

Indukcyjne metody analizy danych

Ćwiczenie 2

Indukcja drzew decyzyjnych C4.5 w R

Prowadzący: dr inż. Paweł Myszkowski

Student: Piotr Bielak, 218137

WT 17:05

Wrocław, 10 kwietnia 2018r.

Spis treści

1	Wprowadzenie	3
1.1	Cel ćwiczenia	3
1.2	Algorytm C4.5	3
2	Analiza zbiorów danych	4
2.1	Zbiór danych – "Diabetes"	4
2.2	Zbiór danych – "Glass"	5
2.3	Zbiór danych – "Wine"	7
3	Eksperyment	9
3.1	Założenia	9
3.2	Badanie parametrów algorytmu C4.5	9
3.3	Wyniki krosvalidacji	10
3.3.1	Zbiór danych – "Diabetes"	10
3.3.2	Zbiór danych – "Glass"	11
3.3.3	Zbiór danych – "Wine"	12
4	Wnioski	13

1 Wprowadzenie

1.1 Cel ćwiczenia

1.2 Algorytm C4.5

2 Analiza zbiorów danych

2.1 Zbiór danych – "Diabetes"

Nazwa klasy	Liczba instancji	% instancji
1 (chory)	500	65 (%)
0 (zdrowy)	268	35 (%)

Tabela 1: Udział procentowy klas w zbiorze "Diabetes".

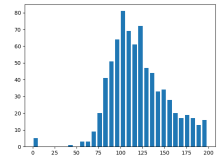
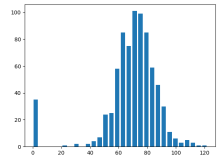
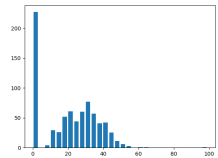
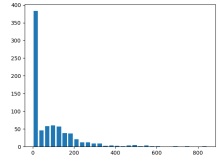
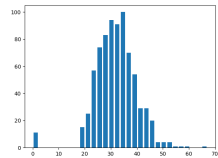
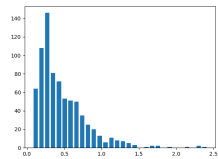
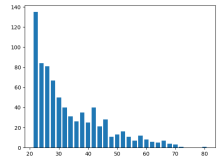
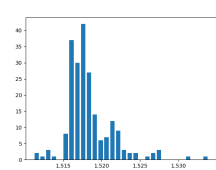
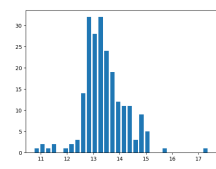
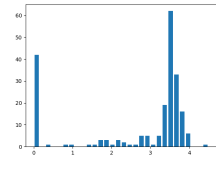
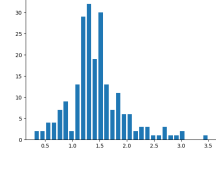
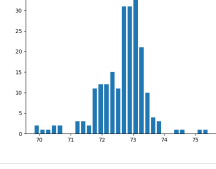
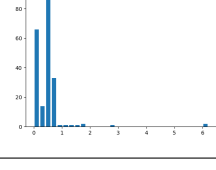
Nazwa atrybutu	Min	Max	Średnia	Ochyl. stand.	Rozkład
Glucose	0	199	120.89	31.95	
BloodPressure	0	122	69.11	19.34	
SkinThickness	0	99	20.54	15.94	
Insulin	0	846	79.80	115.17	
BMI	0	67.1	31.99	7.88	
DiabetesPedigreeFunction	0.08	2.42	0.47	0.33	
Age	21	81	33.24	11.75	

Tabela 2: Atrybuty zbioru danych "Diabetes".

2.2 Zbiór danych – "Glass"

Nazwa klasy	Liczba instancji	% instancji
1 (building_windows_float_processed)	70	33 (%)
2 (building_windows_non_float_processed)	76	36 (%)
3 (vehicle_windows_float_processed)	17	8 (%)
4 (vehicle_windows_non_float_processed)	0	0 (%)
5 (containers)	13	6 (%)
6 (tableware)	9	4 (%)
7 (headlamps)	29	13 (%)

Tabela 3: Udział procentowy klas w zbiorze "Glass".

Name	Min	Max	Mean	Std	Distribution
RI	1.51	1.53	1.52	0.00	
Na	10.73	17.38	13.41	0.81	
Mg	0.00	4.49	2.68	1.44	
Al	0.29	3.50	1.44	0.50	
Si	69.81	75.41	72.65	0.77	
K	0.00	6.21	0.50	0.65	

