaufmann, 2011.
334. Wohlrab L., Fürnkranz J.: A review and comparison of strategies for handling missing values in separate-and-conquer rule lerarning. *Journal of Intelligent Information Systems* 36(1), 2011, s. 73÷98.
335. Wojna A.: Constraint based incremental learning of classification rules. *Lecture Notes in Artificial Intelligence* 2005, 2001, s. 428÷435.
336. Wojtusiak J., Michalski R.S., Kaufman K., Pietrzykowski J.: The AQ21 Natural Induction Program for Pattern Discovery: Initial Version and its Novel Features. *18th IEEE International Conference on Tools with Artificial Intelligence*, IEEE Computer Society 2006, s. 523÷526.
337. Wolpert D.H.: The lack of a priori distinctions between learning algorithms. *Neural Computation* 8(7), 1996, s. 1341÷1390.
338. Wnek J., Michalski R.S: Hypothesis-Driven Constructive Induction in AQ17-HCI: A Method and Experiments. *Machine Learning* 14(1), 1994, s. 139÷168.
339. Wu X., Kumar V., Quinlan J.R., et al.: Top 10 algorithms in data mining. *Knowledge and Information Systems* 14(1), 2008, s. 1÷37.
340. Yao Y.Y., Zhong N.: An analysis of quantitative measures associated with rules. *Lecture Notes in Artificial Intelligence* 1574, 1999, s. 479÷488.
341. Yao Y.Y.: Information-theoretic measures for knowledge discovery and data mining. *Entropy Measures, Maximum Entropy and Emerging Applications. Studies in Fuzziness and Soft Computing*. Springer-Verlag, Heidelberg 2003, s. 115÷136.
342. Yao Y., Zhou B.: Micro and macro evaluation of classification rules. *Proc. 7th IEEE Int. Conf. on Cognitive Informatics*, 2008, s. 441÷448.
343. Yao Y., Zhao Y., Wang J.: On reduct construction algorithms. *Lecture Notes in Computer Science* 5150, 2008, s. 100÷117.
344. Zhao Y., Hang C., Cao L.: Post-Mining of Association rules: Techniques for Effective Knowledge Extraction. *Information Sciences Reference* IGI Global, Hershey 2009.
345. Zhu H., Huang W., Zheng H.: Method for discovering actionable rule. *Proc. of the Fourth International Conference on Fuzzy Systems and Knowledge Discovery*, 2007.
346. Ziarko W.: Variable precision rough set model. *Journal of Computer and Systems Sciences* 461, 1993, s. 39÷59.
347. Zupan B., Demsar J., Kattan M.W., Beck R., Bratko I.: Machine Learning for Survival Analysis: A Case Study on Recurrence of Prostate Cancer. *Lecture Notes in Artificial Intelligence* 1620, 1999, s. 346÷355.

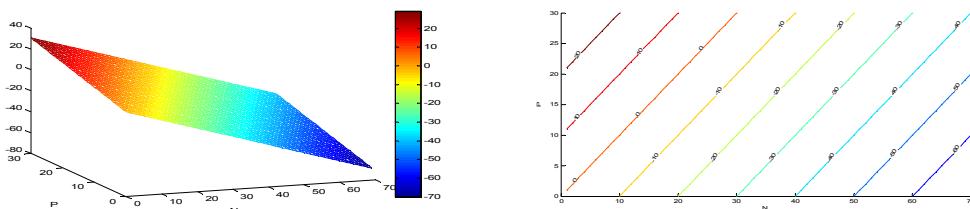


## DODATEK A. WYKRESY MIAR JAKOŚCI

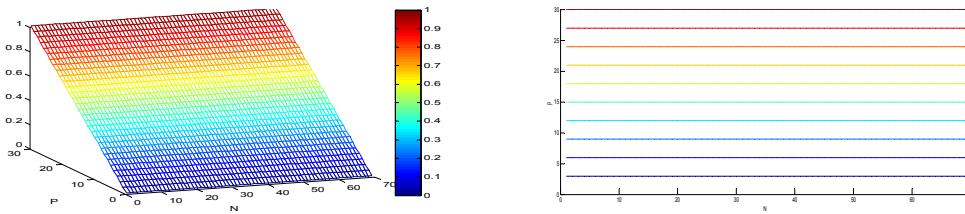
Poniżej zaprezentowano trójwymiarowe i dwuwymiarowe wykresy obiektywnych miar jakości, definiowanych na podstawie tablicy kontyngencji. Kolejność prezentowania wykresów jest zgodna z kolejnością miar w tabeli 3.5. Wykresy utworzono dla hipotetycznego zbioru, złożonego z 30 przykładów pozytywnych i 70 przykładów negatywnych.

Na rysunku A.37 zamieszczono także wykres miary  $nMM(C1,Odds,C2,g)$ . Jako uzupełnienie, na rysunku A.38 przedstawiono wykres miary  $OWS$  dla innego rozkładu przykładów pozytywnych i negatywnych. Wykres ten lepiej pokazuje, że miara nie jest monotoniczna. Na rysunku A.39 przedstawiono wykresy miar  $LS$ ,  $Odds$  oraz  $RR$  w skali logarytmicznej.

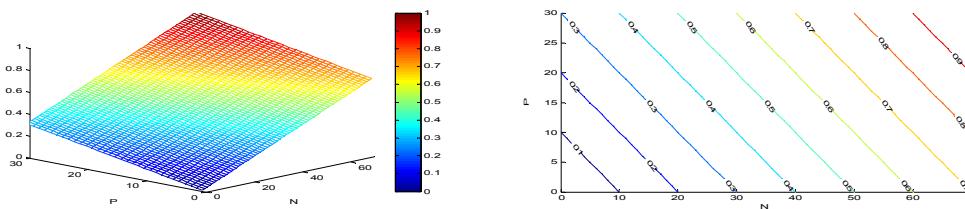
Wszystkie wykresy narysowano dla podstawowych (niemodyfikowanych) postaci miar. Oś przykładów pozytywnych oznaczono jako  $P$ , oś przykładów negatywnych – jako  $N$ . Wykresy wyznaczono dla tych par  $(p,n) \in [0,30] \times [0,70]$ , dla których wartości miar są określone.



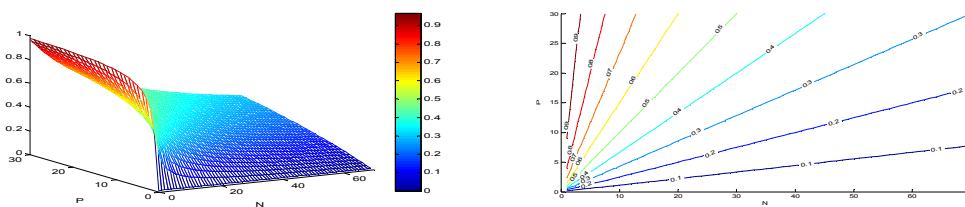
Rys. A.1. Wykres miary *accuracy*  
Fig. A.1. Chart of the *accuracy* measure



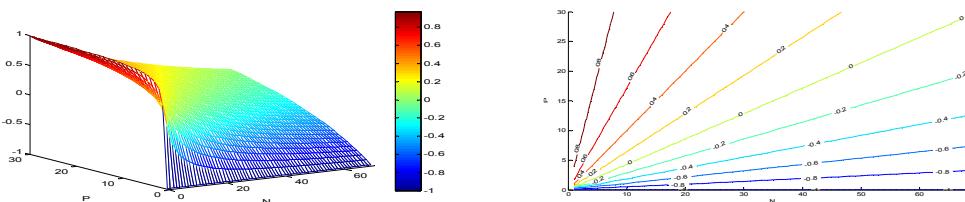
Rys. A.2. Wykres miary *coverage*  
Fig. A.2. Chart of the *coverage* measure



