WSI Drzewo Decyzyjne ID3

Konrad Karpiuk
06.05.2024

1. Opis zadania

Zaimplementować klasyfikator ID3 (drzewo decyzyjne). Podać dokładność i macierz pomyłek na zbiorach: Breast cancer i mushroom. Odpowiedzieć na pytanie dlaczego na jednym zbiorze danych jest znacznie lepszy wynik niż na drugim.

2. Testy na zbiorach

2.1 Breast cancer

Zbiór *Breast cancer* liczy zaledwie 286 wierszy danych, które zostały podzielone losowo w stosunku 3:2 na dane trenujące (171 wierszy) i testujące (115 wierszy). Po stworzeniu drzewa decyzyjnego na podstawie danych trenujących zostało ono przetestowane za pomocą danych testujących. Dokładność uzyskana z 25 uruchomień programu została przedstawiona w poniższej tabeli:

	Minimum	Maksimum	Średnia	Odchylenie standardowe
Dokładność	55,65%	75,65%	65,58%	4,21%

Dokładność jest bardzo niska. Biorąc pod uwagę, że podział klas w zbiorze *Breast cancer* wynosi 201:85, czyli 70,30% danych posiada tę samą klasę, to większą średnią wyników można by było uzyskać wybierając zawsze klasę występującą najczęściej w zbiorze. Świadczyto o braku poprawnej klasyfikacji przez drzewo decyzyjne.

Poniżej przestawiono macierz pomyłek dla zbioru *Breast cancer*. Klasą „pozytywną" jest ta częściej występująca. Liczby podane w tabeli są średnią z 25 uruchomień:

	Rzeczywista klasa pozytywna	Rzeczywista klasa negatywna
Przewidywana klasa pozytywna	66,32	24,96
Przewidywana klasa negatywna	14,62	9,10

Suma przypadków, dla których rzeczywista klasa była pozytywna, wynosi średnio 80,94, co stanowi 70,38% wszystkich wierszy z danych testujących, a więc tylko 0,08% więcej niż procent danych posiadających częstszą klasę w całym zbiorze *Breast cancer*. Świadczy to o tym, że pierwotny zbiór został rzeczywiście losowo podzielony na podzbiór trenujący i podzbiór testujący. Analizując macierz pomyłek można zauważyć skąd wynika taka niska dokładność w przypisywaniu klasy. Drzewo decyzji nie radzi sobie w sytuacji, kiedy powinno danemu wierszowi danych przypisać klasę negatywną. W takiej sytuacji prawie 3 razy częściej błędnie przypisuje temu wierszowi klasę pozytywną.

2.2 Mushroom

Zbiór *mushroom* zawiera 8124 wierszy danych. Jest to o wiele więcej niż w przypadku poprzedniego zbioru. Również rozkład danych jest w tym przypadku bardziej zrównoważony, gdyż na obie klasy przypada po około 50% wierszy. W tym przypadku również zbiór został podzielony w stosunku 3:2 na dane trenujące (4874 wierszy) i testujące (3250 wierszy). Dokładność uzyskana z 25 uruchomień programu została przedstawiona w poniższej tabeli:

	Minimum	Maksimum	Średnia	Odchylenie standardowe
Dokładność	99,29%	99,88%	99,57%	0,17%

W tym przypadku drzewo decyzyjne wykazało się dużo lepszym działaniem niż w przypadku poprzedniego zbioru i prawie we wszystkich przypadkach poprawnie przypisało klasę danemu wierszowi danych.

Poniżej przedstawiono macierz pomyłek dla zbioru mushroom. Liczby podane w tabeli są średnią z 25 uruchomień.

	Rzeczywista klasa pozytywna	Rzeczywista klasa negatywna
Przewidywana klasa pozytywna	1566,84	5,52
Przewidywana klasa negatywna	8,60	1669,04

Tym razem drzewo decyzyjne działa bez zarzutu, około 14 błędnych wyników na 3250 wierszy danych jest bardzo dobrym wynikiem.

3. Różnica w działaniu

Dlaczego dla jednego zbioru danych algorytm ID3 osiągnął dokładność prawie stuprocentową, a dla drugiego nie wykazał się lepszą dokładnością niż przypisanie wszystkim danym testującym częściej występującej w całym zbiorze klasy?

Znaczącą różnicą między zbiorami jest ich wielkość. Ponieważ nie jestem w stanie rozszerzyć zbioru *Breast cancer* o dodatkowe dane, to zbadałem wpływ wielkości wzoru na działanie algorytmu przez ograniczenie zbioru *mushroom* wybierając z niego losowo 286 próbek, tak aby był on równy zbiorowi *Breast cancer* pod względem wielkości.

Dokładność oraz macierz pomyłek dla 25 uruchomień zostały przestawione w poniższych tabelach:

	Minimum	Maksimum	Średnia	Odchylenie standardowe
Dokładność	80,00%	93,91%	88,35%	4,03%

	Rzeczywista klasa pozytywna	Rzeczywista klasa negatywna
Przewidywana klasa pozytywna	48,52	5,76
Przewidywana klasa negatywna	7,64	53,08

Osiągnięta dokładność, mimo że niższa niż ta dla nieograniczonego zbioru, nadal jest dość wysoka, o wiele wyższa niż przy przydziale każdemu wierszowi z danych testujących najczęściej występującej klasy (wtedy byłoby to około 50%). Świadczy to o prawidłowym działaniu drzewa decyzyjnego, oraz o tym, że to nie wielkość zbioru jest kluczową różnicą powodującą brak poprawnego działania algorytmu dla zbioru *Breast cancer*.

Innym powodem słabego działania algorytmu dla zbioru *Breast cancer* może być sama natura problemu opisywanego przez dane. Być może to, czy wystąpi nawrót nowotworu piersi, jest zależne od wszystkich atrybutów zawartych w zbiorze danych, a nawet dla znanej kombinacji wartości atrybutów nie można z całkowitą pewnością przewidzieć, czy taki nawrót nastąpi. Inaczej wygląda sytuacja dla zbioru grzybów. To, czy grzyb jest trujący czy nie, zależy od tego jakiego jest gatunku, a gatunek można precyzyjnie określić na podstawie innych cech grzyba takich jak jego kolor czy kształt.

Podsumowując, klasyfikator jest bardzo przydatnym narzędziem, pozwalającym na dokładne przypisywanie klas obiektom, znając wartości ich atrybutów. Jednak warunkiem poprawnego działania klasyfikatora jest odpowiednia liczba danych trenujących oraz natura problemu, w której brakuje przedstawicieli różnych klas posiadających ten sam lub bardzo podobny zestaw atrybutów.