Indukcja reguł decyzyjnych oparta o selekcję cech

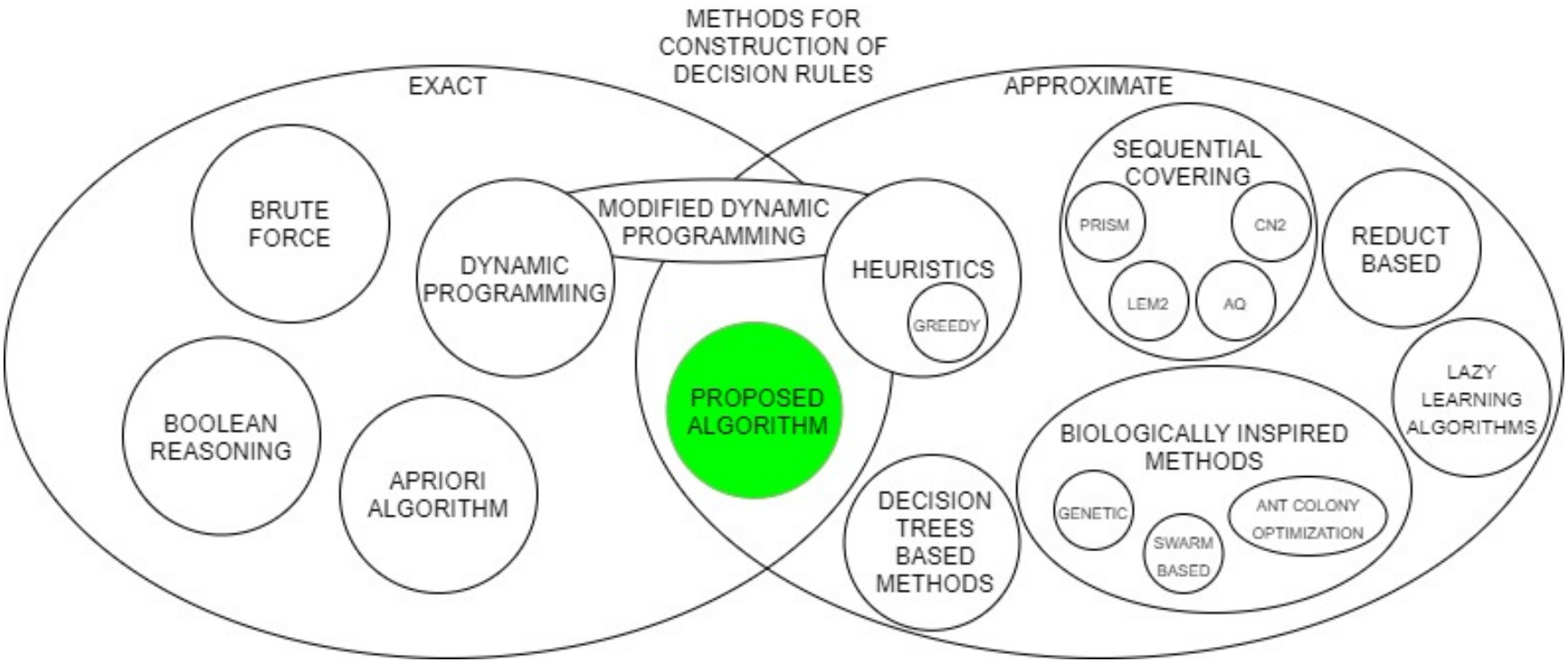
Krzysztof Żabiński

Uniwersytet Śląski
Wydział Nauk Ścisłych i Technicznych
Instytut Informatyki

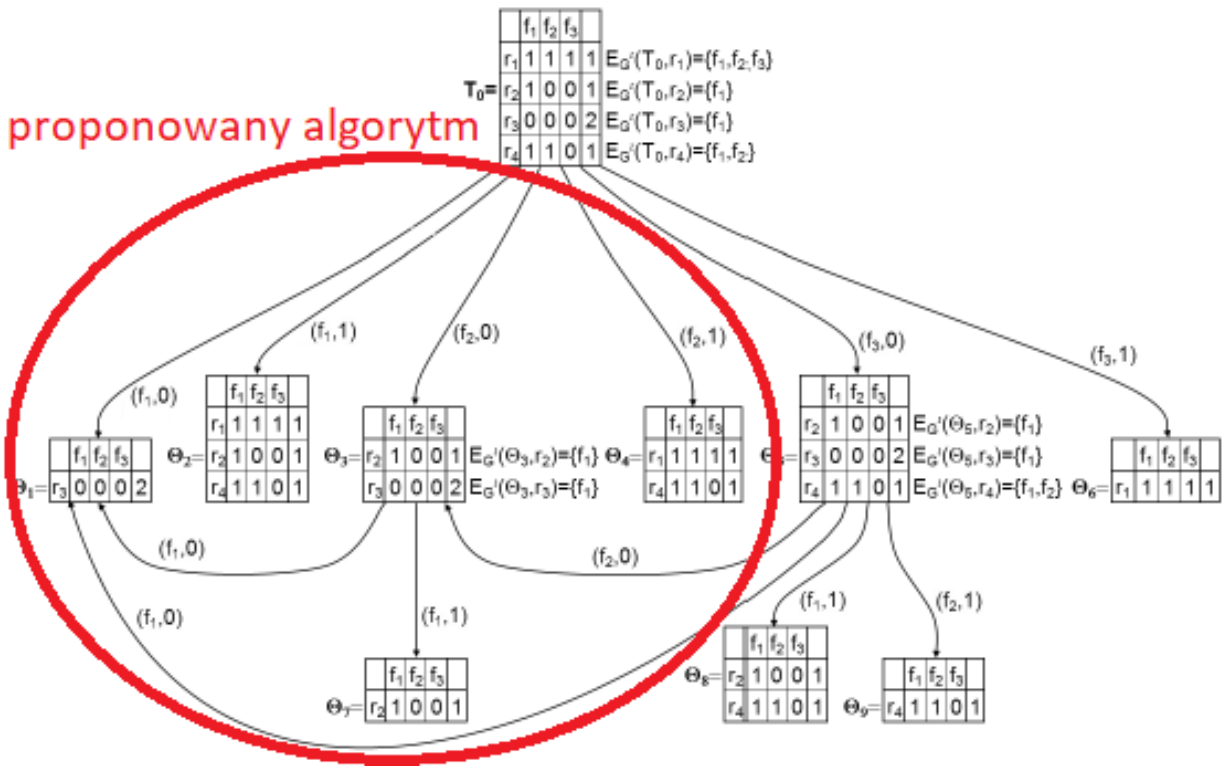
2023

Spis treści

Podejścia do generowania reguł



Umiejscowienie algorytmu wzgledem istniejacych metod



Model EAV

Proponowane podejście generowania reguł decyzyjnych jest oparte na reprezentacji tablicy decyzyjnej w postaci EAV (entity-attribute-value), które pojawia się w literaturze[1]. W każdym wierszu tabeli w postaci EAV wykorzystanej w podejściu znajduje się:

- nazwa atrybutu,
- wartość atrybutu,
