

Laboratorium Matematyka Dyskretna

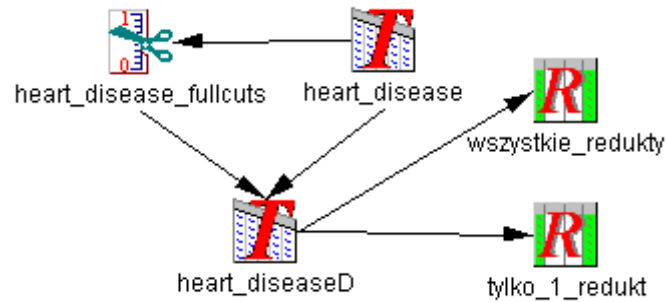
Zbiory przybliżone Zbiór danych Heart disease

15-12-2011r.

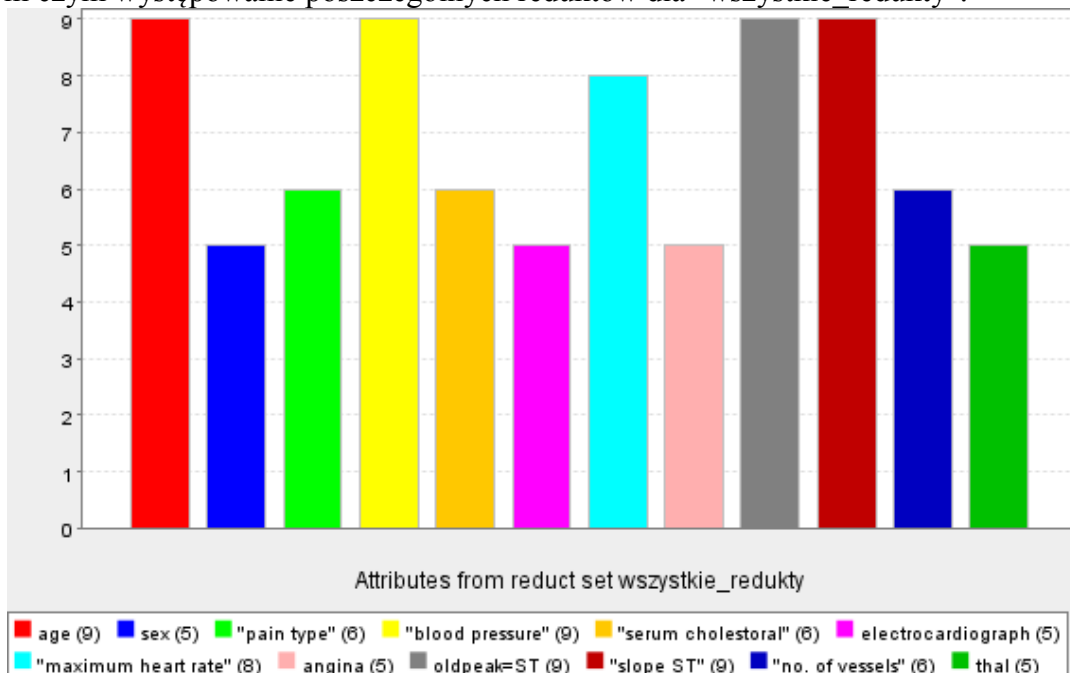
**Skład sekcji:
Remigiusz Szczepanik
Orestes Sporyś
Michał Wesoły
Łukasz Bochenek
Jan Zawada
Sławomir Szymura**

