Mateusz Kołacz, 336360

WSI – Ćwiczenie 4

Zadanie

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.

Poniżej kilka wskazówek ogólnych do tego ćwiczenia

- 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.
- Na tym ćwiczeniu klasyfikator trenuje się na zbiorze trenującym, a ocenia jego jakość na zbiorze testującym. Należy losowo podzielić zbiór danych na trenujący i testujący w stosunku 3:2.
- Jeżeli zbiór danych zawiera numery lub identyfikatory wierszy to należy je wyrzucić - nie chcemy uczyć się identyfikatorów wierszy.
- Brakujące wartości atrybutów taktujemy 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’, ‘?’.
- Tak naprawdę to nie musimy rozumieć dziedziny problemu - na wejściu mamy napisy, na wyjściu napisy, nie ważne czy klasyfikujemy sekwencje DNA, grzyby, czy samochody.
- Nazwa pliku ze zbiorem danych jest parametrem algorytmu klasyfikacji, kod klasyfikatora powinien być w stanie obsłużyć inny zbiór danych o tym samym rozkładzie kolumn (czyli nie należy wpisywać wartości atrybutów „na sztywno” w kodzie).
- W repozytorium ze zbiorami danych zwykle w plikach „.names” jest napisane, który atrybut to klasa (czyli wartości której kolumny mamy się nauczyć przewidywać).

Wyniki

Zaimplementowano klasyfikator ID3 z atrybutami nominalnymi i testami tożsamościowymi. Losowo dzielono dane w stosunku 3:2 na zbiór treningowy i testowy, a także wykonano po 50 iteracji dla każdego ze zbioru danych. Po przeprowadzeniu pierwszych testów na zbiorach danych “Breast_cancer” i “mushroom” otrzymano następujące rezultaty:

[Mushroom] => Dokładność 99.99%

Przewidywany / Rzeczywisty	**Trujący**	**Jadalny**
Trujący	1568	0
Jadalny	0	1681

[Breast_cancer] => Dokładność 63.06%

Przewidywany / Rzeczywisty	**Jednorazowy**	**Powtarzający się**
Jednorazowy	60	22
Powtarzający się	21	12

Wstępna hipoteza, którą postawiono, jest następująca: *wielkość zbioru danych ma wpływ na dokładność predykcji modelu*. W celu zweryfikowania powyższej hipotezy, postanowiono przeprowadzić dodatkowe eksperymenty ze zmodyfikowanymi zbiorami danych.

Dodatkowe eksperymenty

[Zbiór Mushroom zmniejszony do 300 obserwacji] => Dokładność 100,00%

Przewidywany / Rzeczywisty	**Trujący**	**Jadalny**
Trujący	11	0
Jadalny	0	89

Okazuje się, że liczba obserwacji nie ma wpływu na dokładność predykcji modelu. Wyniki dla zbioru około 300 obserwacji były niemal identyczne (a nawet trochę lepsze) co dla zbioru o wielkości ~8100 rekordów.

Wnioski:

Przy wykorzystaniu algorytmu ID3, wielkość zbioru danych nie zawsze wywiera wpływ na uzyskiwane wyniki.