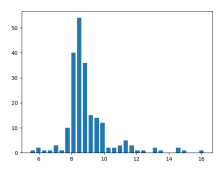
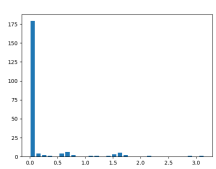
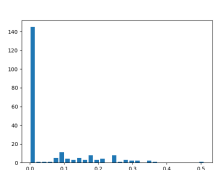
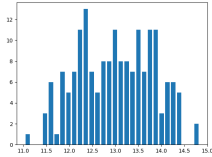
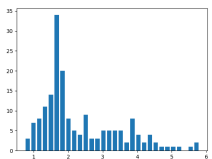
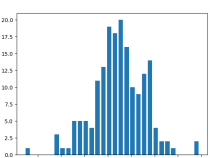
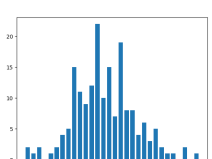
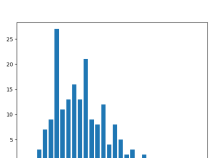
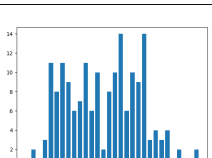
					
Ca	5.43	16.19	8.96	1.42	
					
Ba	0.00	3.15	0.18	0.50	
					
Fe	0.00	0.51	0.06	0.10	

Tabela 4: Atrybuty zbioru danych "Glass".

2.3 Zbiór danych – "Wine"

Nazwa klasy	Liczba instancji	% instancji
1 (Class 1)	59	33 (%)
2 (Class 2)	71	40 (%)
3 (Class 3)	48	27 (%)

Tabela 5: Udział procentowy klas w zbiorze "Wine".

Name	Min	Max	Mean	Std	Distribution
Alcohol	11.03	14.83	13.00	0.81	
Macil_acid	0.74	5.80	2.34	1.11	
Ash	1.36	3.23	2.37	0.27	
Alcalinity_of_ash	10.60	30.00	19.49	3.33	
Magnesium	70.00	162.00	99.74	14.24	
Total_phenols	0.98	3.88	2.30	0.62	

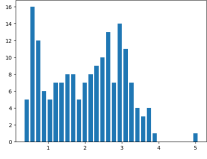
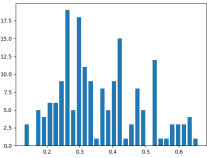
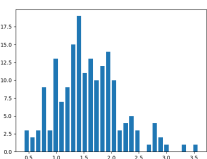
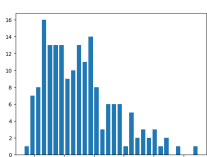
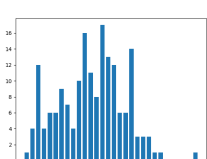
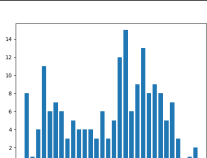
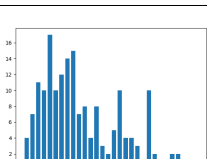
Flavanoids	0.34	5.08	2.03	1.00	
Nonflavanoid_phenols	0.13	0.66	0.36	0.12	
Proanthocyanins	0.41	3.58	1.59	0.57	
Intensity	1.28	13.00	5.06	2.31	
Hue	0.48	1.71	0.96	0.23	
OD280_OD315	1.27	4.00	2.61	0.71	
Proline	278.00	1680.00	746.89	314.02	

Tabela 6: Atrybuty zbioru danych "Wine".

3 Eksperyment

3.1 Założenia

Eksperyment został podzielony na dwie fazy. Pierwsza służyła do zbadania parametrów algorytmu C4.5, natomiast druga miała na celu ocenę jakości działania drzewa decyzyjnego dla wybranych zbiorów danych (**Diabetes**, **Glass** oraz **Wine**). Podobnie jak w przypadku algorytmu klasyfikatora Bayesa została tutaj również zastosowana krowalidacja zwykła oraz stratyfikowana i zostały obliczone miary *accuracy*, *precision*, *recall* oraz *F1*.

Szczegółowe wyniki (wykresy, tabelki, wizualizacje drzew) tego eksperymentu są przedstawione w kolejnych podrozdziałach.

