Zbiory przybliżone w obszarze systemów ekspertowych

Agnieszka Nowak

Institute of Computer Science, University of Silesia
Będzińska 39, 41–200 Sosnowiec, Poland
e-mail: nowak@us.edu.pl

1 Wprowadzenie

Okres ostatnich kilkunastu lat to dość szybki rozwój technologii informatycznych. Przyczyną tego niewątpliwie stało się ciągle rosnące zapotrzebowanie na przechowywanie i przetwarzanie ogromnych zbiorów danych, z jakimi mamy do czynienia w dziedzinie medycyny, finansów czy marketingu. Tak duża ilość zgromadzonych informacji bez odpowiedniej analizy jest jednak bezużyteczna. Niezwykle pomocne stają się więc nowe metody naukowe nazywane eksploracją danych oraz odkrywaniem wiedzy *(ang. knowledge discovery)*. Dzięki nim jesteśmy w stanie pozyskać użyteczną, z określonego punktu widzenia, wiedzę z posiadanych danych. Do procesu odkrywania wiedzy możemy użyć zbiorów przybliżonych, które umożliwiają m.in. indukcję reguł decyzyjnych czy też redukcję zbiorów danych. Teoria zbiorów przybliżonych zaproponowana przez *Pawlaka* stała się niezwykle pomocna w kwestii właśnie kontroli nad dużymi zbiorami danych, których prawidłowa analiza pozwala na skuteczne działanie systemu te dane wykorzystującego. Zjawisko gromadzenia bardzo dużej ilości informacji spowodowało konieczność budowy systemów automatycznego wnioskowania, które zrealizowane na maszynach cyfrowych mogłyby wyręczyć człowieka przy gromadzeniu i analizowaniu informacji potrzebnej do podejmowania decyzji. Systemy takie, często nazywane również systemami ekspertowymi znajdują szerokie zastosowanie przy rozwiązywaniu wielu problemów pojawiających się w takich dziedzinach jak rozpoznawanie obrazów czy uczenie się maszyn. Nie słabnie wdrażanie projektów badawczych i komercyjnych zmierzających do wytworzenia systemów komputerowych, umożliwiających automatyczne podejmowanie decyzji. Powstawanie tych systemów jest w pewnym sensie wymuszone ogromną liczbą ich potencjalnych zastosowań w wielu dziedzinach życia. Systemy ekspertowe są jednym z wielu zagadnień sztucznej inteligencji, jakie omawiane są w ramach zajęć na kierunku informatyka. Celem tych zajęć jest wykazanie możliwości tychże systemów, które mogą przecież zastąpić eksperta w danej dziedzinie lub przynajmniej wspomóc go w podjęciu decyzji dotyczącej rozważanego problemu. Techniki analizy danych proponowane przez drzewa decyzyjne, metody statystyczne, sieci neuronowe czy programowanie w logice pozwalają podejmować zagadnienia związane z komputerowym wspomaganiem decyzji, jednak w przypadku rzeczywistych danych, okazują się

być często nieskuteczne. To powoduje, iż ciągle poszukiwane są nowe próby konstrukcji systemów decyzyjnych, a zagadnienie ilościowego mierzenia niepewności przy prowadzeniu wnioskowania aproksymacyjnych ciągle wzbudza wiele zainteresowania.

2 Zbiory przybliżone

Teoria *prof. Pawlaka* dotycząca zbiorów przybliżonych dostarcza narzędzi matematycznych do formalnego opisu wiedzy, w tym wiedzy niepełnej i niedokładnej. Wykorzystywana skutecznie m.in. w eksploracji danych i odkrywaniu wiedzy, złożonych zadaniach klasyfikacji oraz w komputerowych systemach wspomagania decyzji. Dziedziny, w których teoria ta została zastosowana to nie tylko medycyna czy biznes (bankowość, badania rynku) ale również rozpoznawanie mowy, sieci neuronowe czy ogólnie mówiąc sztuczna inteligencja.

Metodologia zbiorów przybliżonych zyskała sobie dużą popularność. Świadectwem tego może być chociażby fakt, że jest ona przedmiotem badań wielu osób na całym świecie, co udokumentowane zostało ok dwoma tysiącami publikacji. Tematyka ta cieszy się ogromnym zainteresowaniem badaczy, którzy czynnie uczestniczą w prowadzonych w tej dziedzinie cyklicznie międzynarodowych konferencjach i seminariach. Wśród krajów szczególnie zaangażowanych wyróżnić można prócz Polski, m.in. USA, Kanadę, Japonię, Francję czy Norwegię. W Polsce tematyka ta znalazła się w centrum badań naukowych prowadzonych w wielu ośrodkach, również na Uniwersytecie Śląskim w Zakładzie Systemów Informatycznych, szczególnie w zakresie zastosowań medycznych. Na szczególną uwagę zasługują tutaj prace: [4],[5] oraz [6].