- wartość atrybutu decyzyjnego,
- numer wiersza w oryginalnej tabeli decyzyjnej.

[1] Kowalski, M.; Stawicki, S. SQL-Based Heuristics for Selected KDD Tasks over Large Data Sets. Proceedings of the Federated Conference on Computer Science and Information Systems. IEEE, 2012, pp. 303–310.

Postać macierzowa

data table T			
f1	f2	f3	decision
high	bad	small	1
high	good	big	2
high	bad	big	3
low	bad	big	1
low	good	big	2
low	bad	small	3

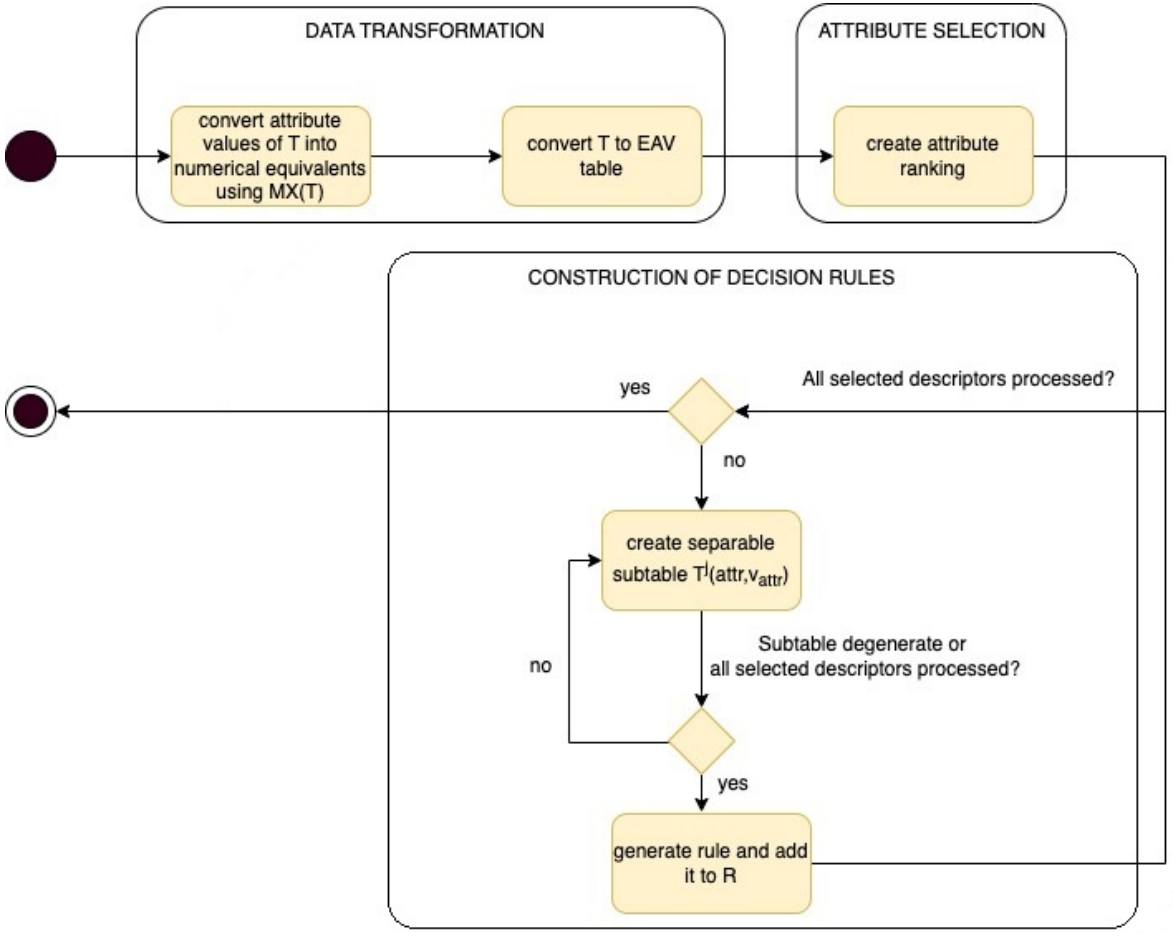
matrix_form						
f1_low	f1_high	f2_bad	f2_good	f3_small	f3_big	decision
0	1	1	0	1	0	1
0	1	0	1	0	1	2
0	1	1	0	0	1	3
1	0	1	0	0	1	1
1	0	0	1	0	1	2
1	0	1	0	1	0	3