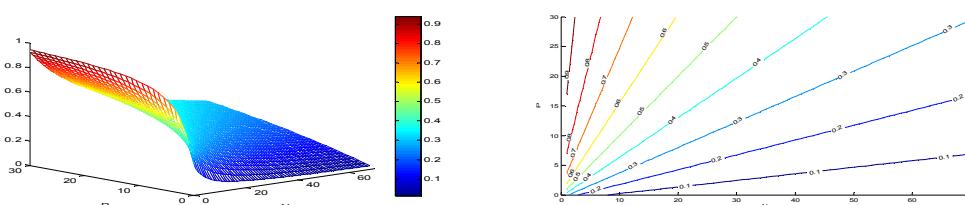
Rys. A.3. Wykres miary *full coverage*  
Fig. A.3. Chart of the *full coverage* measure



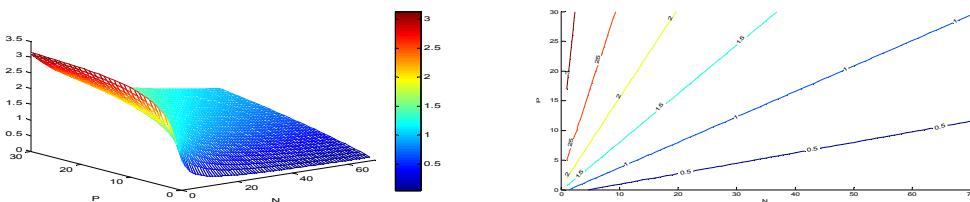
Rys. A.4. Wykres miary *precision*  
Fig. A.4. Chart of the *precision* measure



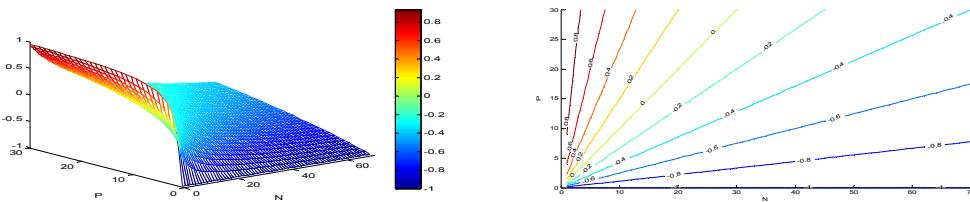
Rys. A.5. Wykres miary *f*  
Fig. A.5. Chart of the *f* measure



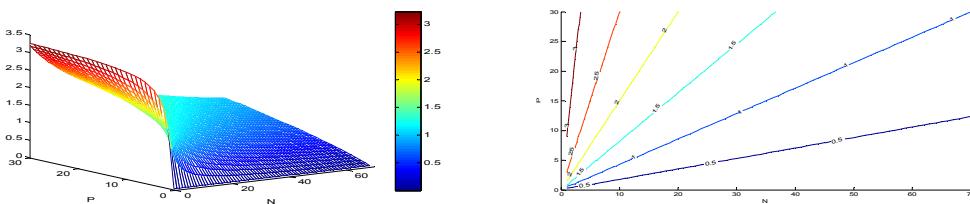
Rys. A.6. Wykres miary *Laplace*  
Fig. A.6. Chart of the *Laplace* measure



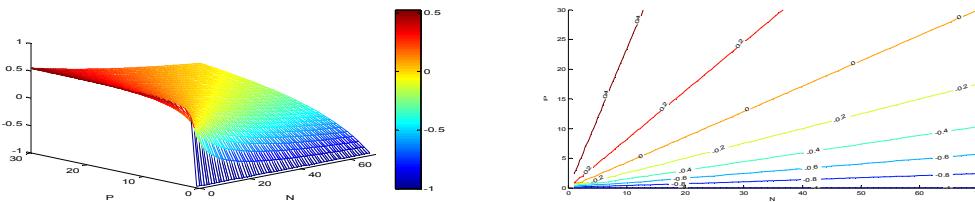
Rys. A.7. Wykres miary *wLap*  
Fig. A.7. Chart of the *wLap* measure



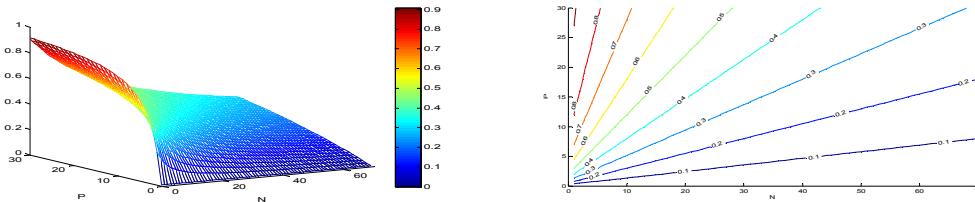
Rys. A.8. Wykres miary *RIPPER*  
Fig. A.8. Chart of the *RIPPER* measure



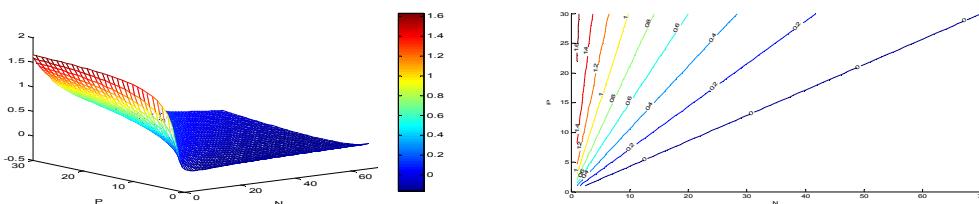
Rys. A.9. Wykres miary *Lift*  
Fig. A.9. Chart of the *Lift* measure



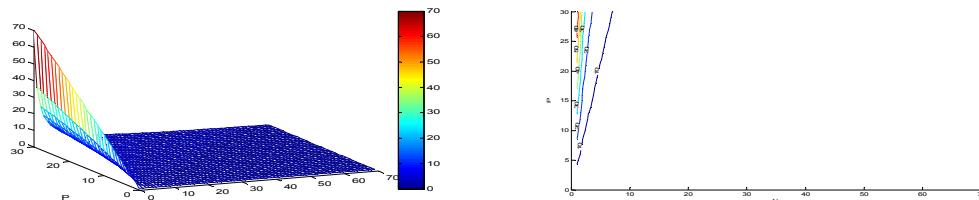
Rys. A.10. Wykres miary *DF*  
Fig. A.10. Chart of the *DF* measure



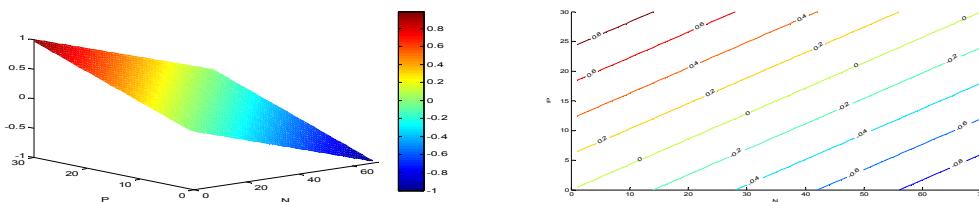
Rys. A.11. Wykres miary *g* (*g*=2)  
Fig. A.11. Chart of the *g* measure



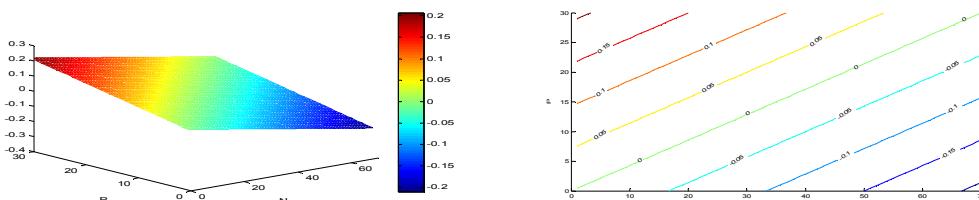
Rys. A.12. Wykres miary *OWS*  
Fig. A.12. Chart of the *OWS* measure