2.1 System informacyjny

Istnieje szereg struktur, które mogą być wykorzystane do przechowywania danych. Niezależnie jednak od tego jaki ostatecznie rodzaj reprezentacji danych przyjmiemy, ważne i zarazem niezbędne wydaje się być spełnienie przez tę formę cech *uniwersalności* oraz *efektywności*. Uniwersalność bowiem zgodnie z założeniem pozwala na gromadzenie i przechowywanie zbiorów różnorodnych danych, opisujących badane zjawiska i procesy zaś efektywność umożliwiać w łatwy sposób komputerową analizę tak zapisanych danych. Te dwie równie ważne cechy posiada znany i często wykorzystywany w praktyce tablicowy sposób reprezentacji danych, który najczęściej przedstawia się za pomocą tablicy, w której kolumny są etykietowane przez atrybuty (parametry, własności, cechy), wiersze odpowiadają zaś obiektom (elementom, sytuacjom, stanom), a na przecięciu wierszy i kolumn znajdują się wartości odpowiednich atrybutów dla poszczególnych obiektów. Mowa oczywiście o strukturze określanej najczęściej mianem systemu informacyjnego (SI) (ang. *information system*) [1]. Formalnie, zbiór ten definiowany jest jako uporządkowana

czwórka:

$$SI =< U, A, V, f >$$

gdzie:

- U - jest niepustym, skończonym zbiorem zwanym uniwersum, przy czym elementy zbioru, U nazywamy obiektami $U = \{x_1, x_2, .., x_n\}$,
- A - jest niepustym, skończonym zbiorem atrybutów: $A = \{a_1, a_2, .., a_m\}$,
- V - jest zbiorem wartości atrybutów ze zbioru A: $V = \cup_{a \in A} V_a$, przy czym V_a nazywamy dziedziną atrybutu $a \in A$,
- f - jest funkcją informacji, odpowiadającą iloczynowi kartezjańskiemu zbioru obiektów i zbioru atrybutów w zbiór wartości atrybutów, co odpowiada formule: $U \times A \to V$, gdzie $\forall_{\substack{x \in U \\ a \in A}} f(x, a) \in V_a$.

Załóżmy, że rozważany przez nas system informacyjny przedstawia tabela nr 1. Obiektami w systemie są klienci banku starający się o kredyt. gdzie: a

Klient	(a)	(b)	(c)	(dec)
1	nie	tak	wysoka	tak
2	tak	nie	wysoka	tak
3	tak	tak	bardzo wysoka	tak
4	nie	tak	bardzo wysoka	tak
5	tak	nie	wysoka	nie
6	nie	tak	normalna	nie

Tablica1. System informacyjny

- płynność finansowa, b - brak kredytów, c - pozycja na rynku oraz dec - decyzja kredytowa. System ten składa się z sześciu obiektów $(1, 2, .., 6)$ oraz czterech atrybutów (płynność finansowa, brak kredytów, sukcesy na rynku, decyzja kredytowa).
Do reprezentacji tego systemu można wykorzystać proponowaną formę czwórki: $SI =< U, A, V, f >$ gdzie:

- $U = \{1, 2, 3, 4, 5, 6\}$,
- $A = \{$*płynność finansowa, brak kredytów, pozycja na rynku, decyzja kredytowa*$\}$
- $V = V_{plynnoscfinansowa} \cup V_{Brakkredytow} \cup V_{pozycjanarynku} \cup V_{decyzjakredytowa}$
  $V_{plynnoscfinansowa} = \{nie, tak\}$
  $V_{brakkredytow} = \{nie, tak\}$
  $V_{pozycjanarynku} = \{normalna, wysoka, bardzowysoka\}$
  $V_{decyzjakredytowa} = \{nie, tak\}$
- $f : U \times A \to V$:
  f(1, płynność finansowa)=nie
  f(3, decyzja kredytowa) = tak

2.2 Tablica decyzyjna

Szczególnym rodzajem systemów informacyjnych są tablice decyzyjne (TD). Tablicą decyzyjną nazywamy uporządkowaną piątkę:

$$TD = (U, C, D, V, f)$$

gdzie:

- $C, D \subset A; C \neq \emptyset; C \cup D = A; C \cap D = \emptyset$,
- elementy zbioru C nazywamy atrybutami warunkowymi,
- elementy zbioru D nazywamy atrybutami decyzyjnymi,
- f nazywamy funkcją decyzyjną.
- interpretacja U oraz V jest taka sama jak w przypadku systemu informacyjnego, ponadto poszczególne wartości v dziedzin atrybutów $D(v \in V_D)$ będziemy nazywać **klasami decyzyjnymi**.

Podstawowa różnica między tablicą decyzyjną a systemem informacyjnym polega więc na tym, że część atrybutów traktujemy jako atrybuty warunkowe (C) a część jako decyzyjne (D).

Przykład
Tabelę 1 będziemy (i nawet powinniśmy jeśli rozpatrzymy przeznaczenie owego systemu i rodzaj informacji w nim przechowywanych) traktować jako tablicę decyzyjną.
Zbiór atrybutów systemu informacyjnego dzielimy więc na dwa podzbiory: podzbiór atrybutów warunkowych (C) oraz podzbiór atrybutów decyzyjnych (D) w następujący sposób:

- $C = \{$*płynność finansowa, brak kredytów, sukcesy na rynku*$\} = \{a, b, c\}$
- $D = \{$*decyzja kredytowa*$\} = \{dec\}$

Z dziedziną systemów informacyjnych nierozerwalnie wiąże się pojęcie relacji nierozróżnialności.

2.3 Relacja nierozróżnialności

Relacja nierozróżnialności, występująca w teorii zbiorów przybliżonych jest generowana przez informację opisującą obiekty należące do zbioru uniwersum U. Pomaga analizować różnego rodzaju systemy informacyjne, zwłaszcza w problemach automatycznego pozyskiwania wiedzy czy generowania reguł decyzyjnych na podstawie danych uczących. Analizując poszczególne obiekty z tabeli nr 1, można zaobserwować, że obiekty o numerach $1, 4$ i 6 mają te same wartości atrybutów: *płynność finansowa* oraz *brak kredytów* zaś obiekty o numerach 1 i 5 mają tę samą wartość atrybutu *pozycja na rynku*. O obiektach numer $1, 4$ i 6 powiemy, że są **nierozróżnialne ze względu na atrybuty**:

płynność finansowa oraz *brak kredytów*, zaś obiekty o numerach 1 i 5 są **nierozróżnialne ze względu na atrybut**: *pozycja na runku*.
Tę obserwację można uogólnić i wyrazić w sposób formalny stosując odpowiednio zdefiniowaną relację.
Niech $SI =< U, A, V, f >$ będzie systemem informacyjnym i niech $A \subseteq B$.