Postać macierzowa - transformacja na EAV

EAV with average values from matrix format			
attribute	value	decision	row
f1	0.50	1	1
f2	0.67	1	1
f3	0.33	1	1
f1	0.50	2	2
f2	0.33	2	2
f3	0.67	2	2
f1	0.50	3	3
f2	0.67	3	3
f3	0.67	3	3
f1	0.50	1	4
f2	0.67	1	4
f3	0.67	1	4
f1	0.50	2	5
f2	0.33	2	5
f3	0.00	2	5
f1	0.50	3	6
f2	0.67	3	6
f3	0.33	3	6

Standard deviation of average values	
attribute	value
f2	0.19
f3	0.10
f1	0.00

Etapy proponowanego algorytmuu



Średnia złożoność obliczeniona $O(n)$, pesymistyczna $O(n^2)$.

Ranking atrybutów

Ranking atrybutów jest tworzony na podstawie obliczania odchylenia standardowego względem każdego atrybutu w ramach klasy decyzyjnej, korzystając z polecenia:

```sql
SELECT attribute, STDDEV(average_value) AS quality FROM
(
    SELECT e.attribute, e.decision, AVG(v.id) AS average_value
    FROM eav e JOIN values v ON e.value = v.value
    GROUP BY attribute, decision
) attribute_average_values
GROUP BY attribute
ORDER BY quality DESC;
```

Rysunek: Komenda SQL obliczająca std dla każdego atrybutu.

Wartość progowa dla wyboru atrybutów

Tworzenie rankingu atrybutów z wartością progową:

1. obliczenie std per klasa decyzyjna dla każdego atrybutu (w języku SQL grupowanie po atrybucie i klasie decyzyjnej),
2. obliczenie std per klasa decyzyjna dla wszystkich wartości atrybutów z tabeli EAV (w języku SQL grupowanie tylko po klasie decyzyjnej) - jest to możliwe ze względu na zastosowanie jednorodnych numerycznych odpowiedników dla wartości każdego atrybutu,
3. przygotowanie rozkładu skumulowanego dla std ad.1 i dla std ad.2,
4. wybór atrybutów, dla których krzywa rozkładu przyrasta wolniej niż krzywa wspólna dla std ad.2.