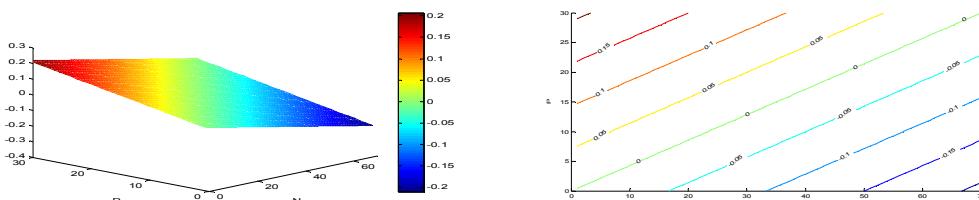
Rys. A.13. Wykres miary *LS*  
Fig. A.13. Chart of the *LS* measure



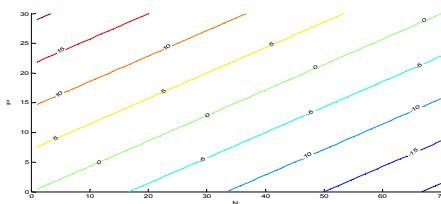
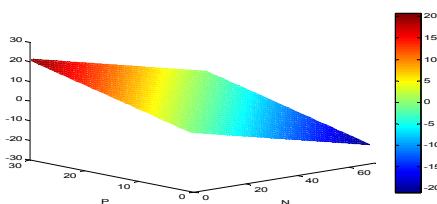
Rys. A.14. Wykres miary *RSS*  
Fig. A.14. Chart of the *RSS* measure



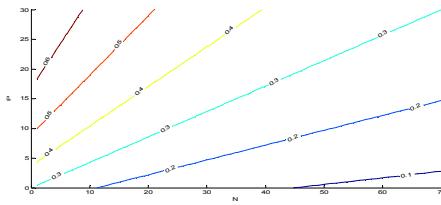
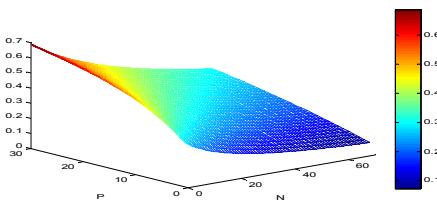
Rys. A.15. Wykres miary *WRA*  
Fig. A.15. Chart of the *WRA* measure



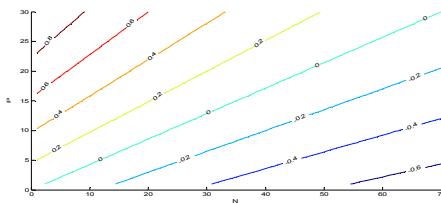
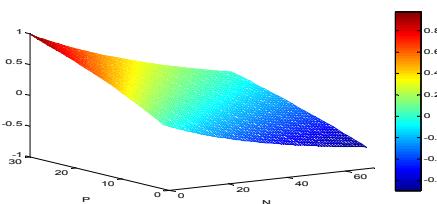
Rys. A.16. Wykres miary *Novelty*  
Fig. A.16. Chart of the *Novelty* measure



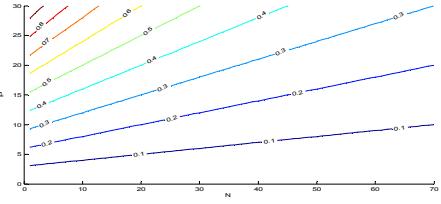
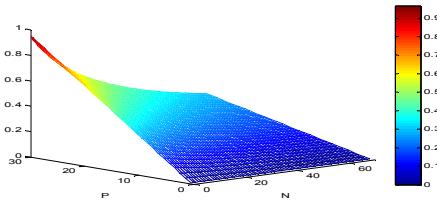
Rys. A.17. Wykres miary  $RI$   
Fig. A.17. Chart of the  $RI$  measure



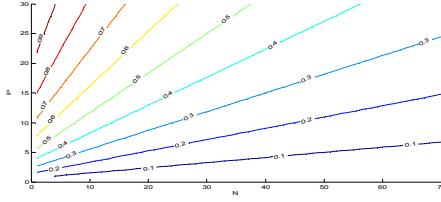
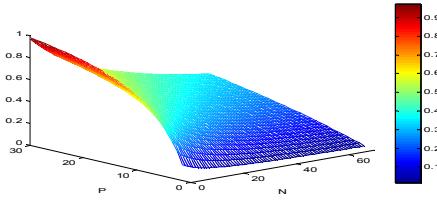
Rys. A.18. Wykres miary  $m$  ( $m=22.466$ )  
Fig. A.18. Chart of the  $m$  measure



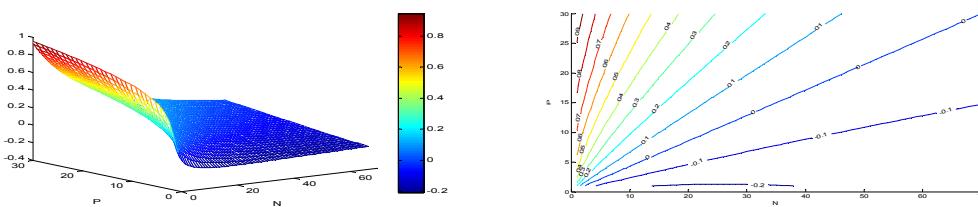
Rys. A.19. Wykres miary  $Cohen$   
Fig. A.19. Chart of the  $Cohen$  measure



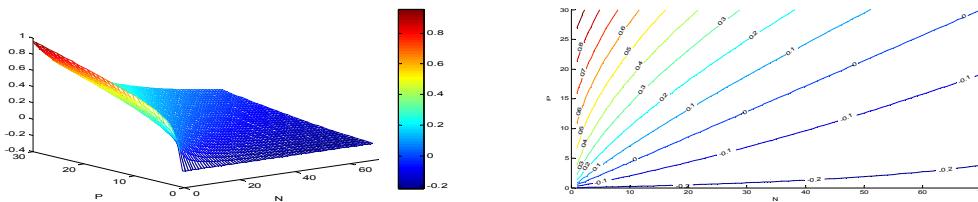
Rys. A.20. Wykres miary  $MS$   
Fig. A.20. Chart of the  $MS$  measure



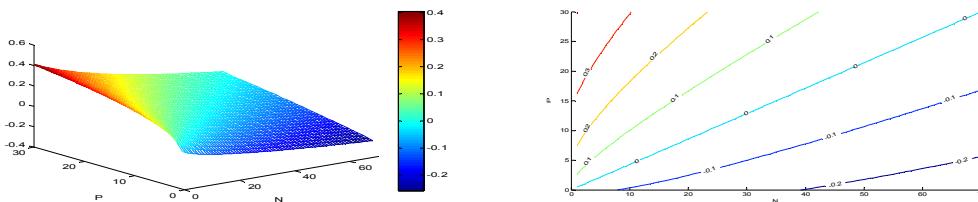
Rys. A.21. Wykres miary  $F$  ( $\beta=0.5$ )  
Fig. A.21. Chart of the  $F$  measure



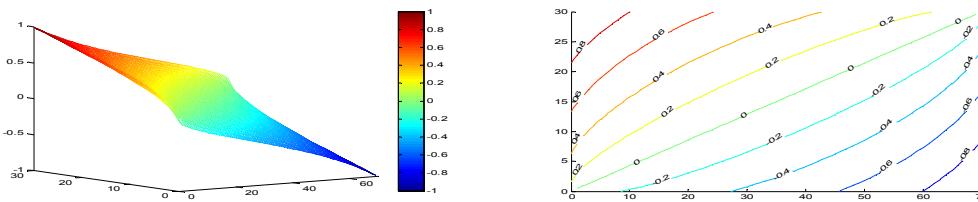
Rys. A.22. Wykres miary  $C1$   
Fig. A.22. Chart of the  $C1$  measure