Relację nierozróżnialności (ang. *indiscernibility relation*) na zbiorze obiektów U generowaną przez zbiór atrybutów B określamy jako:

$$IND_{SI}(B) = \{(x, y) \in U \times U : \forall a \in B f(x, a) = f(y, a)\}$$

gdzie:
$(x, y) \in U \times U$ to obiekty należące do iloczynu kartezjańskiego zbioru Uniwersum, $a \in B$ to każdy atrybut ze zbioru B, $f(x, a)$ to funkcja informacji dla pary: obiekt x i atrybut a [2], [3].

Poszczególne pary obiektów należą do relacji wtedy, gdy posiadają te same wartości dla wszystkich atrybutów ze zbioru B.
Relacja nierozróżnialności $IND_{SI}(B)$ jest relacją równoważności, gdyż jest relacją:

- zwrotną, gdyż: $\overset{\forall}{_{u \in U}}(u, u) \in IND_{SI}(B)$
- symetryczną, gdyż: $\overset{\forall}{_{u,v \in U}}((u, v) \in IND_{SI}(B) \Rightarrow (v, u) \in IND_{SI}(B))$
- przechodnią, gdyż: $\overset{\forall}{_{u,v,w \in U}}((u, v) \in IND_{SI}(B) \bigwedge (v, w) \in IND_{SI}(B) \Rightarrow (u, w) \in IND_{SI}(B))$

Każda relacja równoważności dzieli zbiór, w którym jest określona, na rodzinę rozłącznych podzbiorów zwanych klasami abstrakcji (równoważności) lub zbiorami elementarnymi tej relacji. Klasa abstrakcji elementu $y \in X$ względem relacji równoważności R w zbiorze X to zbiór elementów $x \in X$, które są w relacji R z y.

2.4 Zbiór dokładny oraz zbiór przybliżony

Operowanie pojęciami nieostrymi (nieścisłymi, nieprecyzyjnymi) jest bez wątpienia jednym z głównych problemów rozumowań potocznych. Pojęcia nieostre różnią się tym od pojęć ostrych, że w przeciwieństwie do tych ostatnich nie zawsze możliwe jest jednoznaczne zaklasyfikowanie obiektu do pojęcia, tzn. dla pewnej grupy obiektów z otaczającej nas rzeczywistości nie można — stwierdzić jednoznacznie czy dany obiekt należy do rozpatrywanego pojęcia, czy też nie należy. W rozpatrywanym przykładzie klientów banku trudno będzie określić sytuację klienta na rynku, w przypadku, gdy ten klient w różnych okresach wykazał zarówno duże zyski jak i spore straty. Znane są różne modele reprezentacji pojęć nieostrych, tj. model matematyczny zwany teorią zbiorów rozmytych zaproponowany przez *Zadeha*, czy chociażby teoria ewidencji, zwana również teorią *DemsteraShafera*.

Teoria zbiorów przybliżonych także może być uważana za jeden ze sposobów formalizacji nieostrości pojęć. Wiele prac rozpatrywało na wielu różnych obszarach związki istniejące pomiędzy teorią zbiorów rozmytych a teorią zbiorów przybliżonych. Teoria zbiorów przybliżonych proponuje zastąpienie nieostrego (nieprecyzyjnego) pojęcia, parą pojęć precyzyjnych, zwanych *dolnym* i *górnym przybliżeniem* tego pojęcia. Różnica między górnym i dolnym przybliżeniem jest właśnie tym obszarem granicznym, do którego należą wszystkie przypadki, które nie mogą być prawidłowo zaklasyfikowane na podstawie aktualnej wiedzy. Im większy obszar graniczny pojęcia tym bardziej jest ono nieostre (nieprecyzyjne).

2.5 Aproksymacja zbioru

Jeśli $SI =< U, A, V, f >$ jest systemem informacyjnym takim, że $B \subseteq A$ oraz $X \subseteq U$ to:

- $\underline{B}$ – dolnym przybliżeniem (aproksymacją) zbioru X w systemie informacyjnym SI nazywamy zbiór: $\underline{B}X = \{x \in U : I_{SI,B}(x) \subseteq X\}$
- $\overline{B}$ – górnym przybliżeniem (aproksymacją) zbioru X w systemie informacyjnym SI nazywamy zbiór: $\overline{B}X = \{x \in U : I_{SI,B}(x) \cap X \neq \emptyset\}$
- B – pozytywnym obszarem (ang. *positive area*) zbioru X w systemie informacyjnym SI nazywamy zbiór: $POS_B(X) = \underline{B}X$
- B – brzegiem (granicą) (ang. *boundary*) zbioru X w systemie informacyjnym SI nazywamy zbiór: $BN_B(X) = \overline{B}X - \underline{B}X$
- B – negatywnym obszarem (ang. *negative area*) zbioru X w systemie informacyjnym SI nazywamy zbiór: $NEG_B X = U - \overline{B}X$

Definicje te sprowadzają do następujących wniosków:

1. $\underline{B}X \subseteq X \subseteq \overline{B}X$,
2. zbiór X jest B-dokładny, gdy: $\underline{B}X = \overline{B}X \Leftrightarrow BN_B X = \emptyset$,
3. zbiór X jest B-przybliżony, gdy: $\underline{B}X = \overline{B}X \Leftrightarrow BN_B X \neq \emptyset$.