Przykład

Exemplary decision table

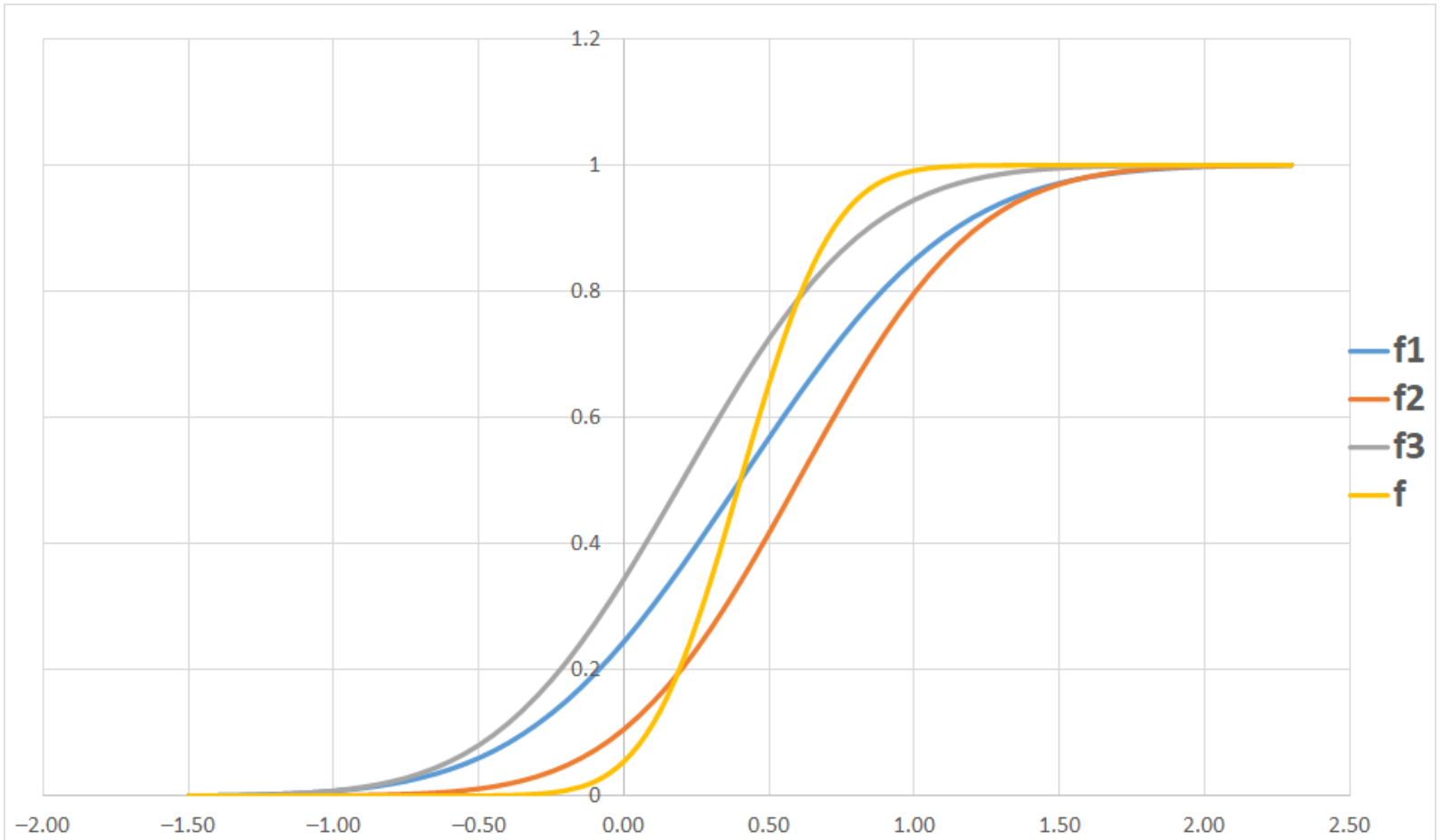
f_1	f_2	f_3	d
1	1	1	1
0	1	0	2
1	1	0	2
0	0	1	3
1	0	0	3

Exemplary decision table in EAVD model

attribute	value	decision	row
f1	1	1	1
f2	1	1	1
f3	1	1	1
f1	0	2	2
f2	1	2	2
f3	0	2	2
f1	1	2	3
f2	1	2	3
f3	0	2	3
f1	0	3	4
f2	0	3	4
f3	1	3	4
f1	1	3	5
f2	0	3	5
f3	0	3	5

Przykład c.d.

Rysunek poniżej pokazuje rozkład skumulowany dla atrybutów i wartości progowej



Brakujące wartości

- Algorytm jest dostosowany do pracy z brakującymi wartościami.
- Możliwość działania bez preprocessingu danych.
- W trakcie transformacji do postaci macierzowej brakujące wartości są zastępowane zerami.

Wyniki eksperymentalne

- Zbiory eksperymentalne wybrane z UCI Machine Learning Repository i Kaggle.
- Cel eksperymetów:
  - analiza uzyskanych wyników z punktu widzenia reprezentacji wiedzy (długość i wsparcie indukowanych reguł);
  - analiza uzyskanych wyników z punktu widzenia klasyfikacji.
- Porównanie z innymi podejściami.
- Wyniki porówane z wykorzystaniem testu Wilcoxona.

Wyniki eksperymentalne

Tabela: Dokładność klasyfikacji dla proponowanego algorytmu - manualne i automatyczne usuwanie atrybutów

data set	100% attr.		80% attr.		60% attr.		automatic	
	accuracy	std	accuracy	std	accuracy	std	accuracy	std
balance-scale	0.89	0.15	0.93	0.14	0.73	0.12	0.93	0.12
breast-cancer	0.92	0.24	0.94	0.24	0.88	0.24	0.94	0.22
cars	0.96	0.15	0.85	0.14	0.84	0.14	0.92	0.13
flags	0.93	0.12	0.92	0.12	0.91	0.11	0.92	0.11
hayes-roth-data	0.92	0.29	0.88	0.29	0.9	0.26	0.91	0.26
house-votes	0.98	0.12	0.98	0.1	0.97	0.1	0.98	0.1
lymphography	0.95	0.11	0.94	0.13	0.92	0.11	0.94	0.11
mushroom	0.87	0.32	0.87	0.3	0.87	0.29	0.87	0.28
nursery	0.85	0.16	0.85	0.15	0.86	0.14	0.86	0.13
shuttle-landing-control	0.78	0.25	0.8	0.27	0.81	0.27	0.81	0.25
soybean-small	0.75	0.28	0.77	0.26	0.78	0.26	0.79	0.24
zoo-data	0.95	0.25	0.95	0.22	0.88	0.2	0.95	0.2
tic-tac-toe	0.94	0.27	0.95	0.26	0.93	0.25	0.95	0.25
$diabetes_{p}rediction_{d}ataset$	0.89	0.15	0.89	0.14	0.91	0.13	0.92	0.13
letter-recognition	0.85	0.19	0.87	0.15	0.89	0.15	0.89	0.15