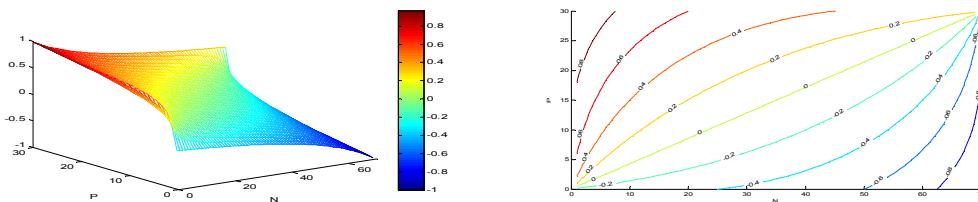
Rys. A.23. Wykres miary  $C2$   
Fig. A.23. Chart of the  $C2$  measure



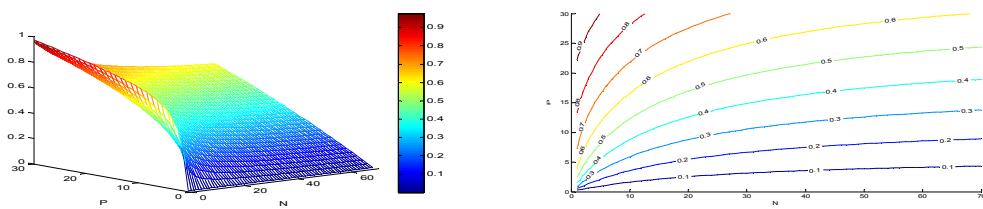
Rys. A.24. Wykres miary  $Klösgen$  ( $\omega=0.4323$ )  
Fig. A.24. Chart of the  $Klösgen$  measure



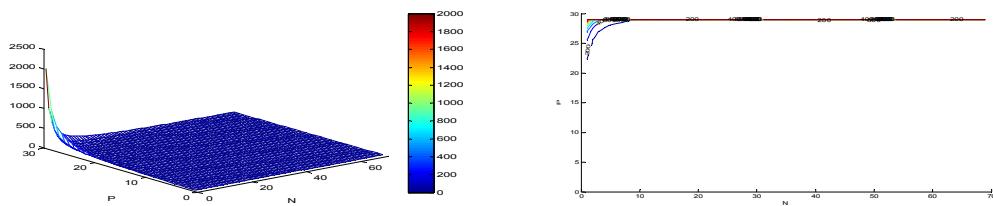
Rys. A.25. Wykres miary  $Corr$   
Fig. A.25. Chart of the  $Corr$  measure



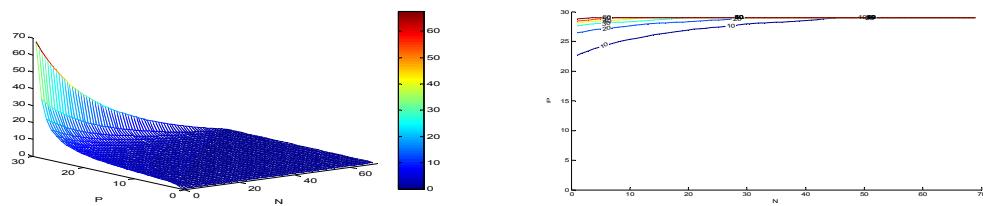
Rys. A.26. Wykres miary  $s$   
Fig. A.26. Chart of the  $s$  measure



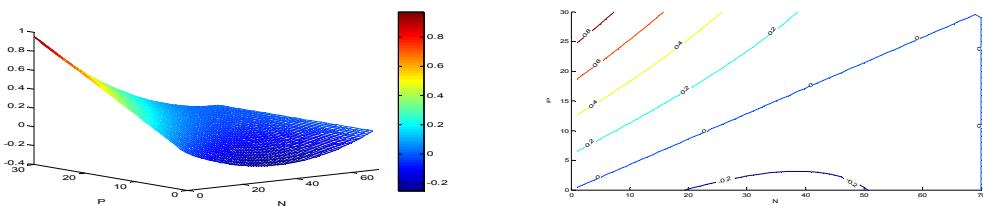
Rys. A.27. Wykres miary *YAILS*  
Fig. A.27. Chart of the *YAILS* measure



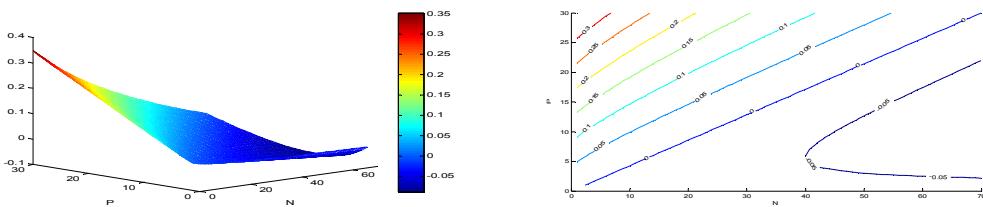
Rys. A.28. Wykres miary *Odds*  
Fig. A.28. Chart of the *Odds* measure



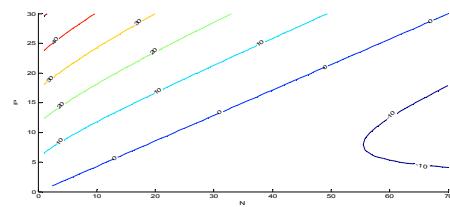
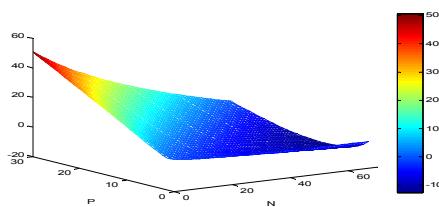
Rys. A.29. Wykres miary *RR*  
Fig. A.29. Chart of the *RR* measure



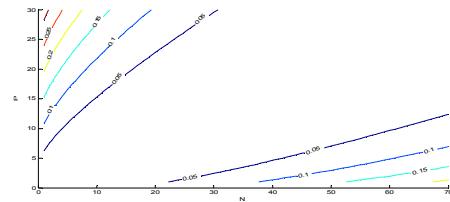
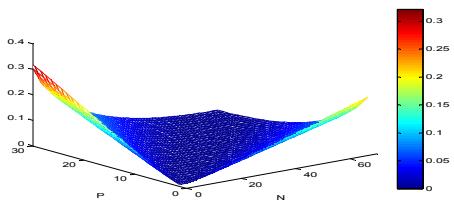
Rys. A.30. Wykres miary *Q2*  
Fig. A.30. Chart of the *Q2* measure



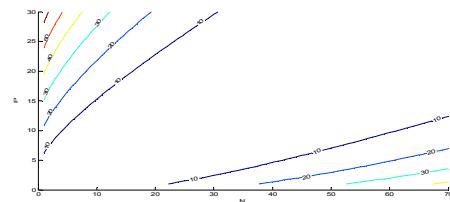
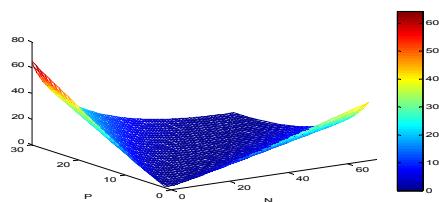
Rys. A.31. Wykres miary *TWS*  
Fig. A.31. Chart of the *TWS* measure