Dolne przybliżenie pojęcia jest to więc obszar definiujący wszystkie obiekty, co do których nie ma wątpliwości, że są one reprezentantami tego pojęcia w świetle posiadanej wiedzy. Do **górnego przybliżenia** należą obiekty, których nie można wykluczyć, że są reprezentantami tego pojęcia. **Brzegiem** zaś pojęcia są wszystkie te obiekty, co do których nie wiadomo czy są czy nie reprezentantami danego zbioru. Istnieje także tzw. liczbowa charakterystyka aproksymacji zbioru, która za pomocą *współczynnika dokładności aproksymacji (przybliżenia)* pozwala ilościowo charakteryzować nam nieostrość pojęć. Współczynnik ten dla zbioru X w systemie informacyjnym SI względem zbioru atrybutów B wyraża się wzorem:

$$\alpha_B(X) = \frac{card(POS_B(X))}{card(\overline{B}X)} = \frac{card(\underline{B}X)}{card(\overline{B}X)}$$

gdzie $card(X)$ oznacza liczność zbioru X.
Łatwo zauważyć, że:

1. $0 \leqslant \alpha_B(X) \leqslant 1$,
2. jeżeli X jest zbiorem dokładnym to: $\alpha_B(X) = 1$,
3. jeżeli X jest zbiorem przybliżonym to: $0 \leqslant \alpha_B(X) < 1$.

3 Usuwanie niespójności z tablicy decyzyjnej

Niespójność w wiedzy to zjawisko dość powszechne wszędzie tam gdzie stosujemy systemy decyzyjne, w systemach z dziedzinowymi bazami wiedzy. Niespójność w tablicy decyzyjnej uniemożliwia prawidłowe podejmowanie decyzji dla tych przypadków, które dotyczą obiektów niespójnych. Zbiory przybliżone dostarczają jednak metod pozwalających na usunięcie obiektów powodujących tę niespójność. Najbardziej znane to metoda nowego podziału systemu informacyjnego oraz metoda jakościowa posługująca się pojęciami dolnego i górnego przybliżenia zbioru. Metoda jakościowa mówi, że z systemu o niespójnej wiedzy należy usunąć ten obiekt bądź obiekty, dla których dokładność dolnego bądź górnego przybliżenia była mniejsza. Dla każdego $X \subseteq U$ i $B \subseteq A$ dokładność dolnego przybliżenia $\gamma_B(X)$ obliczymy ze wzoru:

$$\gamma_B(X) = \frac{|\underline{B}X|}{|U|}$$

Dokładność górnego przybliżenia $\gamma^B(X)$ obliczymy ze wzoru:

$$\gamma_B(X) = \frac{|\overline{B}X|}{|U|}$$

gdzie: $|\underline{B}X|$ to moc zbioru dolnego przybliżenia zbioru X, $|\overline{B}X|$ odpowiednio oznacza moc zbioru górnego przybliżenia zbioru X, zaś $|U|$ to z kolei moc zbioru uniwersum U.

Przykład Dla tabeli numer 1, która przecież jest niespójna postaramy się usunąć niespójność metodą jakościową. Sprawdzimy w tym celu relację między klasami decyzyjnymi a klasami równoważności w systemie dla zbioru atrybutów warunkowych. Zbiór obiektów X dzielimy ze względu na decyzję na dwa rozłączne podzbiory X_1 oraz X_2, gdyż atrybut decyzyjny: *decyzja kredytowa* klasyfikuje nam obiekty na dwie grupy: *decyzja kredytowa = tak* oraz *decyzja kredytowa = nie*.
$X_1 = \{1, 2, 3, 4\}$
$X_2 = \{5, 6\}$
Następnie wyznaczamy klasy równoważności dla całego zbioru atrybutów warunkowych:
$IND(C) = \{\{1\}, \{2, 5\}, \{3\}, \{4\}, \{5\}, \{6\}\}$.
Teraz można już wyznaczyć dla każdego ze zbiorów klas decyzyjnych: X_1 oraz X_2 dolne oraz górne przybliżenie.
$\underline{B}X_1 = \{1, 3, 4\}$
$\overline{B}X_1 = \{1, 2, 3, 4, 5\}$

$\underline{\underline{B}}X_2 = \{6\}$
$\overline{B}X_2 = \{2, 5, 6\}$
Wyliczamy dokładności górnego oraz dolnego przybliżenia:
$\gamma_B(X_1) = \frac{|\underline{\underline{B}}X_1|}{|U|} = \frac{3}{6} = \frac{1}{2}$
$\gamma_B(X_2) = \frac{|\underline{\underline{B}}X_2|}{|U|} = \frac{1}{6}$
$\gamma_B(X_1) = \frac{|\overline{B}X_1|}{|U|} = \frac{5}{6}$
$\gamma_B(X_2) = \frac{|\underline{\underline{B}}X_2|}{|U|} = \frac{3}{6} = \frac{1}{2}$
Zgodnie z założeniami metody usuniemy obiekt, który powodował niespójność i występował w zbiorze X_2. Spójna już teraz tablica decyzyjna wygląda następująco:
gdzie: a - płynność finansowa, b - brak kredytów, c - pozycja na rynku oraz