Wyniki eksperymentalne

Tabela: Dokładność klasyfikacji dla proponowanego algorytmu - zbiory bez i z brakującymi wartościami

data set	no missing		5% missing		10% missing		15% missing	
	accuracy	std	accuracy	std	accuracy	std	accuracy	std
balance-scale	0,93	0,12	0,93	0,12	0,84	0,12	0,8	0,14
breast-cancer	0,94	0,22	0,93	0,23	0,88	0,23	0,83	0,24
cars	0,92	0,13	0,9	0,14	0,9	0,13	0,79	0,14
flags	0,92	0,11	0,91	0,11	0,84	0,12	0,86	0,12
hayes-roth-data	0,91	0,26	0,88	0,26	0,84	0,28	0,79	0,27
house-votes	0,98	0,1	0,96	0,1	0,96	0,1	0,89	0,11
lymphography	0,94	0,11	0,9	0,11	0,9	0,11	0,91	0,12
mushroom	0,87	0,28	0,85	0,29	0,79	0,3	0,77	0,28
nursery	0,86	0,13	0,85	0,13	0,84	0,14	0,83	0,15
shuttle-landing-control	0,81	0,25	0,81	0,26	0,73	0,26	0,75	0,26
soybean-small	0,79	0,24	0,76	0,25	0,78	0,25	0,69	0,25
zoo-data	0,95	0,2	0,8	0,21	0,78	0,21	0,72	0,21
tic-tac-toe	0,95	0,25	0,91	0,25	0,87	0,27	0,82	0,27
diabetes-prediction	0,92	0,13	0,89	0,13	0,86	0,13	0,81	0,13
letter-recognition	0,89	0,15	0,87	0,16	0,89	0,15	0,85	0,17

Wyniki eksperymentalne

Tabela: Dokładność klasyfikacji dla proponowanego algorytmu - zbiory bez i z brakującymi wartościami zastąpionymi przez MCV

data set	no missing		5% missing		10% missing		15% missing	
	accuracy	std	accuracy	std	accuracy	std	accuracy	std
balance-scale	0,93	0,12	0,93	0,12	0,86	0,13	0,86	0,12
breast-cancer	0,94	0,22	0,93	0,23	0,93	0,23	0,88	0,25
cars	0,92	0,13	0,9	0,14	0,88	0,14	0,87	0,13
flags	0,92	0,11	0,91	0,11	0,86	0,12	0,82	0,12
hayes-roth-data	0,91	0,26	0,88	0,26	0,88	0,29	0,87	0,29
house-votes	0,98	0,1	0,96	0,1	0,94	0,11	0,91	0,1
lymphography	0,94	0,11	0,9	0,11	0,85	0,11	0,85	0,11
mushroom	0,87	0,28	0,85	0,29	0,81	0,31	0,84	0,32
nursery	0,86	0,13	0,85	0,13	0,86	0,13	0,77	0,13
shuttle-landing-control	0,81	0,25	0,81	0,26	0,77	0,27	0,78	0,27
soybean-small	0,79	0,24	0,76	0,25	0,76	0,25	0,73	0,25
zoo-data	0,95	0,2	0,8	0,21	0,78	0,22	0,74	0,2
tic-tac-toe	0,95	0,25	0,91	0,25	0,91	0,26	0,82	0,27
diabetes-prediction	0,92	0,13	0,89	0,13	0,91	0,14	0,88	0,13
letter-recognition	0,89	0,15	0,87	0,16	0,81	0,16	0,86	0,16

Wyniki eksperymentalne

Tabela: Dokładność klasyfikacji dla wybranych heurystyk

data set	M		RM		log	
	accuracy	std	accuracy	std	accuracy	std
balance-scale	0.94	0.06	0.95	0.05	0.95	0.05
breast-cancer	0.94	0.03	0.95	0.03	0.95	0.03
cars	0.97	0.11	0.97	0.11	0.97	0.11
flags	0.97	0.08	0.99	0.08	0.99	0.08
hayes-roth-data	0.94	0.07	0.94	0.07	0.94	0.07
house-votes	0.99	0.11	0.99	0.11	0.99	0.11
lymphography	0.94	0.05	0.98	0.06	0.98	0.06
mushroom	0.89	0.06	0.89	0.06	0.89	0.06
nursery	0.88	0.12	0.88	0.12	0.88	0.12
shuttle-landing-control	0.85	0.11	0.85	0.11	0.85	0.11
soybean-small	0.81	0.08	0.81	0.08	0.81	0.08
zoo-data	0.96	0.05	0.96	0.05	0.96	0.05
tic-tac-toe	0.97	0.04	0.98	0.05	0.98	0.05
$diabetes_{p}rediction_{d}ataset$	0.95	0.08	0.95	0.08	0.95	0.08
letter-recognition	0.91	0.09	0.91	0.09	0.91	0.09

