

Raport

Temat: Reguły decyzyjne w Rough Set Exploration System

Wykonał: Artur Kompata

Przygotowanie Zbioru Danych

Opis: Dla wybranego zbioru danych z poprzedniego laboratorium

Zbiór danych: WATER_POTABILITY

- Liczba przykładów: 419 rekordów
- Liczba atrybutów warunkowych: 9 (ph, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic_carbon, Trihalomethanes, Turbidity)
- Atrybut decyzyjny: Potability (wartości: 0, 1)
- Format: RSES (dyskretyzacja wartości na zakresy: range1, range2)

Table: WATER_POTABILITY										
419 / 10	ph	Hardn...	Solids	Chlora...	Sulfate	Condu...	Organ...	Trihal...	Turbid...	Potabi...
O:1	range2	range2	range1	range2	range2	range2	range1	range2	range1	0
O:2	range1	range1	range1	range1	range2	range2	range1	range1	range2	0
O:3	range2	range2	range1	range2	range2	range1	range2	range2	range1	0
O:4	range2	range2	range1	range2	range2	range1	range2	range2	range2	0
O:5	range2	range1	range1	range1	range2	range1	range1	range1	range1	0
O:6	range1	range2	range1	range2	range2	range1	range1	range1	range1	0
O:7	range2	range2	range1	range2	range2	range1	range1	range2	range1	0
O:8	range2	range2	range1	range1	range1	range2	range1	range2	range2	0
O:9	range2	range1	range1	range2	range1	range1	range1	range1	range1	0
O:10	range2	range2	range1	range2	range2	range2	range2	range2	range2	0
O:11	range2	range1	range2	range2	range2	range1	range2	range2	range1	0
O:12	range2	range1	range1	range1	range1	range1	range2	range2	range1	0
O:13	range2	range1	range1	range2	range1	range1	range2	range2	range2	0
O:14	range2	range2	range1	range1	range2	range1	range1	range2	range2	0
O:15	range1	range2	range2	range2	range2	range2	range1	range2	range2	0
O:16	range2	range2	range2	range2	range2	range1	range2	range1	range2	0
O:17	range2	range2	range1	range2	range2	range2	range1	range2	range2	0
O:18	range2	range2	range1	range1	range2	range1	range1	range2	range1	0
O:19	range2	range2	range1	range1	range2	range2	range1	range2	range1	0
O:20	range2	range2	range1	range2	range2	range2	range2	range1	range2	0

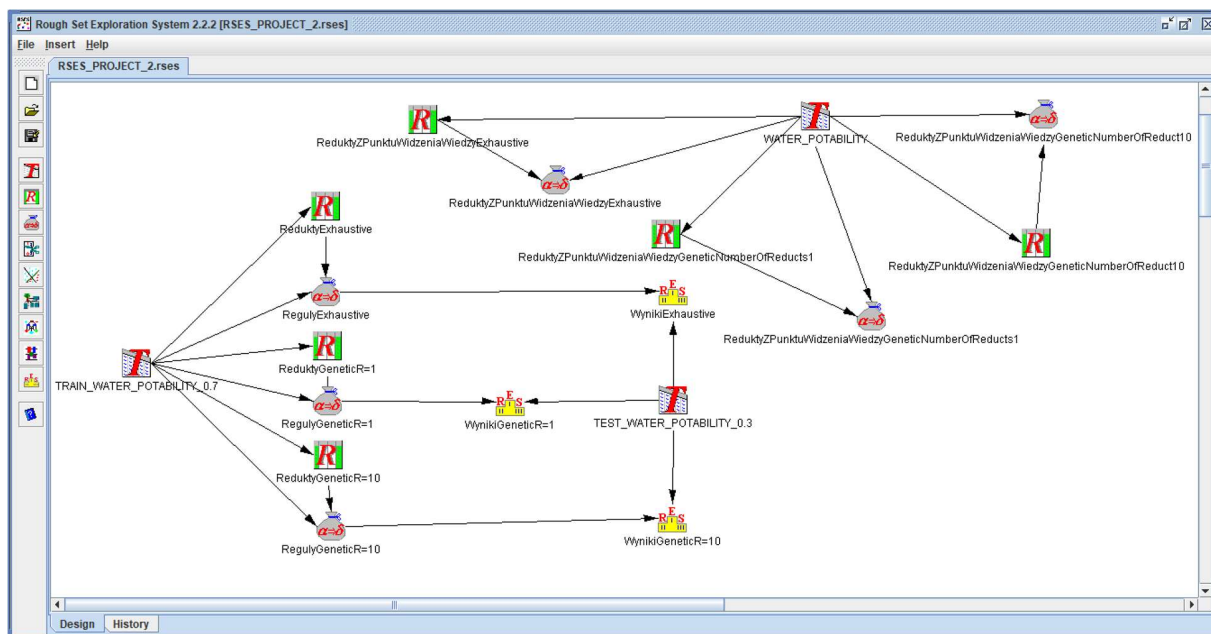
Przygotowane zbiory:

- WATER_POTABILITY (pełny zbiór: 419 rekordów)
- WATER_POTABILITY_0.7 (zbiór testowy: 293 rekordów, 70% danych) z poprzedniego zadania
- WATER_POTABILITY_0.3 (zbiór testowy: 126 rekordów, 30% danych) z poprzedniego zadania

Generowanie reduktów i reguł

Opis: Wygeneruj redukt za pomocą oprogramowania RSES, korzystając z dwóch wybranych algorytmów, określ parametry algorytmów. Na podstawie reduktu wygeneruj reguły

decyzyjne, przedstaw zestawienie w postaci tabeli opisujące parametry zbioru danych, uzyskane reduktory oraz uzyskany zbiór reguł z punktu widzenia reprezentacji wiedzy



Algorytm 1: Exhaustive Algorithm

Parametry:

*Przy zaznaczenie opcji Full discernibility powodowało, że algorytm zwracał pusty zbiór reduktów. Stało się tak mimo tego, że zbiór danych został wcześniej oczyszczony z

duplikatów, niespójności oraz braków danych. W konsultacji z prowadzącymi nie udało się jednoznacznie zidentyfikować przyczyny tego zachowania ani znaleźć stabilnego obejścia problemu, dlatego w dalszych eksperymentach zrezygnowano z tej opcji na rzecz Object related discernibility.

Algorytm 2: Genetic Algorithm (Genetyczny)

Parametry pierwszego:

- Number of reducts = 1
- HighSpeed

Parametry drugiego:

- Number of reducts = 10
- HighSpeed

Algorytm	Reduktów z całego zbioru	Reduktów z testowego zbioru	Reguł z całego zbioru	Reguł z testowego zbioru
Exhaustive	169	230	131	137
GeneticR=1	139	119	107	610
GeneticR=10	168	228	130	135

Klasyfikacja na podstawie wygenerowanych reguł

Opis: Uwzględniając z poprzedniego laboratorium podział zbioru na części trenujące i testowe w procesie klasyfikacji

- zastosuj do części treningowych zbioru danych wygenerowany redukt
- wygeneruje reguł decyzyjne na podstawie reduktu- dokonaj klasyfikacji korzystając z modelu w postaci reguł decyzyjnych, określ parametry klasyfikacji

Przedstaw zestawienie w postaci tabeli opisujące parametry zbioru danych oraz uzyskane wyniki klasyfikacji.

Parametry klasyfikacji (train and test):

Test table "WATER_POTABILITY_0.3" using rule set

General test mode

☒ Generate confusion matrix

☐ Classify new cases

Conflicts resolve by:

☒ Simple voting

☐ Standard voting

Rules from set:

RulesExhaustive

OK

Results to:

ResultsExhaustive

Cancel

Dokładne parametry klasyfikacji:

- Zbiór testowy: WATER_POTABILITY_0.3 (126 obiektów)
- Strategie: **Simple voting**
- Tryb: **Generate confusion matrix**

Metryka	Exhaustive	GeneticR=1	GeneticR=10
Accuracy	0.667	0.651	0.667
Dokładność klasy 0	0.703	0.676	0.703
Dokładność klasy 1	0.615	0.615	0.615
TPR klasa 0	0.72	0.71	0.72
TPR klasa 1	0.59	0.57	0.59
Liczba prawdziwych pozytywów	32	32	32
Liczba fałszywych pozytywów	20	20	20
Liczba prawdziwych negatywów	52	50	52
Liczba fałszywych negatywów	22	24	22
Liczba testowanych obiektów	126	126	126

Pokrycie	1	1	1
----------	---	---	---

Wnioski z porównania

Opis: Dokonaj porównania wyników klasyfikacji, dla systemów regułowych skonstruowanych na podstawie dwóch różnych reduktów.