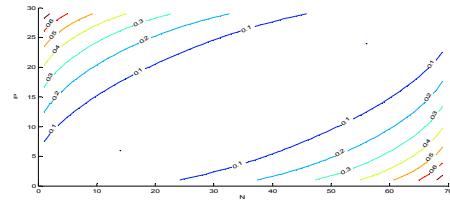
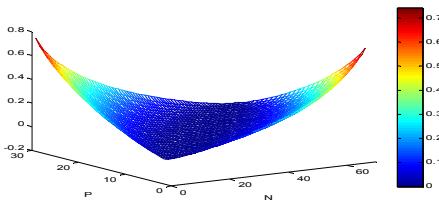
Rys. A.32. Wykres miary *cFoil*  
Fig. A.32. Chart of the *cFoil* measure



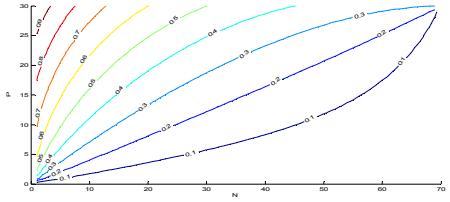
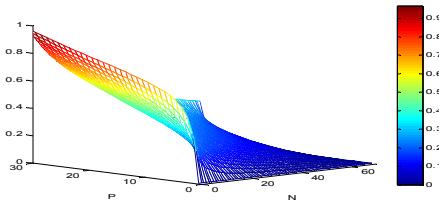
Rys. A.33. Wykres miary *J*  
Fig. A.33. Chart of the *J* measure



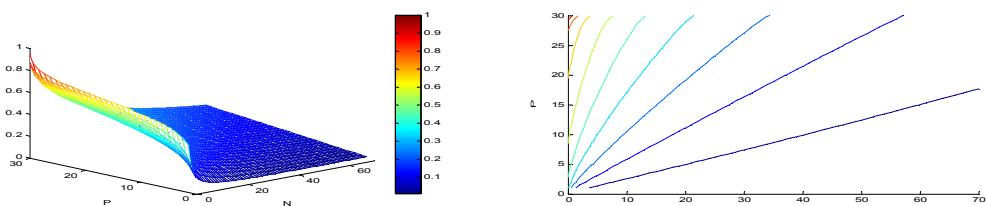
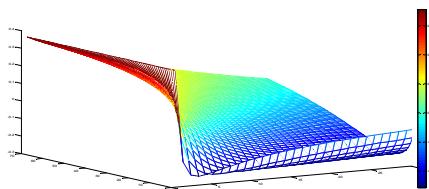
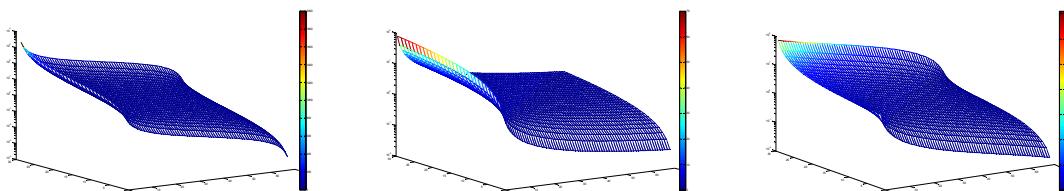
Rys. A.34. Wykres miary *CN2*  
Fig. A.34. Chart of the *CN2* measure



Rys. A.35. Wykres miary *Gain*  
Fig. A.35. Chart of the *Gain* measure



Rys. A.36. Wykres miary *E<sup>φ</sup>*  
Fig. A.36. Chart of the *E<sup>φ</sup>* measure

Rys. A.37. Wykres miary  $nMM(C1, Odds, C2, g)$ Fig. A.37. Chart of the  $nMM(C1, Odds, C2, g)$  measureRys. A.38. Wykres miary  $OWS$  dla zbioru zawierającego 70 przykładów pozytywnych i 30 przykładów negatywnychFig. A.38. Graph of the  $OWS$  measure for the set containing 70 positive examples and 30 negative examplesRys. A.39. Wykresy miar  $Odds, LS, RR$  – skala logarytmicznaFig. A.39. Graphs of the measures:  $Odds, LS, RR$  – logarithmic scale



## **DODATEK B. PRZYKŁAD INDUKCJI REGUŁ REGRESYJNYCH**

W tabeli B.1 zaprezentowano wyniki zastosowania miar jakości należących do zbioru *Max* w pokryciowym algorytmie indukcji reguł regresyjnych. Algorytm działa na zasadzie identycznej z q-ModLEM. W konkluzji reguły znajduje się wyrażenie  $d = Me$ , gdzie *Me* jest medianą wartości atrybutu decyzyjnego, wyznaczoną na podstawie przykładów pokrywanych przez regułę. Podczas indukcji reguły zbiór wszystkich przykładów pozytywnych stanowią przykłady, których wartości atrybutu decyzyjnego mieszczą się w przedziale  $[Me - \sigma, Me + \sigma]$ . Liczba  $\sigma$  jest odchyleniem standardowym od *Me*. Przykłady, których wartości atrybutu decyzyjnego nie mieszczą się w przedziale  $[Me - \sigma, Me + \sigma]$ , stanowią zbiór przykładów negatywnych. Na podobnej zasadzie wyznacza się zbiór przykładów pozytywnych pokrywanych przez regułę i zbiór przykładów negatywnych pokrywanych przez regułę. Sposób definiowania zbiorów przykładów pozytywnych i negatywnych oznacza, że liczebność klasy decyzyjnej zmienia się w fazie wzrostu (wraz z konkretyzacją kolejnych warunków elementarnych) oraz w fazie przycinania (wraz z usuwaniem warunków zbędnych). Niemniej jednak przedstawiona strategia pozwala na określenie wartości  $p, n, P, N$  na każdym etapie budowy reguły, a to z kolei pozwala na obliczenie wartości każdej z zawartych w zbiorze *Max* miar jakości.

Podczas „klasyfikacji” stosowano strategię polegającą na przyporządkowaniu przykładowi testowemu średniej arytmetycznej z wartości *Me*, zawartych w konkluzjach pokrywających go reguły. Jeżeli przykład nie był pokrywany przez żadną z reguł, to wartość atrybutu decyzyjnego była dla tego przykładu równa medianie wartości atrybutu decyzyjnego w zbiorze treningowym.

Rezultaty prezentowane w tabeli B.1 otrzymano po zastosowaniu 10-krotnej walidacji krzyżowej. Analizie poddano 35 zbiorów danych pochodzących z bazy UCI. Do eksperymentów użyto następujących zbiorów danych: auto93, auto-mpg, auto-price, baseball, bodyfat, breasttumor, cholesterol, cloud, compressive, concrete, cpu, dee, diabetes, echomonths, ele-1, ele-2, elusage, fishcatch, friedman, fruitfly, housing, kidney, laser, lowbwt, machine, mbagrade, meta, pbc, pharynx, pollution, pyrim, sensory, strike, triazines, veteran.

Wymienione zbiory opisywane są przez różne wektory cech (tylko symboliczne, tylko numeryczne i mieszane) oraz charakteryzują się odmiennymi rozkładami wartości atrybutu decyzyjnego.

Kolumny oznaczone jako *prec.* i *full cov.* oznaczają odpowiednio średnią dokładność reguł oraz ich średnie całkowite pokrycie.