Wyniki eksperymentalne

Tabela: Długość i wsparcie dla wybranych heurystyk

data set	M		RM		log	
	length	support	length	support	length	support
balance-scale	3.41	3.38	3.41	3.38	3.41	3.38
breast-cancer	2.97	2.81	2.97	2.81	2.97	2.81
cars	5.57	6.69	5.57	6.69	5.57	6.69
flags	2.04	2.04	2.04	2.04	2.04	2.04
hayes-roth-data	2.88	2.33	2.88	2.33	2.88	2.33
house-votes	3.17	22.86	3.17	22.86	3.17	22.86
lymphography	2.32	5.34	2.32	5.34	2.32	5.34
mushroom	2.22	1112.55	2.22	1112.55	2.22	1112.55
nursery	3.35	253.33	3.35	253.33	3.35	253.33
shuttle-landing-control	3.25	2.5	3.25	2.5	3.25	2.5
soybean-small	2.95	12.34	2.95	12.34	2.95	12.34
zoo-data	4.22	12.35	4.22	12.35	4.22	12.35
tic-tac-toe	4.12	7.32	4.12	7.32	4.12	7.32
diabetes-prediction	2.95	21222,25	2.95	21222,25	2.95	21222,25
letter-recognition	3.84	563.2	3.84	563.2	3.84	563.2

Wyniki eksperymentalne

Tabela: Dokładność klasyfikacji dla Random Forest - zbiory bez i z brakującymi wartościami zastąpionymi przez MCV

data set	no missing		5% missing		10% missing		15% missing	
	accuracy	std	accuracy	std	accuracy	std	accuracy	std
balance-scale	0,91	0,03	0,91	0,03	0,91	0,03	0,91	0,03
breast-cancer	0,87	0,16	0,87	0,16	0,87	0,16	0,87	0,16
cars	0,89	0,08	0,89	0,08	0,89	0,08	0,89	0,08
flags	0,9	0,11	0,9	0,11	0,9	0,11	0,9	0,11
hayes-roth-data	0,91	0,12	0,89	0,12	0,89	0,12	0,89	0,12
house-votes	0,96	0,06	0,96	0,06	0,96	0,06	0,96	0,06
lymphography	0,82	0,08	0,82	0,08	0,82	0,08	0,82	0,08
mushroom	0,86	0,05	0,86	0,05	0,86	0,05	0,86	0,05
nursery	0,81	0,09	0,81	0,09	0,81	0,09	0,81	0,09
shuttle-landing-control	0,81	0,14	0,81	0,14	0,81	0,14	0,81	0,14
soybean-small	0,79	0,11	0,79	0,11	0,79	0,11	0,79	0,11
zoo-data	0,78	0,06	0,78	0,06	0,78	0,06	0,78	0,06
tic-tac-toe	0,9	0,03	0,9	0,03	0,9	0,03	0,9	0,03
diabetes-prediction	0,91	0,06	0,91	0,06	0,91	0,06	0,91	0,06
letter-recognition	0,83	0,12	0,83	0,12	0,83	0,12	0,83	0,12

Wyniki eksperymentalne

Tabela: Średnie wsparcie dla proponowanego algorytmu - zbiory bez i z brakującymi wartościami

data set	no missing	5% missing	10% missing	15% missing
balance-scale	2,44	2,48	2,52	2,68
breast-cancer	2,61	2,70	2,61	2,81
cars	79,88	80,96	80,19	82,85
flags	1,78	1,82	1,87	2,01
hayes-roth-data	3,81	3,94	3,99	4,04
house-votes	31,23	31,48	30,89	34,58
lymphography	2,85	2,98	3,00	3,06
mushroom	816,58	827,66	809,47	870,72
nursery	111,73	115,41	115,23	120,24
shuttle-landing-control	1,53	1,58	1,58	1,58
soybean-small	9,46	9,76	9,85	10,76
zoo-data	8,22	8,40	8,39	8,95
tic-tac-toe	6,43	6,73	6,80	7,32
diabetes-prediction	9233,43	9452,59	9476,72	9936,11
letter-recognition	99,3	99,7	100,15	102,42

Wyniki eksperymentalne

Tabela: Średnie wsparcie dla proponowanego algorytmu - zbiory bez i z brakującymi wartościami zastąpionymi przez MCV

data set	no missing	5% missing	10% missing	15% missing
balance-scale	2,44	2,41	2,52	2,60
breast-cancer	2,61	2,59	2,61	2,63
cars	79,88	76,95	80,19	82,05
flags	1,78	1,81	1,87	1,89
hayes-roth-data	3,81	3,94	3,99	4,04
house-votes	31,23	30,87	30,89	31,11
lymphography	2,85	2,91	3,00	3,10
mushroom	816,58	796,73	809,47	826,88
nursery	111,73	115,10	115,23	117,23
shuttle-landing-control	1,53	1,54	1,58	1,63
soybean-small	9,46	9,61	9,85	9,98
zoo-data	8,22	8,00	8,39	8,59
tic-tac-toe	6,43	6,64	6,80	6,84
diabetes-prediction	9233,43	9278,51	9476,72	9520,38
letter-recognition	99,3	97,23	100,15	101,76

