WSI, Ćwiczenie 4

Michał Kwarciński

Polecenie:

Zaimplementować klasyfikator ID3 (drzewo decyzyjne). Atrybuty nominalne, testy tożsamościowe. Podać dokładność i macierz pomyłek na zbiorach: Breast cancer i mushroom. Dlaczego na jednym zbiorze jest znacznie lepszy wynik niż na drugim? Do potwierdzenia lub odrzucenia postawionych hipotez konieczne może być przeprowadzenie dodatkowych eksperymentów ze zmodyfikowanymi zbiorami danych. Sformułować i spisać wnioski.

Za wynik pozytywny uznałem ten, który niesie ze sobą większe ryzyko, jeśli zostanie błędnie odczytany (dla grzybów - „p", dla raków piersi – „recurrence-events").

Wyniki:

Testy na zbiorze uczącym przy 10 uruchomieniach:

Dla zbioru z grzybami:

- Średnia dokładność wyniosła 100%.
- Macierz pomyłek:

		Klasa rzeczywista	
	Wynik	Pozytywny	Negatywny
Klasa predykowana	Pozytywny	2531,1	0
	Negatywny	0	2342,9

Dla zbioru z nowotworami piersi:

- Średnia dokładność wyniosła 98%.
- Macierz pomyłek:

		Klasa rzeczywista	
	Wynik	Pozytywny	Negatywny
Klasa predykowana	Pozytywny	119,9	0,2
	Negatywny	2,5	49,4

Obserwacje:

Widać, że dla zbioru danych o grzybach drzewo idealnie odzwierciedliło w przewidywaniach rzeczywistość. Przy zbiorze danych o nowotworach zdarzały się nieliczne przypadki błędów.

Przy 10 uruchomieniach dla podziału zbioru 3:2 (dane uczące / dane testujące):

Dla zbioru z grzybami:

- Średnia dokładność wyniosła 100%.
- Macierz pomyłek:

		Klasa rzeczywista	
	Wynik	Pozytywny	Negatywny
Klasa predykowana	Pozytywny	1265,1	0
	Negatywny	0	1171,9

Dla zbioru z nowotworami piersi:

- Średnia dokładność wyniosła 59%.
- Macierz pomyłek:

		Klasa rzeczywista	
	Wynik	Pozytywny	Negatywny
Klasa predykowana	Pozytywny	56,7	14,8
	Negatywny	32,5	10,0

Dla podziału 1:4:

- Średnia dokładność wyniosła 55%.
- Macierz pomyłek:

		Klasa rzeczywista	
	Wynik	Pozytywny	Negatywny
Klasa predykowana	Pozytywny	107,0	32,0
	Negatywny	71,0	19,0

Dla podziału 2:3:

- Średnia dokładność wyniosła 56%.
- Macierz pomyłek:

		Klasa rzeczywista	
	Wynik	Pozytywny	Negatywny
Klasa predykowana	Pozytywny	82,1	25,2
	Negatywny	50,6	14,1

Dla podziału 4:1:

- Średnia dokładność wyniosła 62%.
- Macierz pomyłek:

		Klasa rzeczywista	
	Wynik	Pozytywny	Negatywny
Klasa predykowana	Pozytywny	28,8	6,8
	Negatywny	15,1	6,3

Obserwacje i wnioski:

Widać, że dla zbioru o grzybach, dla danych testujących dokładność predykcji algorytmu jest idealna, a dla zbioru nowotworowego jest duży spadek. Część błędów może wynikać z potrzeby użycia przez drzewo atrybutu, który ma niesprecyzowaną klasę wyjściową (to znaczy wyniki z danych uczących były zarówno pozytywne, jak i negatywne). Dla tego zbioru można zauważyć dużą przewagę liczebną w wynikach pozytywnych w klasie rzeczywistej, gdy w przypadku drugiego rodzaju danych stosunek wyników pozytywnych i negatywnych jest zbliżony. Podczas testów zmian proporcji widać również, że średnie dopasowanie wzrosło z 55% do 62%, lecz równie dobrze może wynikać to z losowości i zmniejszającej się, wraz ze wzrostem części danych uczących, ilości danych testujących.

Można dojść do wniosku, że zbiór danych o nowotworach piersi nie jest odpowiedni do stworzenia poprawnie działającego drzewa ID3. Dane są zbyt złożone (3 na 9 atrybutów ma powyżej 8 możliwych parametrów), istnieje zbyt duża dysproporcja wyników pozytywnych do negatywnych (201 wyników negatywnych do 85 pozytywnych) oraz w zbiorze uczącym jest zbyt mało rekordów.