3.2 Badanie parametrów algorytmu C4.5

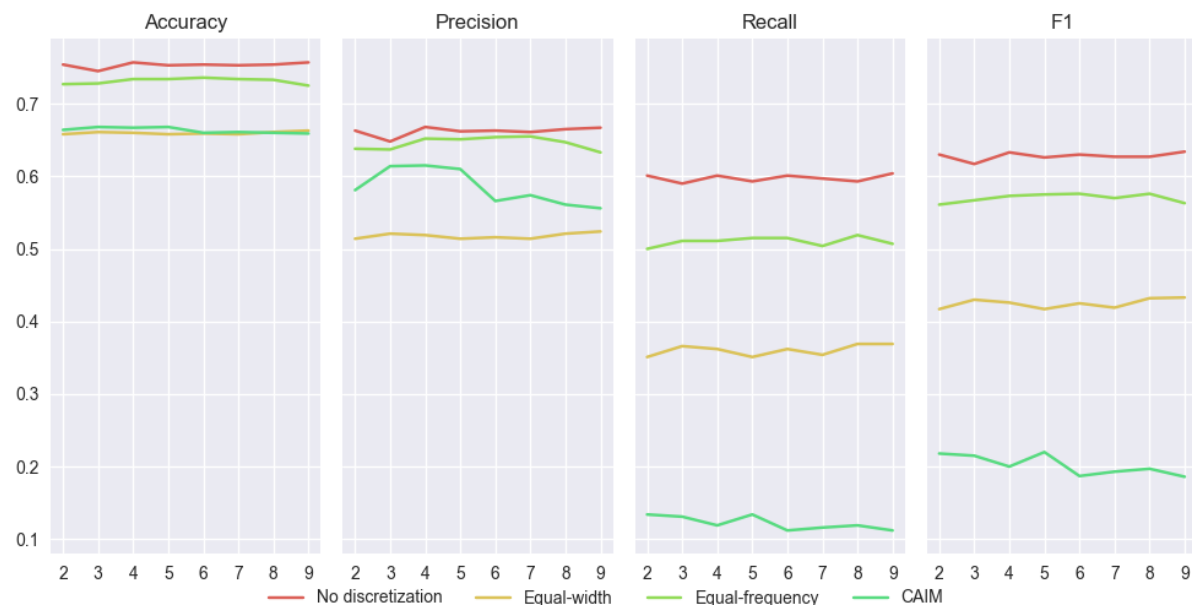
3.3 Wyniki krosvalidacji

Poniżej zostały przedstawione wyniki zastosowania krosvalidacji (z parametrem w postaci liczby podzbiorów; zmieniający się od 2 do 9 ze skokiem 1) zbiorów danych, które:

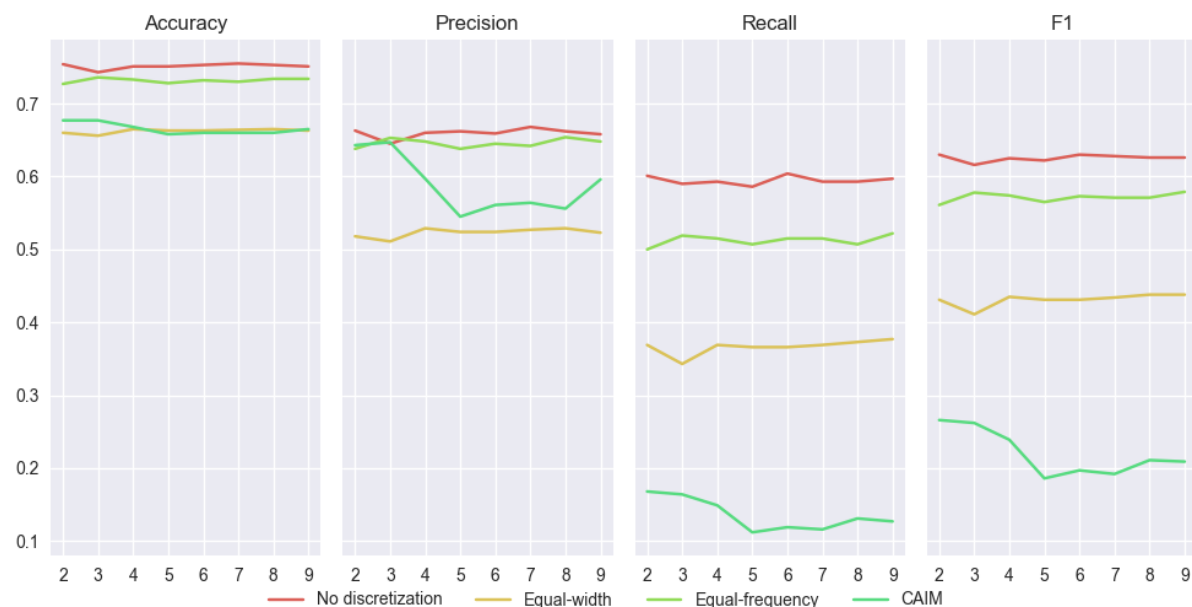
-

a następnie w ramach danego procesu krosvalidacji, wyznaczono wartości miar oceny jakości klasyfikatora. Dodatkowo zostały zamieszczone tabelki z dokładnymi wartościami tych miar.

3.3.1 Zbiór danych – "Diabetes"



Rysunek 1: Wykresy wartości metryk dla zbioru "Diabetes" – krosvalidacja zwykła.



Rysunek 2: Wykresy wartości metryk dla zbioru "Diabetes" – krosvalidacja stratyfikowana.

3.3.2 Zbiór danych – "Glass"



Rysunek 3: Wykresy wartości metryk dla zbioru "Glass" – krosvalidacja zwykła.



Rysunek 4: Wykresy wartości metryk dla zbioru "Glass" – krosvalidacja stratyfikowana.

3.3.3 Zbiór danych – "Wine"



Rysunek 5: Wykresy wartości metryk dla zbioru "Wine" – krosvalidacja zwykła.



Rysunek 6: Wykresy wartości metryk dla zbioru "Wine" – krosvalidacja stratyfikowana.

4 Wnioski

-