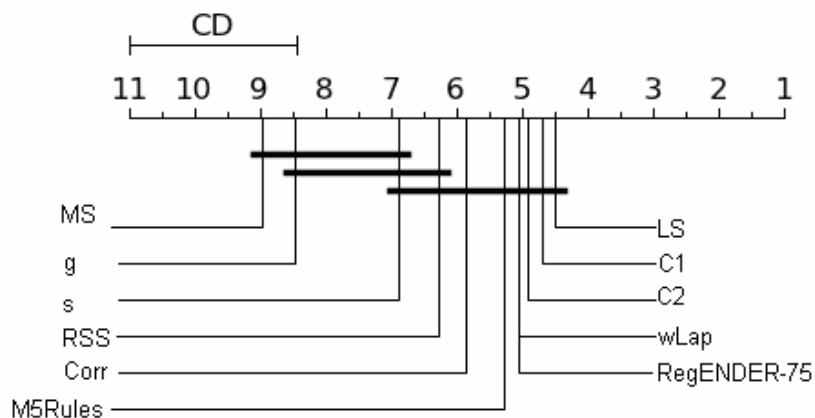
Tabela B.1  
Wyniki indukcji reguł regresyjnych,  
otrzymane za pomocą algorytmu  
pokryciowego i miar jakości, należących  
do zbioru *Max*

Miara	Indukcja reguł			
	rRMSE	Reguły		
		liczba	<i>prec.</i>	<i>full cov.</i>
<i>LS</i>	0,7253	73	0,82	0,07
<i>C1</i>	0,7331	50	0,87	0,11
<i>wLap</i>	0,7377	75	0,75	0,06
<i>C2</i>	0,7477	37	0,86	0,15
<i>Corr</i>	0,7765	20	0,79	0,24
<i>RSS</i>	0,7850	15	0,76	0,29
<i>s</i>	0,8215	29	0,83	0,27
<i>g (g=2)</i>	0,8861	15	0,87	0,29
<i>MS</i>	0,9078	5	0,78	0,56
M5Rules	0,7730	9		
RegENDER-75	0,7685	75		

Rezultaty algorytmu pokryciowego skonfrontowano z rezultatami dwóch algorytmów indukcji reguł regresyjnych. Pierwszy z nich to algorytm M5, w którym wyłączono fazę budowy modelu regresyjnego w konkluzjach reguł, a drugi to przeznaczona do indukcji reguł regresyjnych modyfikacja algorytmu ENDER [62]. Algorytm RegENDER stosuje metodykę *boosting* i uzyskuje bardzo dobore wyniki prognoz. W algorytmie RegENDER liczbę tworzonych reguł arbitralnie ustalono na 75.

Wyniki otrzymane przez miary jakości, zawarte w zbiorze *Max*, potwierdzają efektywność tych miar także w nadzorowaniu procesu indukcji reguł regresyjnych. Wyjątek stanowi miara *g* z wartością parametru *g=2*. Wynik ten nie jest zaskakujący, gdyż wartość parametru dobierana była na podstawie publikacji opisujących próby rzeczywistego oszacowania dokładności reguł definiowanych dla celów klasyfikacji [88,146]. Pozostałe miary jakości zachowują się podobnie jak w realizacji zadań klasyfikacyjnych. Można zauważyc podobną zależność pomiędzy błędem, liczbą oraz dokładnością i pokryciem (tym razem całkowitym pokryciem) reguł wyznaczanych za pomocą poszczególnych miar.

Porównanie efektywności algorytmu pokryciowego z innymi algorytmami pokazuje (rys. B.1), że indukcja reguł regresyjnych za pomocą algorytmu pokryciowego jest interesującym kierunkiem dalszych badań. Uzyskane rezultaty uzasadniają celowość podjęcia dalszych badań nad zidentyfikowaniem zbioru miar najbardziej efektywnych z punktu widzenia indukcji reguł regresyjnych.



Rys. B.1. Porównanie zbiorów reguł utworzonych za pomocą algorytmu pokryciowego, stosującego miary jakości zawarte w zbiorze *Max*, oraz algorytmów M5Rules i RegENDER. Porównanie na podstawie 35 zbiorów danych. Poziom istotności 0.05

Fig. B.1. Comparison of the rule sets get by the sequential covering algorithm that uses the quality measures included in the *Max* set, and the algorithms M5Rules and RegENDER. The comparison was based on 35 data sets with the significance level equal to 0.05



## SPIS RYSUNKÓW

Rys. 2.1.	Krzywa ROC klasyfikatora dyskretnego .....	56
Rys. 3.1.	Przestrzeń wartości miary .....	77
Rys. 3.2.	Wykres miary <i>precision</i> .....	78
Rys. 3.3.	Wizualizacja procesu tworzenia reguły w przestrzeni pokrycia .....	79
Rys. 3.4.	Wykres miary $p_{val}$ .....	121
Rys. 3.5.	Wykres izolinii miar $E^\psi$ i $E^\varphi$ .....	125
Rys. 4.1.	Ilustracja fazy wzrostu i przycinania reguły .....	149
Rys. 4.2.	Ilustracja zależności pomiędzy liczbą wyznaczanych reguł a iloczynem ich średniej dokładności i średniego pokrycia.....	153
Rys. 4.3.	Ilustracja zależności pomiędzy liczbą wyznaczanych reguł a średnią liczbą warunków elementarnych .....	155
Rys. 4.4.	Schemat adaptacyjnego doboru miary jakości w pokryciowym algorytmie indukcji reguł .....	162
Rys. 4.5.	Porównanie klasyfikatorów w grupie 34 zbiorów treningowych .....	164
Rys. 4.6.	Porównanie klasyfikatorów w grupie 14 zbiorów testowych .....	165
Rys. 4.7.	Porównanie dwóch schematów klasyfikacji: głosowanie i największe zaufanie .....	169
Rys. 4.8.	Porównanie całkowitej dokładności klasyfikacji pokryciowych algorytmów indyukcji reguł – 34 zbiory treningowe .....	171
Rys. 4.9.	Porównanie całkowitej dokładności klasyfikacji pokryciowych algorytmów indyukcji reguł – 14 zbiorów testowych .....	171
Rys. 4.10.	Diagram ilustrujący podobieństwo pomiędzy grupami miar .....	177
Rys. 4.11.	Zależność pomiędzy wartościami miar <i>DF</i> i <i>Lift</i> .....	179
Rys. 5.1.	Przykład agregacji reguł zbudowanych z ciągłych warunków elementarnych... <td>200</td>	200
Rys. 5.2.	Reguły przed i po agregacji .....	206
Rys. 5.3.	Wizualizacja procesu dostrajania zagregowanej reguły .....	206
Rys. 5.4.	Zasada tworzenia reguł w zbiorze danych syntetycznych .....	213
Rys. 5.5.	Średnia procentowa redukcja liczby reguł po zastosowaniu filtracji.....	234
Rys. 6.1.	Typowa infrastruktura stacji geofizyki górniczej .....	240
Rys. 6.2.	Metodyka tworzenia klasyfikatora prognozującego zagrożenia sejsmiczne.....	249
Rys. 6.3.	Krzywe przeżycia dla przykładów pokrywanych przez reguły R1–R4.....	263
Rys. 6.4.	Krzywe przeżycia reguł R1 i R3 oraz ich dopełnień .....	264
Rys. 6.5.	Fragment ontologii BP .....	269