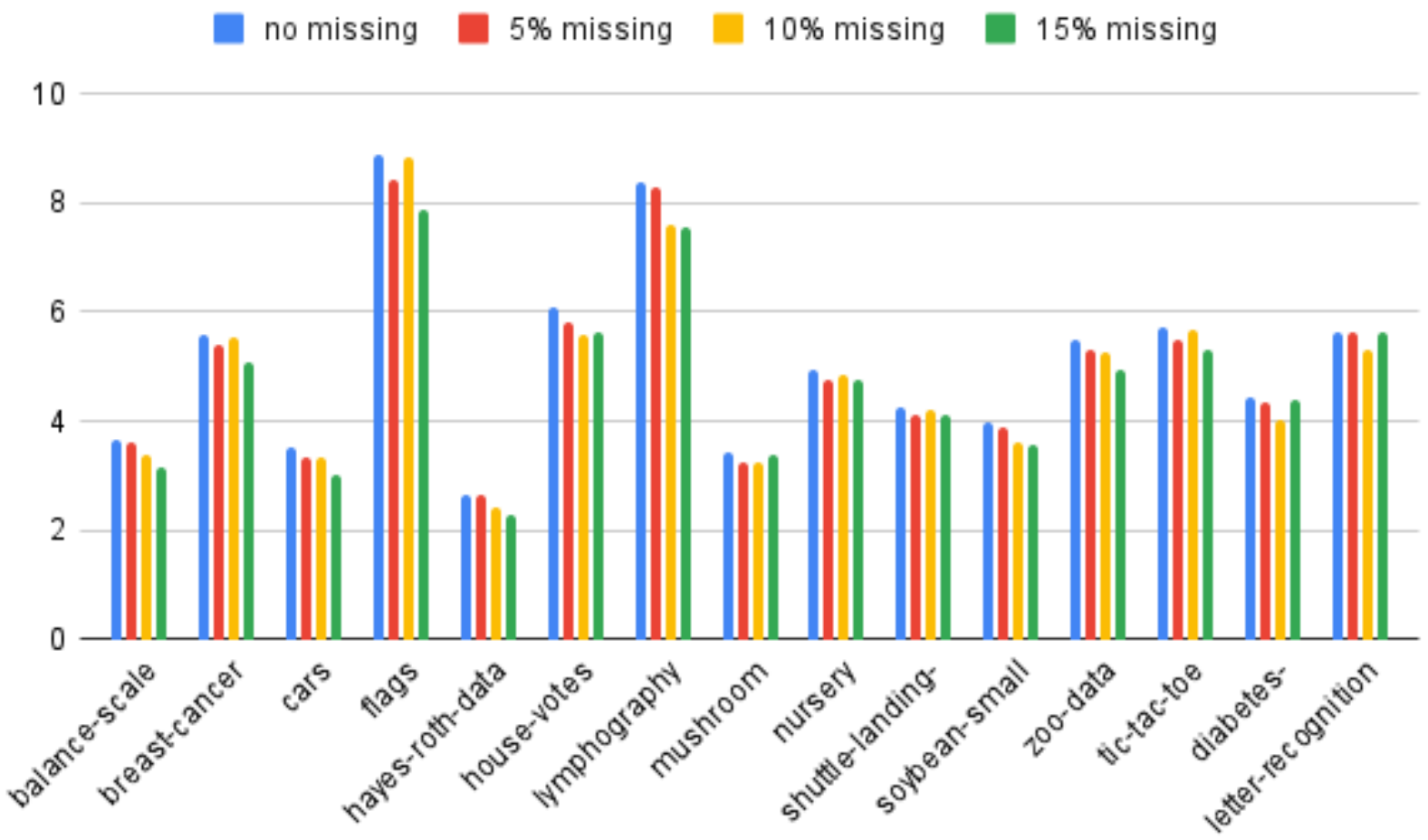
Wyniki eksperymentalne

Tabela: Średnie wsparcie dla Random Forest - zbiory bez i z brakującymi wartościami zastąpionymi przez MCV

data set	no missing	5% missing	10% missing	15% missing
balance-scale	2,76	2,76	2,76	2,76
breast-cancer	4,74	4,74	4,74	4,74
cars	86,73	86,73	86,73	86,73
flags	5,22	5,22	5,22	5,22
hayes-roth-data	3,57	3,57	3,65	3,65
house-votes	61,61	61,61	61,61	61,61
lymphography	7,73	7,73	7,73	7,73
mushroom	930,64	930,64	930,64	930,64
nursery	174,11	174,11	174,11	174,11
shuttle-landing-control	2,1	2,1	2,1	2,1
soybean-small	11,87	11,87	11,87	11,87
zoo-data	14,6	14,6	14,6	14,6
tic-tac-toe	11,98	11,98	11,98	11,98
diabetes-prediction	9400,91	9400,91	9400,91	9400,91
letter-recognition	186,6	186,6	186,6	186,6