Klient	(a)	(b)	(c)	(dec)
1	nie	tak	wysoka	tak
2	tak	nie	wysoka	tak
3	tak	tak	bardzo wysoka	tak
4	nie	tak	bardzo wysoka	tak
6	nie	tak	normalna	nie

Tablica2. System informacyjny / tablica decyzyjna po usunięciu niespójności

dec - decyzja kredytowa. Metoda tworzenia nowego podziału (Systemu informacyjnego) proponuje, aby bez usuwania czegokolwiek z danego systemu informacyjnego pozbyć się niespójności w wiedzy. Wiedząc, że decyzja d wyznacza klasyfikację: $Class_A(d) = \{X_1, ..., X_{r(d)}\}$, (gdzie (d) - to ilość różnych wartości atrybutu decyzyjnego), tworzymy nowy podział: $App-Class_A(d) = \{A_|X1, ..., A_{Xr(d)}\} \bigcup \{Bd_A() : || > 1\}$ gdzie ten nowy podział tworzy tablicę decyzyjną spójną.
Tabela nr 1 (niespójna) po dodaniu do systemu informacyjnego, nowego, uogólnionego atrybutu decyzyjnego wygląda następująco:
gdzie: a - płynność finansowa, b - brak kredytów, c - pozycja na rynku

Klient	(a)	(b)	(c)	(dec)
1	nie	tak	wysoka	tak
2	tak	nie	wysoka	tak,nie
3	tak	tak	bardzo wysoka	tak
4	nie	tak	bardzo wysoka	tak
5	tak	nie	wysoka	tak,nie
6	nie	tak	normalna	nie

Tablica3. System informacyjny / tablica decyzyjna z uogólnionym atrybutem decyzyjnym

oraz *dec* - decyzja kredytowa. Rozwiązanie to jednak w wielu przypadkach nie jest efektywne. Dzieje się tak dlatego, iż najczęściej systemy informacyjne mające charakter decyzyjny klasyfikują obiekty do jednej z dwóch grup decyzyjnych. Podobnie jest w rozpatrywanym w niniejszej pracy przykładzie systemu, który dla każdego obiektu proponuje jedną z dwóch wartości atrybutu decyzyjnego *decyzja kredytowa*: *tak* lub *nie*. W tym przypadku omawiana metoda dodaje do systemu nową wartość będącą połączeniem dwóch pozostałych wartości: tak, nie. Nie trzeba chyba uzasadniać bezcelowości takiego postępowania, które wyraźnie nie rozwiązuje problemu.

4 Macierz oraz funkcja rozróżnialności systemu informacyjnego

Zbiory przybliżone pozwalają nie tylko w prosty sposób przechowywać informacje o obiektach tworzących ten system. Przede wszystkim teoria ta wspomaga procesy pozyskiwania wiedzy dzięki metodom redukcji wiedzy oraz generowania reguł decyzyjnych (w tym reguł minimalnych). Te elementy wsparte metodami usuwania niespójności z systemu stanowią silny samowystarczalny aparat matematyczny kierujący działaniem systemy ekspertowego. Dzięki redukcji wiedzy udaje nam się zminimalizować liczbę informacji przechowywanych w systemie, co z kolei przyspiesza procesy podejmowania decyzji dzięki zminimalizowaniu danych określających zdefiniowane wcześniej przypadki.

4.1 Macierz rozróżnialności

Jeśli $SI =< U, A, V, f >$ jest systemem informacyjnym takim, że $U = \{u_1, u_2, .., u_n\}$ i $A = \{a_1, a_2, .., a_m\}$, to *macierz rozróżnialności* systemu informacyjnego SI $M(SI)$ (ang. *indiscernibility matrix*) definiujemy następująco:

$$M(SI) = (H_{i,j})_{i,j=1,..,n} = \{a \in A : f(u_i, a) \neq f(u_j, a)\}$$

dla $i, j = 1, .., n$, gdzie $n = |U|$.

Macierz ta jest dwuwymiarową macierzą kwadratową o wymiarach: $|U| \times |U|$. Pojedyncza komórka będąca przecięciem i-tego wiersza z j-tą kolumną, co oznaczamy: $M(SI)[i, j]$, zawiera zbiór tych atrybutów, dla których obiekty uniwersum u_i i u_j mają różne wartości (są rozróżnialne przy pomocy tych atrybutów). Macierz ta dla rozpatrywanego systemu wygląda następująco:

Własności macierzy rozróżnialności:

- macierz $M(SI)$ ma zawsze na przekątnej zbiory puste ($\emptyset$),
- macierz $M(SI)$ jest symetryczna względem przekątnej,
- każdy element macierzy $M(SI)$ jest zbiorem,
- rozmiar macierzy rośnie w sposób kwadratowy wraz ze wzrostem liczby obiektów w systemie informacyjnym.

	1	2	3	4	6
1	$\emptyset$				
2	a,b	$\emptyset$			
3	a,c	b,c	$\emptyset$		
4	c	a,b,c	a	$\emptyset$	
6	c	a,b,c	a,c	c	$\emptyset$

Tablica4. Macierz rozróżnialności dla systemu informacyjnego

Generowanie macierzy rozróżnialności

Wejście: $A = (U, A)$ system informacyjny taki, że $U = \{u_1, .., u_n\}$ i $A = \{a_1, .., a_m\}$.
Wyjście: $M(A) = (C_{ij})_{i,j=1,..,n}$ macierz rozróżnialności systemu A, przy czym $M(A)$ ma obliczone tylko te pola C_{ij} dla których $1 \leqslant j < i \leqslant n$.
Metoda:

For i=1 to n do
For j=1 to i-1 do
Wstaw do C_{ij} atrybuty, na których różnią się obiekty u_i i u_j.