## SPIS TABEL

Tabela 3.1	Tablica kontyngencji dla reguły $r \equiv \varphi \rightarrow \psi$ .....	76
Tabela 3.2	Metody normalizacji miar konfirmacji .....	85
Tabela 3.3	Własności związane z monotonicznością miar .....	90
Tabela 3.4	Własności miar związane z oceną zdolności opisowych reguł decyzyjnych ..	94
Tabela 3.5	Wybrane miary oceniające jakości reguł decyzyjnych .....	110
Tabela 3.6	Tablica kontyngencji dla wektora progów tolerancji .....	118
Tabela 3.7	Tablica kontyngencji przeznaczona do badania podobieństwa reguł .....	129
Tabela 4.1	Charakterystyka zbiorów danych użytych do badań .....	151
Tabela 4.2	Charakterystyka klasyfikatorów regułowych wyznaczonych przez algorytm q-ModLEM(MMM) .....	152
Tabela 4.3	Korelacja pomiędzy liczbą reguł a iloczynem ich średniej dokładności i średniego pokrycia .....	154
Tabela 4.4	Liczba zwycięstw i porażek na 34 treningowych zbiorach danych .....	156
Tabela 4.5	Grupy miar najbardziej obiecujących .....	158
Tabela 4.6	p-wartość testu Wilcoxona pomiędzy najlepszymi miarami ze względu na całkowitą dokładność klasyfikacji ( <i>Acc</i> ) .....	159
Tabela 4.7	p-wartość testu Wilcoxona pomiędzy najlepszymi miarami ze względu na średnią dokładność klas decyzyjnych ( <i>BAcc</i> ) .....	159
Tabela 4.8	p-wartość testu Wilcoxona pomiędzy najlepszymi miarami ze względu na <i>AvC</i> .....	159
Tabela 4.9	Liczba porażek dla: najlepszych miar, grup miar najbardziej obiecujących oraz miary złożonej ( <i>Acc</i> ) .....	161
Tabela 4.10	Liczba porażek dla: najlepszych miar, grup miar najbardziej obiecujących oraz miary złożonej ( <i>BAcc</i> ) .....	161
Tabela 4.11	Liczba porażek dla: najlepszych miar, grup miar najbardziej obiecujących oraz miary złożonej ( <i>AvC</i> ) .....	161
Tabela 4.12	Wyniki szczegółowe dla najlepszych miar, miary złożonej oraz metody <i>AMS</i> – optymalizacja i ocena klasyfikatora ze względu na <i>Acc</i> .....	163
Tabela 4.13	Wyniki szczegółowe dla najlepszych miar, miary złożonej oraz metody <i>AMS</i> – optymalizacja i ocena klasyfikatora ze względu na <i>BAcc</i> .....	163
Tabela 4.14	Wyniki szczegółowe dla najlepszych miar, miary złożonej oraz metody <i>AMS</i> – optymalizacja i ocena klasyfikatora ze względu na <i>AvC</i> .....	163
Tabela 4.15	p-wartość testu Wilcoxona pomiędzy metodą <i>AMS-Max</i> a najlepszymi miarami jakości .....	164
Tabela 4.16	Charakterystyka klasyfikatorów regułowych wyznaczonych przez algorytm q-ModLEM(EMM) – 34 zbiory treningowe – głosowanie .....	166
Tabela 4.17	Dokładność klasyfikacji klasyfikatorów regułowych wyznaczonych przez q-ModLEM(EMM) – 34 zbiory treningowe – klasyfikacja wg największego zaufania .....	167

Tabela 4.18	Algorytm q-ModLEM(EMM) – wyniki szczegółowe dla najlepszych miar jakości oraz metody adaptacyjnej.....	169
Tabela 4.19	Porównanie dokładności klasyfikacji adaptacyjnych wersji algorytmu q-ModLEM z innymi algorytmami pokryciowymi.....	170
Tabela 4.20	Grupy miar podobnych.....	173
Tabela 4.21	Minimalna korelacja pomiędzy reprezentacjami porządków reguł utworzonych na podstawie miar należących do grupy 3.....	173
Tabela 4.22	Minimalna korelacja pomiędzy reprezentacjami porządków reguł utworzonych na podstawie miar należących do grupy 4.....	173
Tabela 4.23	Minimalna korelacja pomiędzy reprezentacjami porządków reguł utworzonych na podstawie miar należących do grupy 5.....	174
Tabela 4.24	Minimalna korelacja pomiędzy reprezentacjami porządków reguł utworzonych na podstawie miar należących do grupy 6.....	174
Tabela 4.25	Minimalna korelacja pomiędzy reprezentacjami porządków reguł utworzonych na podstawie miar należących do grupy 8.....	174
Tabela 4.26	Minimalne i maksymalne korelacje pomiędzy reprezentacjami porządków reguł.....	175
Tabela 4.27	Przykłady ilustrujące brak równoważności miar ze względu na sposób rozstrzygania konfliktów klasyfikacji.....	181
Tabela 4.28	Analiza własności miar definiowanych na podstawie tablicy kontyngencji..	185
Tabela 4.29	Zwycięstwa miar zawartych w zbiorze $Max \cup \{CN2\}$ ze względu na całkowitą dokładność klasyfikacji.....	191
Tabela 5.1	Wyniki agregacji reguł wyznaczonych przez algorytm RIPPER (5 3 -), max-dim=5.....	209
Tabela 5.2	Rezultaty agregacji reguł wyznaczonych przez algorytm RIPPER .....	209
Tabela 5.3	Wyniki agregacji reguł wyznaczonych przez algorytm q-ModLEM (0 - 0), maxdim=3.....	210
Tabela 5.4	Redukcja liczby reguł wyznaczonych przez algorytm q-ModLEM.....	211
Tabela 5.5	Rezultaty agregacji reguł wyznaczonych przez algorytm q-ModLEM.....	211
Tabela 5.6	Wynik agregacji reguł wyznaczonych na podstawie syntetycznych zbiorów danych .....	214
Tabela 5.7	Rezultaty redefinicji reguł – dokładność klasyfikacji.....	223
Tabela 5.8	Rezultaty redefinicji reguł – liczba reguł .....	224
Tabela 5.9	Rezultaty redefinicji reguł dla różnych konfiguracji algorytmu .....	225
Tabela 5.10	Wartość współczynnika korelacji $\tau$ Kendalla pomiędzy rankinami warunków elementarnych utworzonych przez podstawowe, uproszczone i zmodyfikowane postaci indeksu Banzhafa.....	227
Tabela 5.11	Wyniki filtracji zbiorów reguł utworzonych przez algorytm q-ModLEM(MMM) .....	233
Tabela 5.12	Wyniki filtracji zbiorów reguł utworzonych przez algorytm q-ModLEM(EMM) .....	233
Tabela 5.13	Wyniki filtracji – zmiany liczby przykładów pokrywanych unikalnie oraz poziomu statystycznej istotności reguł.....	235
Tabela 6.1	Charakterystyka analizowanych zbiorów danych .....	242
Tabela 6.2	Wyniki klasyfikacji zagrożeń sejsmicznych – indukcja i filtracja.....	243
Tabela 6.3	Wyniki klasyfikacji zagrożeń sejsmicznych – indukcja, filtracja i redefinicja reguł.....	244

Tabela 6.4	Porównanie dokładności prognoz – klasyfikator regułowy vs metoda rutynowa.....	245
Tabela 6.5	Wybrane atrybuty warunkowe opisujące pacjentów.....	259
Tabela 6.6	Charakterystyka klasyfikatorów utworzonych na podstawie różnych miar jakości .....	261
Tabela 6.7	Porównanie jakości klasyfikatorów utworzonych przez różne algorytmy indukcji reguł i drzew decyzyjnych.....	261
Tabela 6.8	Porównanie wartości wskaźnika Briera dla modeli przeżycia otrzymanych na podstawie algorytmów indukcji drzew i reguł przeżycia ....	266
Tabela 6.9	System informacyjny utworzony na podstawie analizy struktury ontologii genowej przedstawionej na rysunku 6.5.....	270
Tabela 6.10	Wyniki indukcji reguł dla problemu opisu grup genów za pomocą reguł decyzyjnych .....	278
Tabela 6.11	Porównanie efektywności reguł utworzonych przez serwis Genecodis i modyfikowaną wersję algorytmu Explore uzupełnioną o algorytm filtracji.....	279
Tabela 6.12	Wyniki redefinicji reguł na podstawie informacji o ważności terminów GO znajdujących się w przesłankach reguł.....	280



# **WYBRANE METODY OCENY I PRZCINANIA REGUŁ DECYZYJNYCH**

## **STRESZCZENIE**

Pierwsza część monografii poświęcona jest pokryciowym algorytmom indukcji reguł i obiektywnym miarom oceny jakości reguł decyzyjnych. Przedstawiono dwa algorytmy, które indukcję reguł prowadzą w kierunku maksymalizacji wartości miar przeznaczonych do oceny reguł decyzyjnych. Dokonano analizy własności miar definiowanych na podstawie tablicy kontyngencji i na tej podstawie określono minimalne zbiory własności, pożądane dla miar nadzorujących proces indukcji oraz oceniających zdolności opisowe reguł decyzyjnych. Przeprowadzono analizę równoważności i podobieństwa miar. Równoważność analizowano zarówno ze względu na uporządkowanie reguł, jak i na sposób rozstrzygania konfliktów klasyfikacji. W części eksperimentalnej zweryfikowano efektywności miar, zidentyfikowano zbiory miar najbardziej efektywnych oraz zaproponowano adaptacyjną metodę doboru miary w algorytmie indukcji reguł. Przeprowadzono także analizę własności teoretycznych najefektywniejszych miar. W pierwszej części pracy omówiono również miary niedefiniowane bezpośrednio na podstawie tablicy kontyngencji. Przedyskutowano możliwość złożonej oceny reguł, a także przedstawiono propozycję wielokryterialnej oceny reguł na podstawie tzw. funkcji użyteczności.