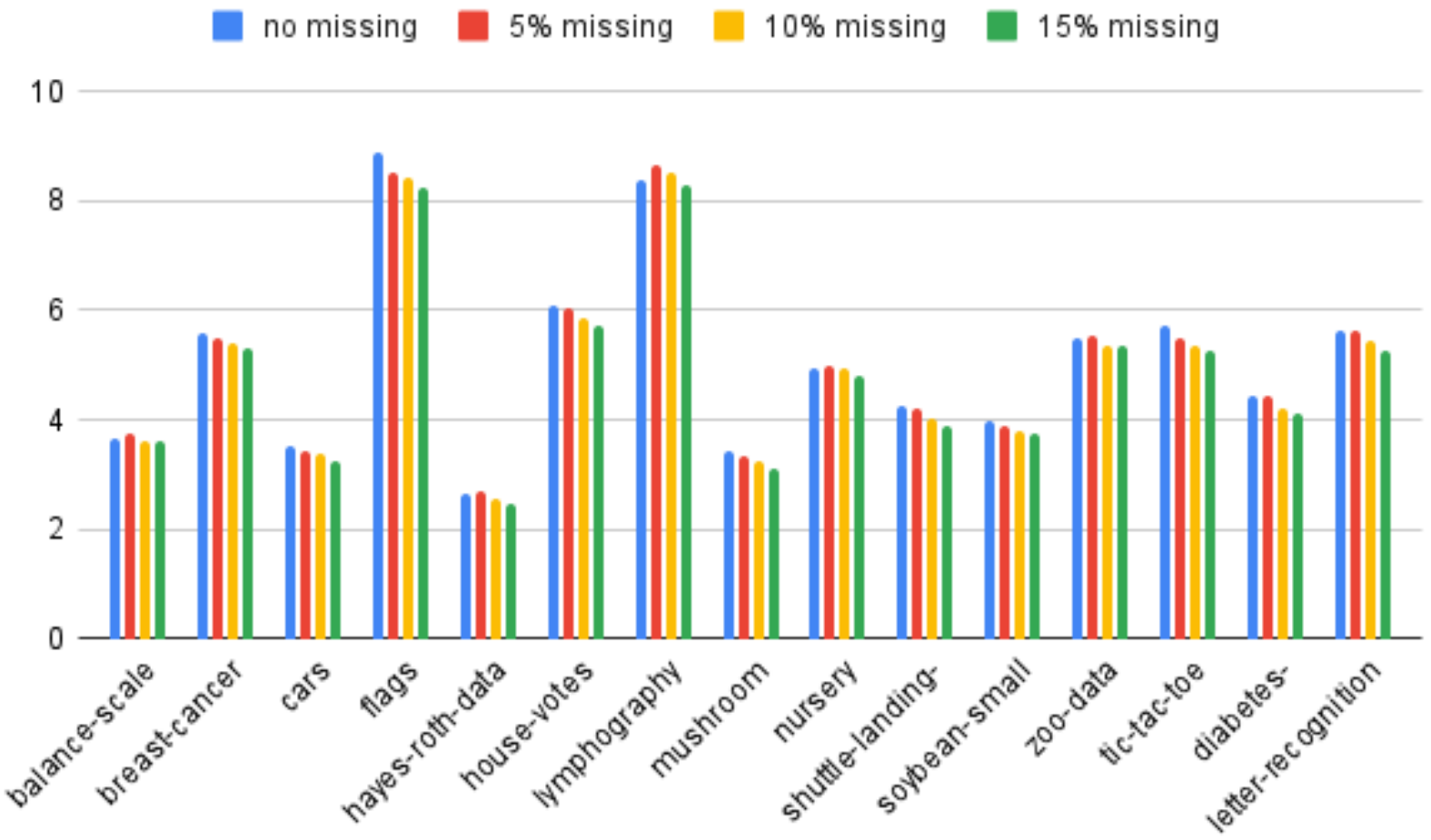
Wyniki eksperymentalne

Średnia długość dla proponowanego algorytmu - zbiory bez i z brakującymi wartościami:



Wyniki eksperymentalne

Średnia długość dla proponowanego algorytmu - zbiory bez i z brakującymi wartościami zastąpionymi przez MCV:



Wnioski

- Zaproponowany algorytm umożliwia generowanie reguł o jakości klasyfikacji zbliżonej do innych znanych metod przy zachowaniu reguł dobrych z punktu widzenia reprezentacji wiedzy.
- Przewagą algorytmu jest możliwość bezpośredniej pracy ze zbiorami o brakujących wartościach.
- Algorytm jest w stanie pracować z dowolnymi atrybutami kategorycznymi.
- Algorytm pozwala na efektywną pracę z dużymi zbiorami danych - potwierdza to niewielka złożoność obliczeniowa.
- Zaproponowana metoda budowania rankingu atrybutów została porównana z metodą opartą o algorytm Relief.