Złożoność. Złożoność pamięciowa wynika z elementów tworzących samą macierz $M(A)$, gdzie należy wyznaczyć zawartość $\frac{n^2-n}{2}$ pól. Złożoność obliczeniowa wyznaczania każdego pola jest zależna od liczby atrybutów m opisujących obiekty. Dlatego złożoność obliczeniowa czasowa algorytmu jest rzędu $O(n^2 * m)$, natomiast złożoność obliczeniowa pamięciowa algorytmu jest rzędu $O(C)$, gdzie C jest pewną stałą.

4.2 Funkcja rozróżnialności

Wiedzę zawartą w macierzy rozróżnialności można także przedstawić w postaci funkcji rozróżnialności. Funkcją rozróżnialności systemu informacyjnego SI (ang. *discernibility function*) nazywamy funkcję boolowską f_{SI} zmiennych $a_1^*, .., a_m^*$ odpowiadających odpowiednio atrybutom (systemu informacyjnego) $a_1, .., a_m$ zdefiniowaną następująco:
$f_{SI}(a_1^*, .., a_m^*) = \bigcap\{\bigcup(X_{i,j} : 1 \leqslant j \leqslant n \wedge Hi, j \neq \emptyset)\}$
gdzie: $n = |U|, m = |A|, \bigcup X_{i,j}$ jest alternatywą wszystkich zmiennych $a^* \in \{a_1^*, .., a_m^*$ takich, że $a \in Hi, j$.
Funkcja rozróżnialności staje się niezwykle przydatna w procesie generowania reguł minimalnych. Dzięki minimalizacji funkcji boolowskiej, jaką niewątpliwie omawiana funkcja jest uzyskuje się również optymalną postać reguły decyzyjnej, którą nazywa się regułą minimalną.
Przykład
Obliczmy funkcję rozróżnialności dla macierzy rozróżnialności z tabeli 4:

$f_{SI}(a^*, b^*, c^*, dec^*) = (a^* \vee b^*) \wedge (a^* \vee c^*) \wedge (c^*) \wedge (a^* \vee b^* \vee c^*) \wedge (c^* \vee c^*) \wedge (b^* \vee c^*) \wedge (a^* \vee b^* \vee c^*) \wedge (c^*) \wedge (a^* \vee b^* \vee c^* \vee c^*) \wedge (a^*) \wedge (b^* \vee c^* \vee c^*) \wedge (a^* \vee c^* \vee c^*) \wedge (a^* \vee b^* \vee c^* \vee c^*) \wedge (c^* \vee c^*) \wedge (a^* \vee b^* \vee c^*)$

Wyrażenie to można uprościć stosując m.in. prawo pochłaniania $(a \vee (a \cup b)) = a$ do postaci:
$f_{SI}(a^*, b^*, c^*, dec^*) = (c^* \wedge a^* \wedge dec^*)$

5 Redukt i rdzeń zbioru atrybutów

Niech $SI = \{U, A, V, f\}$ będzie systemem informacyjnym oraz $C \subseteq A$.

Definicja. Atrybut zbędny (zależny)
Atrybut $a \subseteq C$ jest zbędny, jeżeli $IND(C) = IND(C - \{a\})$.
W przeciwnym wypadku (tzn. jeżeli $IND(C) \neq IND(C - \{a\})$ jest niezbędny.

Definicja. Atrybut niezależny (niezbędny)
A - zbiór atrybutów jest niezależny wtedy i tylko wtedy, gdy dla każdego $a \subseteq A$, a jest niezbędny. W przeciwnym wypadku zbiór jest zależny.

Definicja. Redukt i rdzeń (jądro) $B \subseteq A$ nazywamy **reduktem** A wtedy i tylko wtedy, gdy B jest niezależny oraz $IND(B) = IND(A)$. Zbiór wszystkich reduktów oznaczamy przez $RED(A)$.
Zbiór wszystkich niezbędnych atrybutów w B będziemy nazywali **rdzeniem** (jądrem) B i oznaczali przez $CORE(B)$.
Powiązanie między reduktami i jądrem
Zachodzi następujący związek:
$CORE(A) = \bigcap RED(A)$,
gdzie $RED(A)$ to zbiór wszystkich reduktów B, tzn. jądro atrybutów to przekrój po wszystkich reduktach.

5.1 Generowanie reduktu i rdzenia z definicji

Redukujemy zbiór atrybutów w systemie do takiego podzbioru, który zapewnia nam dotychczasową klasyfikację obiektów. Najpierw wyznaczamy klasy równoważności dla pełnego zbioru atrybutów:
$IND(C) = \{\{1\}, \{2\}, \{3\}, \{4\}, \{6\}\}$
Teraz będziemy sprawdzać czy zmieni się dotychczasowa klasyfikacja obiektów, jaką mamy dla pełnego zbioru atrybutów, jeśli usuniemy jakiś atrybut ze zbioru C.
$IND((C) - \{a\}) = \{\{1\}, \{2\}, \{3, 4\}, \{6\}\}$
czyli:
$IND((C) - \{a\}) \neq IND(C)$
więc atrybut $\{a\}$ jest niezbędny w systemie, ponieważ jeśli go usuniemy to stracimy informacje o rozróżnialności dwóch obiektów 3 i 4.