Zbiór danych Heart disease

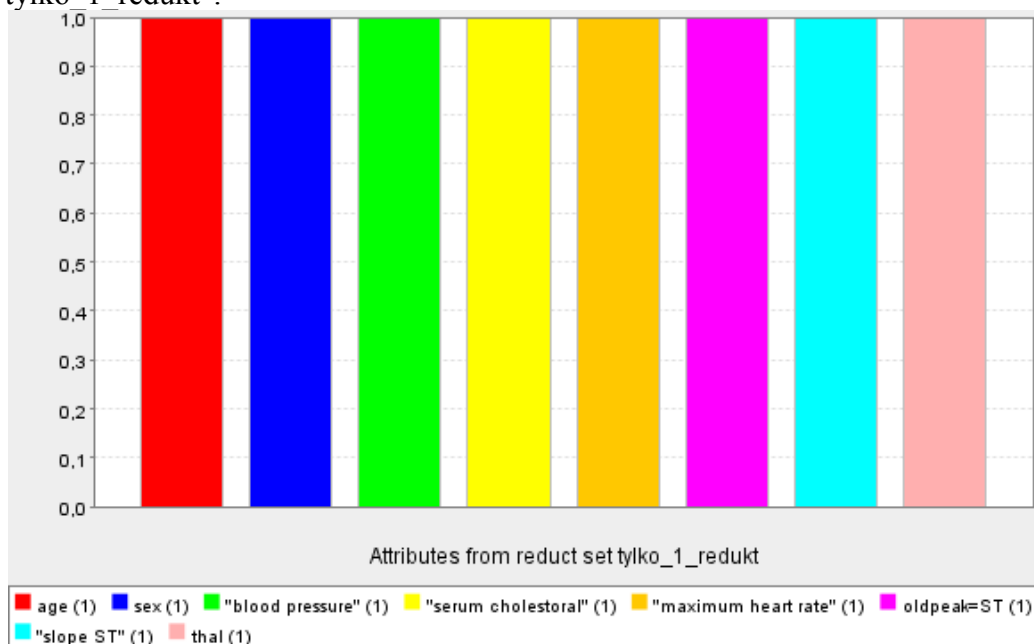
a-b) Dyskretyzacja i wyznaczenie na całej tablicy wszystkich reduktów oraz tylko 1 reduktu przedstawia się następująco:



Przyczyną występowania poszczególnych reduktów dla "wszystkie_redukty":






oraz dla "tylko_1_redukt":



Niezredukowana tablica, 10-krotna krosvalidacja, algorytm genetyczny

Niezredukowana tablica, 10-krotna krosvalidacja, algorytm LEM2

Niezredukowana tablica, 10-krotna krosvalidacja, algorytm genetyczny, skracanie reguł ze współczynnikiem 0,9

Results of experiments by cross-validation method: "heart_disease_G...   

		Predicted			No. of obj.	Accuracy	Coverage
		2	1				
Actual	2	10.3	1.7	12	0.868	1	
	1	2.4	12.2	15	0.85	0.976	
	True positive rate	0.83	0.88				

Total number of tested objects: 27
Total accuracy: 0.845
Total coverage: 0.985

The screenshot shows a software window with the title "Results of experiments by cross-validation method: heart_disease_ge...". The window contains a confusion matrix and summary statistics.

		Predicted					
			2	1	No. of obj.	Accuracy	Coverage
Actual			2	1			
	2		9.7	2.3	12	0.805	1
	1		1.4	13.6	15	0.902	1
	True positive rate		0.88	0.85			

Below the confusion matrix, the following summary statistics are displayed:

- Total number of tested objects: 27
- Total accuracy: 0.863
- Total coverage: 1

Results of experiments by cross-validation method: "heart_disease_1..."

		Predicted				
Actual				No. of obj.	Accuracy	Coverage
	2	9.6	2.4	12	0.796	1
	1	2.7	12	15	0.818	0.982
	True positive rate	0.78	0.83			

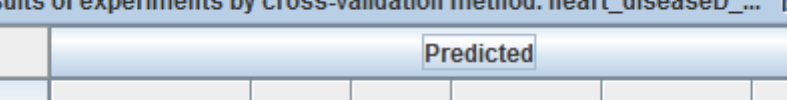
Total number of tested objects: 27
 Total accuracy: 0.809
 Total coverage: 0.989

Results of experiments by cross-validation method: "heart_disease_L..."

		Predicted				
		2	1	No. of obj.	Accuracy	Coverage
Actual	2	9	3	12	0.762	1
	1	1.6	13.4	15	0.893	1
	True positive rate	0.86	0.81			

Total number of tested objects: 27
 Total accuracy: 0.83
 Total coverage: 1

Zredukowana tablica (1 redukt), 10-krotna krosvalidacja, algorytm genetyczny



Results of experiments by cross-validation method: heart_diseaseD_...

