Kacper Ptaszek
331 427

Sprawozdanie WSI ćwiczenia 4.

Treść zadania:
Zaimplementować klasyfikator ID3 (drzewo decyzyjne). Atrybuty nominalne, testy tożsamościowe. Podać dokładność i macierz pomyłek na zbiorach: Breast cancer i mushroom.
Założenia dotyczące implementacji klasyfikatora ID3:

- Atrybuty nominalne - każdy atrybut może przyjmować jedną z kilku dozwolonych wartości, zakładamy, że wartość atrybutu to napis, np. "kot", "a", "20-34", ">40".
- Testy tożsamościowe - jeżeli atrybut testowany w danym węźle ma np. 3 dozwolone wartości, np. a, b, c, to z węzła tego wychodzą 3 krawędzie oznaczone: a, b, c.
- brakujące wartości atrybutów traktujemy jako wartość, np. jeżeli symbol '?' oznacza brakującą wartość, a symbole 'a', 'b' wartości normalne, to z naszego punktu widzenia mamy 3 wartości normalne (fachowo: 3 wartości atrybutu): 'a', 'b', '?'.

Zbiór danych dzielę za każdym razem losowo w stosunku 3:2 na zbiór trenujący i zbiór testujący działanie klasyfikatora.

Wyniki dla zbioru Breast Cancer (286 próbek)

Wyniki otrzymałem z 25 uruchomień predykcji klasy i w tabeli umieszczone są dokładności procentowe predykcji zaimplementowanego klasyfikatora.

Min	Max	Średnia	Odchylenie standardowe
59,1	73,2	65,8	3,1

Macierz pomyłek
Dwie różne klasy - no-recurrence-events, recurrence-events

Spodziewane / Otrzymane	no-recurrence-events	recurrence-events
no-recurrence-events, recurrence-events	60	18
recurrence-events	18	19

Wyniki dla zbioru agaricus-lepiota (8124 próbki)

Wyniki również otrzymałem z 25 uruchomień predykcji klasy, a w tabeli umieszczone są dokładności w procentach.

Min	Max	Średnia	Odchylenie standardowe
99,42	99,94	99,6	0,091

Macierz pomyłek
Dwie różne klasy - edible=e, poisonous=p

Spodziewane / Otrzymane	e	p
e	1706	0
p	10	1534

Z otrzymanych wyników widać że dla zbioru danych agaricus-lepiota klasyfikator ID3 znacznie częściej przyporządkowuje odpowiednią klasę. Zbiór breast cancer ma znacznie mniej próbek (około 27 razy mniej), więc drzewo klasyfikatora dla tego zbioru jest na pewno mniej przetrenowane niż drzewo zbioru danych agaricus-lepiota. Aby sprawdzić jeszcze działanie klasyfikatora dla zbioru danych o większej ilości próbek niż zbiór breast cancer i mniejszej niż dla zbioru agaricus-lepiota, przeprowadziłem testy dla zbioru Car Evaluation.

Wyniki dla zbioru Car Evaluation (1728 próbek)

Wyniki również otrzymałem z 25 uruchomień predykcji klasy, a w tabeli umieszczone są dokładności w procentach.

Min	Max	Średnia	Odchylenie standardowe
84,53	92,21	88,57	1,54

Macierz pomyłek
Cztery różne klasy - unacc, acc, good, vgood

Spodziewane / Otrzymane	acc	good	unacc	vgood
acc	108	1	51	1
good	4	10	12	6
unacc	11	1	452	0
vgood	2	9	9	15

Wnioski:

Jeśli dane wejściowe dla klasyfikatora są wysokiej jakości, czyli mamy dużą ilość próbek i wartości atrybutów są znane dla większości próbek to klasyfikator ID3 działa bardzo dobrze. Dla zbioru danych Car Evaluation, gdzie liczba próbek była znacznie większa niż dla zbioru breast cancer, ale mniejsza niż dla zbioru agaricus-lepiota uzyskaliśmy również dobre wyniki w okolicach 88% dokładności predykcji. Za to drzewo decyzyjne dla zbioru breast cancer dawało około 65% dokładności predykcji. Ta niższa dokładność wynika przede wszystkim z małej liczby próbek oraz z mniejszej liczby atrybutów niż dla zbioru agaricus-lepiota, gdzie dokładność predykcji była prawie stu procentowa. Na przykładzie zbioru agaricus-lepiota widać, że z danych wysokiej jakości można utworzyć drzewo decyzyjne, które będzie poprawnie przewidywać klasę pojedynczej próbki w ogromnej większości przypadków.