$IND((C) - \{b\}) = \{\{1\}, \{2\}, \{3\}, \{4\}, \{6\}\}$
czyli:
$IND((C) - \{b\}) = IND(C)$
więc atrybut $\{b\}$ jest zbędny w systemie, ponieważ jeśli go usuniemy to nie stracimy informacji o rozróżnialności obiektów.
$IND((C) - \{c\}) = \{\{1, 4, 6\}, \{2\}, \{3\}\}$
czyli:
$IND((C) - \{c\}) \neq IND(C)$
więc atrybut $\{c\}$ jest niezbędny w systemie, ponieważ jeśli go usuniemy to stracimy informację o rozróżnialności obiektów.

Zatem $CORE(C)$ to zbiór atrybutów niezbędnych w systemie więc w naszym przypadku stanowią go dwa atrybuty:
$CORE(C) = \{ac\}$
Redukt zgodnie z definicją jest to taki zbiór atrybutów niezbędnych, dla którego zapewniona jest dotychczasowa klasyfikacja obiektów, a wiec na pewno redukt musi zawierać w sobie jądro.
Sprawdzamy więc dla jakiej kombinacji atrybutów uzyskamy taki sam podział obiektów jaki dała $IND(C)$.
$IND(ac) = \{\{1\}, \{2\}, \{3\}, \{4\}, \{6\}\}$
Skoro $IND(ac) = IND(C)$, to ten zbiór atrybutów $\{ac\}$ jest reduktem zbioru atrybutów.
$RED(C) = \{ac\}$.

5.2 Generowanie reduktu i rdzenia z macierzy rozróżnialności

Mając wygenerowaną macierz rozróżnialności dla tablicy decyzyjnej i wiedząc, że każda komórka $M(SI)[i, j]$ zawiera zbiór tych atrybutów, dla których obiekty uniwersum u_i i u_j mają różne wartości (są rozróżnialne przy pomocy tych atrybutów) możemy wygenerować redukt oraz rdzeń takiego systemu.

Rdzeniem będzie połączenie wszystkich tych komórek macierzy, które zawierały zbiory jednostkowe: $CORE(A) = \{a \subseteq A : c_{ij} = \{a\}\}$, dla pewnego $0 < i, j < n + 1$, tzn. takie, które występują w macierzy rozróżnialności pojedynczo.
$B \subseteq A$ jest reduktem A wtedy i tylko wtedy, gdy B jest minimalny (w sensie zawierania zbiorów) oraz z każdym niepustym elementem macierzy rozróżnialności $M(S)$ ma niepuste przecięcie. Innymi słowy redukt jest to najmniejszy zbiór atrybutów, przy którym zostaje zachowana dotychczasowa klasyfikacja (rozróżnialność) obiektów. $RED(C) = \{ac\}$ oraz $CORE(C) = \{ac\}$.

6 Generowanie reguł decyzyjnych

Każdy obiekt $u \subset U$ tablicy decyzyjnej $TD = (U, C, D, V, f)$ może zostać zapisany w postaci zdania warunkowego (postaci: *jeżeli warunki to decyzja*)

i być traktowany jako **reguła decyzyjna**.

Regułą decyzyjną w tablicy decyzyjnej TD nazywamy funkcje:$q : C \cup D \rightarrow V$ jeżeli istnieje $x \in U$, taki, że $q = f_x$.
Obcięcie g do C $(q|C)$ oraz q do D $(q|D)$ nazywamy odpowiednio warunkami oraz decyzjami reguły decyzyjnej q.

Przykład
Z przykładowej tablicy decyzyjnej z tabeli 1 po uprzednim usunięciu niespójności, możemy wyprowadzić następujące reguły (odpowiadające konkretnym obiektom):

1. *jeżeli* (a="nie") *i* (b="tak") *i* (c="wysoka") *to* (dec="tak")
2. *jeżeli* (a="tak") *i* (b="nie") *i* (c="wysoka") *to* (dec="tak")
3. *jeżeli* (a="tak") *i* (b="tak") *i* (c="bardzo wysoka") *to* (dec="tak")
4. *jeżeli* (a="nie") *i* (b="tak") *i* (c="bardzo wysoka") *to* (dec="tak")
5. *jeżeli* (a="nie") *i* (b="tak") *i* (c="normalna") *to* (dec="nie")

6.1 Reguły minimalne

Tworzymy reguły minimalne dla $\delta_{decyzjakredytowa} = \{tak\}$ czyli reguły postaci: $\alpha \Rightarrow \delta_{decyzjakredytowa} = \{tak\}$

Aby stworzyć te reguły musimy utworzyć uogólnione macierze rozróżnialności dla obiektów zbioru X (mających wartość decyzji {tak}):
$MG(A, \{tak\}, X_1)$,
$MG(A, \{tak\}, X_2)$,
$MG(A, \{tak\}, X_3)$,
$MG(A, \{tak\}, X_4)$.
Funkcja rozróżnialności odpowiadająca tej macierzy ma postać:
$f_{MG}(A, \{tak\}, X_1)(a, b, c) = c$
$f_{MG}(A, \{tak\}, X_2)(a, b, c) = a \vee b \vee c$
$f_{MG}(A, \{tak\}, X_3)(a, b, c) = a \vee c$
$f_{MG}(A, \{tak\}, X_4)(a, b, c) = c$

Tworzymy reguły minimalne dla $\delta_{decyzjakredytowa} = \{nie\}$ czyli reguły postaci:
$\alpha \Rightarrow \delta_{decyzjakredytowa} = \{nie\}$
Aby stworzyć te reguły musimy utworzyć uogólnione macierze rozróżnialności dla obiektów zbioru X (mających wartość decyzji {tak}):
$MG(A, \{nie\}, X_6)$.
Funkcja rozróżnialności odpowiadająca tej macierzy ma postać:
$f_{MG}(A, \{nie\}, X_6)(a, b, c) = (c) \wedge (a \vee b \vee c) \wedge (a \vee c) \wedge (c)$