		Predicted				
		2	1	No. of obj.	Accuracy	Coverage
Actual	2	9.4	2.4	12	0.776	0.984
	1	3.2	11.3	15	0.783	0.97
	True positive rate	0.73	0.82			

Total number of tested objects: 27
 Total accuracy: 0.787
 Total coverage: 0.974

Zredukowana tablica (1 redukt), 10-krotna krosvalidacja, algorytm LEM2

Results of experiments by cross-validation method: heart_diseaseD_1...

		Predicted				
Actual				No. of obj.	Accuracy	Coverage
	2	6.8	1.1	12	0.862	0.649
	1	1.8	8	15	0.823	0.654
	True positive rate	0.77	0.88			

Total number of tested objects: 27

Total accuracy: 0.843

Total coverage: 0.656

Zredukowana tablica (1 redukt), 10-krotna krosvalidacja, algorytm genetyczny, skracanie reguł ze współczynnikiem 0,9

Results of experiments by cross-validation method: heart_diseaseDX...

		Predicted				
Actual		2	1	No. of obj.	Accuracy	Coverage
	2	10.2	1.8	12	0.855	1
	1	2.8	11.9	15	0.808	0.981
	True positive rate	0.79	0.87			

Total number of tested objects: 27
Total accuracy: 0.828
Total coverage: 0.989

Zredukowana tablica (1 redukt), 10-krotna krosvalidacja, algorytm genetyczny, skracanie reguł ze współczynnikiem 0,8

Results of experiments by cross-validation method: heart_diseaseD_...

		Predicted				
Actual				No. of obj.	Accuracy	Coverage
	2	10	2	12	0.826	1
	1	2.5	12.5	15	0.837	1
	True positive rate	0.79	0.86			

Total number of tested objects: 27
Total accuracy: 0.833
Total coverage: 1

Zredukowana tablica (1 redukt), 10-krotna krosvalidacja, algorytm LEM2, skracanie reguł ze współczynnikiem 0,9

The screenshot shows a software window with the title "Results of experiments by cross-validation method: heart_diseaseD_1...". The window contains a confusion matrix and summary statistics.

		Predicted				
			2	1	No. of obj.	Accuracy
Actual	2	9.9	1.9	12	0.843	0.983
	1	2.1	12.7	15	0.85	0.986
	True positive rate	0.84	0.87			

Summary statistics:

- Total number of tested objects: 27
- Total accuracy: 0.85
- Total coverage: 0.985

Zredukowana tablica (1 redukt), 10-krotna krosvalidacja, algorytm LEM2, skracanie reguł ze współczynnikiem 0,8

Results of experiments by cross-validation method: "heart_diseaseD_...

		Predicted				
Actual		2	1	No. of obj.	Accuracy	Coverage
	2	9.1	2.9	12	0.775	1
	1	1.6	13.4	15	0.892	1
	True positive rate	0.85	0.82			

Total number of tested objects: 27
Total accuracy: 0.833
Total coverage: 1

o) Opracowanie oraz porównanie uzyskanych wyników prezentujemy w formie tabelki:

Czy udało się zredukować zbiory danych?

Poprzez 10-krotną krosvalidację liczba obiektów w zbiorach danych uległa zmniejszeniu 10-krotnemu. Z 270 do 27 obiektów dla każdego typu klasyfikacji.

Dokładność klasyfikacji, liczba reguł oraz liczba obiektów nierozpoznanych:

		10-krotna krosvalidacja					
		Algorytm genetyczny					
		Nie zredukowana tablica			Zredukowana tablica		
		Bez skracania reguł	Skracanie reguł ze współczynnikiem 0,9	Skracanie reguł ze współczynnikiem 0,8	Bez skracania reguł	Skracanie reguł ze współczynnikiem 0,9	Skracanie reguł ze współczynnikiem 0,8
liczba reguł dla kolejnych fold	Dokładność	0,812	0,845	0,863	0,787	0,828	0,833
	Pokrycie	0,985	0,985	1	0,974	0,989	1
	Liczba obiektów nierozpoznanych	4	4	0	7	3	0
	1	154	124	89	175	128	97
	2	142	132	84	171	131	87
	3	120	118	87	158	125	93
	4	156	121	80	170	127	102
	5	144	117	99	140	138	106
	6	140	129	100	169	125	83
	7	150	118	70	165	128	98
	8	149	109	92	162	120	100
	9	136	120	96	156	124	94
	10	152	125	98	142	122	89
	Średnia	144,3	121,3	89,5	160,8	126,8	94,9

		10-krotna krosvalidacja					
		Algorytm LEM2					
		Nie zredukowana tablica			Zredukowana tablica		
		Bez skracania reguł	Skracanie reguł ze współczynnikiem 0,9	Skracanie reguł ze współczynnikiem 0,8	Bez skracania reguł	Skracanie reguł ze współczynnikiem 0,9	Skracanie reguł ze współczynnikiem 0,8
liczba reguł dla kolejnych fold	Dokładność	0,869	0,809	0,83	0,843	0,85	0,833
	Pokrycie	0,619	0,989	1	0,656	0,985	1
	Liczba obiektów nierozpoznanych	103	3	0	93	4	0
	1	64	62	56	70	64	48
	2	73	55	56	65	67	58
	3	71	57	57	64	61	46
	4	67	57	45	69	63	53
	5	69	58	58	71	66	51
	6	80	68	51	70	60	64
	7	72	60	53	76	73	58
	8	66	67	48	66	66	65
	9	65	59	61	78	62	60
	10	68	60	50	67	60	56
	Średnia	69,5	60,3	53,5	69,6	64,2	55,9

Na podstawie wyników zawartych w tabelkach możemy w prosty sposób przedstawić różnicę pomiędzy algorytmem genetycznym, a algorytmem LEM2.

Dokładność obu tych algorytmów jest bardzo zbliżona, dlatego możemy powiedzieć, że nie jest to na pewno decydującym czynnikiem w wyborze pomiędzy tymi dwoma algorytmami.

Pokrycie natomiast ma już zasadniczy wpływ na zredukowane zbiory danych. W przypadku braku skracania reguł LEM2 ma o prawie 1/3 mniejsze pokrycie niż algorytm genetyczny. Skracanie reguł ze współczynnikiem 0,9 w obu przypadkach jest zbliżone, natomiast użycie współczynnika 0,8 powoduje, że oba algorytmy mają 100% pokrycie.

Ze współczynnika pokrycia wynika liczba obiektów nierozpoznanych. Zależność między algorytmami jest podobna jak przy pokryciu.

Ostatnim czynnikiem porównawczym algorytmów jest liczba reguł użytych do klasyfikacji. Z powodu 10-krotnej krosvalidacji dla każdej klasyfikacji składowo przypada 10 różnych ilości reguł, dlatego zdecydowaliśmy się wprowadzić średnią by ułatwić porównywanie. Z tabelki jasno wynika, że algorytm LEM2 potrzebuje znacząco (czasami nawet ponad 2-krotnie mniej) liczby reguł do wyprowadzenia klasyfikacji.

p) Wnioski:

Po tym laboratorium dowiedzieliśmy się jak działają zbiory danych oraz jak możemy je redukować za pomocą danych algorytmów (w naszym przypadku były to algorytm genetyczny i LEM2).

Dodatkowo poszerzyliśmy nasze słownictwo o znajomość słów takich jak dyskretyzacja czy krosvalidacja – jesteśmy pewni, że teraz już nas nikt nimi nie zaskoczy.

Przeprowadzając praktyczne prace w programie RSES poznaliśmy jego działanie jak i potrafimy już używać algorytmów na zbiorach danych. W naszym przypadku, używając algorytmów LEM2 i genetycznego, znamy już ich zalety jak i wady użycia.

Uważamy, że dla zadanego zbioru danych z chorobami serca użycie algorytmu genetycznego byłoby odpowiedniejsze mimo, iż używa on więcej reguł do klasyfikacji.