Druga część monografii koncentruje się na wybranych metodach przycinania reguł. W części tej zaprezentowano dwa algorytmy agregacji reguł, algorytm redefinicji reguł na podstawie informacji o ważności tworzących je warunków elementarnych oraz cztery algorytmy filtracji reguł. Dzięki agregacji i redefinicji w przesłankach reguł mogą pojawić się złożone warunki elementarne, co w szczególnych przypadkach lepiej odzwierciedla zależności, jakimi charakteryzują się dane.

Efektywność wszystkich algorytmów proponowanych w częściach pierwszej i drugiej zweryfikowano eksperimentalnie.

Ostatnia część publikacji przedstawia przykłady nowych zastosowań algorytmów indukcji reguł decyzyjnych. Zaprezentowano trzy nowe obszary zastosowań: prognozowanie zagrożeń sejsmicznych, analizę danych okołoprzeszczepowych oraz funkcjonalny opis genów. W zastosowaniach tych wykorzystano rezultaty badań przedstawionych w częściach pierwszej i drugiej. W ostatniej części monografii przedstawiono także dwie, ukierunkowane dziedzinowo, modyfikacje algorytmów indukcji reguł. Pierwsza z nich umożliwia indukcję reguł sterowaną hipotezami definiowanymi przez użytkownika. Druga dostosowuje algorytm indukcji do hierarchicznej struktury analizowanych danych. Rezultatem badań nad funkcjonalnym opisem genów jest także metoda redukcji atrybutów, biorąca pod uwagę semantykę ich wartości.



# **SELECTED METHODS OF DECISION RULE EVALUATION AND PRUNING**

## **ABSTRACT**

The first part of the book is devoted to sequential covering rule induction algorithms and objective rule evaluation measures. Two algorithms that maximize values of rule evaluation measures are presented. The properties of measures defined on the basis of the contingency table were analyzed and minimal sets of the properties desired for measures controlling the process of rule induction and evaluating descriptive quality of decision rules were specified. The analysis of equivalence and similarity of measures was carried out. The equivalence was analyzed both due to the rule ordering and the classification conflicts resolving. In the experimental part the efficiency of measures was verified, sets of the most efficient measures were identified. The adaptive method of measure selection in sequential covering rule induction algorithm was proposed. Moreover, theoretical properties of most effective measures were analyzed. In the first part of the paper measures that are not defined directly from the contingency table were also discussed. Furthermore, the possibility of complex evaluation of rules was discussed and the proposal of multi-criteria rule assessment on the basis of so-called utility function was presented.

The second part of the book focuses on algorithms of rule pruning. This part presents two algorithms of rule aggregation, the algorithm of rule redefinition based on information about the importance of the rule elementary conditions, and four algorithms of rule filtration. Through aggregation and redefinition complex elementary conditions may appear in rule premises which, in specific cases, better reflect dependencies in data.

The effectiveness of all the algorithms proposed in the first and second parts was verified experimentally.

The last part of the work shows examples of new applications of the decision rule induction algorithms. The following three new areas of application are presented: forecasting of seismic hazards, analysis of bone marrow transplantation data and functional description of genes. The results of study presented in the first two parts of the book were used there. Two domain-oriented modifications of rule induction algorithms are proposed. The first one allows for the rule induction controlled by hypothesis defined by the user. The second adjusts the induction algorithm to the hierarchical structure of the analyzed data. The method of attribute reduction that takes into consideration the semantics of the attributes values is also the result of research on the functional description of genes.

## **INFORMATION FOR AUTHORS**

The journal *STUDIA INFORMATICA* publishes both fundamental and applied Memoirs and Notes in the field of informatics. The Editors' aim is to provide an active forum for disseminating the original results of theoretical research and applications practice of informatics understood as a discipline focused on the investigations of laws that rule processes of coding, storing, processing, and transferring of information or data.

Papers are welcome from fields of informatics inclusive of, but not restricted to *Computer Science, Engineering, and Life and Physical Sciences*.

All manuscripts submitted for publication will be subject to critical review. Acceptability will be judged according to the paper's contribution to the art and science of informatics.

In the first instance, all text should be submitted as hardcopy, conventionally mailed, and for accepted paper accompanying with the electronically readable manuscript to:

**Dr. Marcin SKOWRONEK**  
Institute of Informatics  
Silesian University of Technology  
ul. Akademicka 16  
44-100 Gliwice, Poland  
Tel.: +48 32 237-12-15  
Fax: +48 32 237-27-33  
e-mail: marcin.skowronek@polsl.pl

## **MANUSCRIPT REQUIREMENTS**

All manuscripts should be written in Polish or in English. Manuscript should be typed on one side paper only, and submitted in duplicate. The name and affiliation of each author should be followed by the title of the paper (as brief as possible). An abstract of not more than 50 words is required. The text should be logically divided under numbered headings and subheadings (up to four levels). Each table must have a title and should be cited in the text. Each figure should have a caption and have to be cited in the text. References should be cited with a number in square brackets that corresponds to a proper number in the reference list. The accuracy of the references is the author's responsibility. Abbreviations should be used sparingly and given in full at first mention (e.g. "Central Processing Unit (CPU)"). In case when the manuscript is provided in Polish (English) language, the summary and additional abstract (up to 300 words with reference to the equations, tables and figures) in English (Polish) should be added.

After the paper has been reviewed and accepted for publication, the author has to submit to the Editor a hardcopy and electronic version of the manuscript.

It is strongly recommended to submit the manuscript in a form downloadable from web site <http://zti.polsl.pl/makiety/>.

**To subscribe:** *STUDIA INFORMATICA* (PL ISSN 0208-7286) is published by Silesian University of Technology Press (Wydawnictwo Politechniki Śląskiej) ul. Akademicka 5, 44-100 Gliwice, Poland, Tel./Fax +48 32 237-13-81. 2012 annual subscription rate: US\$60. Single number price approx. US\$10-20 according to the issue volume.

## **INSTYTUT INFORMATYKI prowadzi:**

- Studia stacjonarne I stopnia (inżynierskie)
- Studia stacjonarne II stopnia (magisterskie)
- Studia niestacjonarne I stopnia (inżynierskie)
- Studia niestacjonarne II stopnia (magisterskie)
- Studia podyplomowe:
  - *Sieci i systemy komputerowe, bazy danych*
  - *Systemy informacji geograficznej*
  - *Teleinformatyka w transporcie lotniczym*
  - *Technologie internetowe i technologie mobilne*
  - *Metody eksploracji baz danych przedsiębiorstw*
- Studia doktoranckie

## **Informacje:**

**POLITECHNIKA ŚLĄSKA  
Instytut Informatyki**

44-100 Gliwice, ul. Akademicka 16  
tel. (32) 237 24 05; 237 21 51;  
fax (32) 237 27 33 (czynny całą dobę)

e-mail: [rau2@polsl.pl](mailto:rau2@polsl.pl)

<http://www.inf.polsl.pl> (dydaktyka)