Korzystając z praw algebry Boole'a oraz zastąpując symbol $\wedge$ symbolem $*$, a symbol $\vee$ symbolem $+$, otrzymujemy: $f_{ba}(A, \{nie\}, X_6)(a, b, c) = (c) * (a + b + c) * (a + c) * (c) = cc * (a + b + c) * (a + c) = (cca + ccb + ccc)(a + c) =$

$(ccaa + ccca + ccba + cccb + ccca + cccc) = (ca + ca + cba + cb + ca + c) = (ca + cba + cb + c) = ca(1 + b) + c(1 + b) = ca + c = c(a + 1) = c$
Czyli:

- funkcja dla reguły nr 1: $[c]$
- funkcja dla reguły nr 2: $[a + b + c]$
- funkcja dla reguły nr 3: $[a + c]$
- funkcja dla reguły nr 4: $[c]$
- funkcja dla reguły nr 6: $[c]$

Funkcja $f_{MG}(A, \{tak\}, X_1) = c$ oznacza, że możemy zbudować dla decyzji $\{tak\}$ 1 regułę minimalną:
$if\ c = wysoka\ then\ decyzjakredytowa = tak$

Odpowiednio teraz:
Funkcja $f_{MG}(A, \{tak\}, X_2) = a + b + c$ oznacza, że możemy zbudować dla decyzji $\{tak\}$ 3 reguły minimalne:

$if\ a = tak\ then\ decyzjakredytowa = tak$
$if\ b = nie\ then\ decyzjakredytowa = tak$
$if\ c = wysoka\ then\ decyzjakredytowa = tak$

Funkcja $f_{MG}(A, \{tak\}, X_3) = a + c$ oznacza, że możemy zbudować dla decyzji $\{tak\}$ 2 reguły minimalne:

$if\ a = tak\ then\ decyzjakredytowa = tak$
$if\ c = bardzowysoka\ then\ decyzjakredytowa = tak$

Funkcja $f_{MG}(A, \{tak\}, X_4) = c$ oznacza, że możemy zbudować dla decyzji $\{tak\}$ 1 regułę minimalną:

$if\ c = bardzowysoka\ then\ decyzjakredytowa = tak$
Ostatecznie otrzymamy optymalną regułę decyzyjną dla decyzji $dec = tak$:
$if\ c = wysoka \vee a = tak \vee b = nie \vee c = bardzowysoka\ then\ decyzjakredytowa = tak$

Funkcja $f_{MG}(A, \{nie\}, X_6) = c$ oznacza, że możemy zbudować dla decyzji $\{nie\}$ 1 regułę minimalną:

$if\ c = normalna\ then\ decyzjakredytowa = nie$

7 Posumowanie

Teoria zbiorów przybliżonych (ang. *rough sets*) została zaproponowana przez Pawlaka jako narzędzie analizy informacji granularnej. Granularność dostępnej informacji może powodować niespójność opisu obiektów, a teoria zbiorów

przybliżonych dostarcza właśnie podstaw do uwzględniania tego typu niespójności. Teoria oparta jest na założeniu, że posiadając informację reprezentowaną za pomocą atrybutów i ich wartości na obiektach, możliwe jest określenie relacji zachodzącej pomiędzy tymi obiektami. Obiekty posiadające ten sam opis, wyrażony za pomocą atrybutów, są nierozróżnialne ze względu na dostępną informację. Zakładamy, że informacja o obiektach dostępna jest w postaci tablicy informacyjnej W przypadku uczenia nadzorowanego informacja o klasyfikacji obiektów może być wyrażona za pomocą atrybutu decyzyjnego d. Prowadzi to do reprezentacji w postaci tablicy decyzyjnej $DT = (U, A)$. Teoria zbiorów przybliżonych jest wykorzystywana w różnym stopniu zarówno w indukcji reguł, jak i we wstępnym przetwarzaniu danych. W rezultacie zaproponowano wiele algorytmów indukcji reguł korzystających z elementów zbiorów przybliżonych. Zagadnienia przedstawione w niniejszej pracy pokazują użyteczne funkcje teorii *prof Pawlaka*, które pozwalają nie tylko gromadzić ale i skutecznie analizować wiedzę zapisaną w reprezentacji dostarczanej przez zbiory przybliżone.

Literatura

1. Pawlak Z., (1983) Information Systems - theoretical foundations [polish], WNT, W-wa.
2. Pawlak Z., (1982) Rough Sets. Int. J. of Information and Computer Sci 11: 344-356.
3. Skowron A., Grzymała-Busse J., (1994) From the Rough Set Theory to the Evidence Theory. In R.R. Yager, M. Fedrizzi, J. Kacprzyk (eds.), Advances in the Dempster-Shafer Theory of Evidence. New York: Wiley, 193-236.
4. Paszek P., (1999) Zastosowanie teorii zbiorów przybliżonych w wielostopniowym diagnozowaniu medycznym, rozprawa doktorska, Uniwersytet Śląski w Katowicach Instytut Informatyki.
5. Wakulicz-Deja A., Paszek P., (1997) Diagnose Progressive Encephalopathy Applying the Rough Set Theory. International Journal of Medical Informatics, 46, 119-127
6. Wakulicz-Deja A., Boryczka M., Paszek P., (1998) Discretization of continuous attrib-utes on Decision System in Mitochondrial Encephalomyopathies. Lecture Notes in Computer Science 1424, Springer-Verlag, Berlin, 483-490.