Krzysztof Terlecki 318454

Politechnika Warszawska

Wprowadzenie do sztucznej inteligencji - ćwiczenie 4

1. Wstęp

1.1. 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.

1.2. Opis algorytmu

Implementujemy klasyczny algorytm drzewa decyzyjnego ID3. Klasa ID3 ma dwie zasadnicze funkcje: 1. do trenowania przyjmującą za argumenty ramkę danych z parametrami i ramkę ze zbiorem odpowiedzi (class), funkcja na podstawie adnych tworzy drzewo; 2. do predykcji na podstawie stworzonego drzewa, przyjmuje za argument dataframe z danymi do predykcji

2. Badanie algorytmu

Na trzech data setach przeprowadziłem testy algorytmu

2.1. Prosty data set

Wpierw testowałem działanie algorytmu na prostym, 8-kolumnowym data secie stworzonym specjalnie z myślą o drzewach decyzyjnych. Skuteczność predykcji wynosiła 100% dla danych treningowych zawierających wszystkie kolumny, a także dla 6 kolumn (był to wynik oczekiwany, świadczący o poprawnym działaniu algorytmu).
Następnie przeszedłem do testów na większych data setach (tych z treści zadania).

2.2. Breast cancer data set

Problem z datasetem był taki, że pojawiały się w nim błędy. Np występowały wartości jakiegoś parametru: 1. a-b, 2. b-c, 3. a-c, przy czym trzecia wartość była prawdopodobnie błędem i pojawiała się tak rzadko, że zdażały się przypadki, że nie było jej w danych treningowych, a w testowych już tak. Uodporniłem algorytm na takie przypadki. W przypadku wartości z którą nie spotkał się w trakcie trenigu zwraca napis "noData".
Po udoskonaleniu algorytmu przeprowadziłem test na rozłącznych, losowo dobranych zbiorach: treningowym (moc: 186), testowym (moc: 100). Skuteczność predykcji wyniosła 65%.

2.3. Mushroom data set

losowo dobranych zbiorach: treningowym (moc: 5123), testowym (moc: 3000). Skuteczność predykcji wyniosła 99,6%.

2.4. Wnioski

Pierwszym wnioskiem jest to, że algorytm został zaimplementowany poprawnie. Widać jednak, że na zbiorze breast cancer miał dużo niższą skuteczność niż na pozostałych. Wynika to z typu problemu opisywanego przez dany zbiór. Drzewo decyzyjne jest narzędziem świetnie radzącym sobie w klasyfikacji. Kiedy na podstawie zadanych cech trzeba zidentyfikować konkretny gatunek (lub tak jak w datasecie mushrom powiedzieć czy jest on trujący lub nie, co jest cechą konkretnych gatunków) lub kiedy trzeba na podstawie parametów pogodowych podjąć decyzję czy wychodzić na dwór czy nie (pierwszy, prosty dataset) algorytm radzi sobie wyśmienicie i osiąga skuteczność bliską 100%. Inaczej jest w przypadku datasetu breast cancer, który dotyczy predykcji zagadnień nowotworowych. Te zagadnienia wpadają w zupełnie inną klasę problemów niż poprzednio omówione. Są to problemy gdzie żadna kombinacja parametrów nie zapewni nam stuprocentowej, lub bliskiej jej pewności o tym, że wystąpi nowotwór czy jego reemisja. Nawet jak byśmy mieli kompletny genom osobnika, niewiele by to nam dało, bo aktywacja protoonkogenów zależy od wielu czynników, także środowiskowych i nie sposób jej przewidzić. W przypadku problemów tej klasy możemy mówić co najwyżej o czynnikach ryzyka. Bądąc w jakiejś grupie wiekowej zwiększa się prawdopodobieństwo na wystąpienie, lub wznowienie choroby, jednak nie jest to wzrost aż tak znaczny. Poza tym bycie w innej grupie wiekowej nie chroni nas w pełni przed takimi nieprzyjemnymi zdarzeniami. W przypadku tego typu problemów predykcja staje się bardzo trudna, a dokładna predykcja wręcz niemożliwa. W tym kontekście wynik 65% wydaje się całkiem niezły.