GeneticR=1 vs. Exhaustive:

- Exhaustive osiąga wyższą dokładność (0.667 vs 0.651)
- Exhaustive ma lepszą czułość dla klasy 0 (0.72 vs 0.71)
- Exhaustive ma porównywalne reguły do GeneticR=10 (137 vs 135)
- Różnica w dokładności wynosi 1.6 pp na korzyść Exhaustive

GeneticR=10 vs. Exhaustive:

- Rezultaty są praktycznie identyczne (0.667 dla obu)
- GeneticR=10 generuje 168 reduktów vs 169 Exhaustive
- GeneticR=10 generuje 135 reguł vs 137 Exhaustive
- Metody są zbliżone wydajnościowo

Ocena Poszczególnych Algorytmów

Opis: Który z algorytmów indukcji reduktów jest lepszy z punktu widzenia reprezentacji wiedzy, który z punktu widzenia klasyfikacji, uzasadnij odpowiedź

Z punktu widzenia reprezentacji wiedzy Exhaustive i GeneticR=10 (ex aequo)

Uzasadnienie:

Liczba reduktów:

- Exhaustive: 169 (pełny zbiór) - gwarantuje niezawodność
- GeneticR=10: 168 (praktycznie identyczne) - heurystyka prawie równa pełnemu przeszukiwaniu
- GeneticR=1: 139 (znacznie mniej) - niedostateczna eksploracja

Liczba reguł (złożoność reprezentacji):

- Exhaustive: 137 reguł - zwięzła wiedza
- GeneticR=10: 135 reguł – najzwięźlejsza
- GeneticR=1: 610 reguł - katastrofalna redundancja

Stosunek reguł do reduktów:

- Exhaustive: 0.811 - wysoka jakość
- GeneticR=10: 0.804 - najwyższa jakość
- GeneticR=1: 4.388 - zbyt wiele reguł na redukt

Wnioskowanie:

- Exhaustive i GeneticR=10 reprezentują wiedzę efektywnie
- Obie metody znajdują minimalne atrybuty niezbędne do klasyfikacji
- GeneticR=10 osiąga rezultaty porównywalne do pełnego przeszukiwania przy mniejszej złożoności obliczeniowej

Ocena: GeneticR=10 jest lepszy dla praktyki (niższa złożoność obliczeniowa), ale Exhaustive zapewnia teoretyczną gwarancję pełności.

Z Punktu Widzenia Jakości Klasyfikacji Exhaustive i GeneticR=10 (ex aequo)

Uzasadnienie:

Dokładność całkowita:

- Exhaustive: 0.667 (66.7%)
- GeneticR=10: 0.667 (66.7%)
- GeneticR=1: 0.651 (65.1%)
- Różnica: Exhaustive i GeneticR=10 lepsze o 1.6 pp

Czułość dla klasy 0 (TPR):

- Exhaustive: 0.72
- GeneticR=10: 0.72
- GeneticR=1: 0.71
- Przewaga: Exhaustive i GeneticR=10

Czułość dla klasy 1 (TPR):

- Exhaustive: 0.59
- GeneticR=10: 0.59
- GeneticR=1: 0.57
- Przewaga: Exhaustive i GeneticR=10

Macierz konfuzji:

- Exhaustive i GeneticR=10: Lepsze prawdziwe negatywy (52 vs 50)

- Exhaustive i GeneticR=10: Mniej fałszywych negatywów (22 vs 24)

Wnioskowanie:

- GeneticR=10 osiąga zbieżność do wyniku Exhaustive

Ocena: GeneticR=10 jest praktycznie optymalny - osiąga wyniki Exhaustive przy mniejszych wymaganiach obliczeniowych.