# WSI Ćwiczenie 4 – Drzewo ID3

Kamil Marszałek 331401

## Zbiór breast cancer

Zbiór breast cancer zawiera dane dotyczące nawrotu raka piersi. Za klasę pozytywną przyjąłem: no-recurrence-event, negatywna: recurrence-event. Zawiera on 10 kolumn danych, przy czym w pierwszej znajduje się klasa. Dla losowo wybranych 60% danych ze zbioru breast cancer budowałem drzewo zgodnie z algorytmem ID3. Pozostałe 40% pozwoliło na pomiar dokładność oraz generację macierzy pomyłek. Wykonałem 100 eksperymentów, gdzie dla każdej wylosowanej kombinacji budowałem nowe drzewo, a później używając zbioru testującego szacowałem dokładność oraz macierz pomyłek.

W tabeli poniżej przedstawiam, jaką dokładność udało się uzyskać:

| Minimum | Średnia | Maksimum | Odchylenie std |
|---|---|---|---|
| 0,54 | 0,65 | 0,74 | 0,04 |

Uśredniona macierz pomyłek – próba testowa zawierała 115 rekordów

| Prawdziwie pozytywne | Prawdziwie negatywne | Fałszywie pozytywne | Fałszywie negatywne |
|---|---|---|---|
| 61,67 | 12,52 | 21,35 | 19,46 |

## Zbiór mushroom

Zbiór mushroom zawiera dane dla jadalnych oraz trujących grzybów. Klasa pozytywna w tym przypadku to: e, klasa negatywna: p. Zawiera 23 kolumny, przy czym w pierwszej znajduje się klasa. Dla losowo wybranych 60% danych ze zbioru Agaricus Lepiota budowałem drzewo zgodnie z algorytmem ID3. Pozostałe 40% posłużyło jako zbiór testujący do oceny dokładności klasyfikatora oraz do generacji macierzy pomyłek. Wykonałem 100 eksperymentów.

W tabeli poniżej przedstawiam jaką dokładność udało się uzyskać:

| Minimum | Średnia | Maksimum | Odchylenie std |
|---|---|---|---|
| 0,998 | 0,999 | 1 | 0,0002 |

Uśredniona macierz pomyłek – próba testowa zawierała 3250 rekordów

| Prawdziwie pozytywne | Prawdziwie negatywne | Fałszywie pozytywne | Fałszywie negatywne |
|---|---|---|---|
| 1682,31 | 1567,63 | 0,06 | 0 |

Można zauważyć, że dla zbioru mushroom osiągnęliśmy nieporównywalnie wyższą dokładność klasyfikacji. Średnia dokładność wynosi praktycznie 1. Może to wynikać ze specyficznego rozłożenia danych w zbiorze mushroom. Jest tam bardzo dużo atrybutów, jednak w procesie debugowania zauważyłem, że drzewa, które są tworzone dla zbioru mushroom są zazwyczaj bardzo płytkie. Może to oznaczać, że pewne atrybuty bardzo dobrze rozdzielają grzyby na trujące i jadalne. Przeprowadziłem poszukiwanie pewnego podzbioru atrybutów dla którego będzie osiągana jak najlepsza dokładność, zbliżona do tej osiąganej przez drzewo stworzone na podstawie wszystkich dostarczonych danych. W wyniku eksperymentów, natrafiłem na podzbiór kolumn o indeksach od 1 do 5. Dla tych pięciu kolumn udało mi się uzyskać bardzo wysoką dokładność. Przeprowadziłem, tak jak dla kompletnych danych, 100 eksperymentów, gdzie dla danych trenujących zostały usunięte wszystkie kolumny, z wyjątkiem tych o indeksach od 1 do 5.

W tabeli poniżej przedstawiam jaką dokładność uzyskałem:

| Minimum | Średnia | Maksimum | Odchylenie std |
|---|---|---|---|
| 0,991 | 0,994 | 0,997 | 0,001 |

Uśredniona macierz pomyłek – próba testowa zawierała 3250 rekordów

| Prawdziwie pozytywne | Prawdziwie negatywne | Fałszywie pozytywne | Fałszywie negatywne |
|---|---|---|---|
| 1679,49 | 1551,33 | 13,84 | 5,34 |

Z przedstawionych danych widać, że wyniki się praktycznie nie zmieniły, pomimo znaczącego ograniczenia ilości analizowanych atrybutów, średnia dokładność spadła bardzo nieznacznie. Oznacza to, że podzbiór pierwszych pięciu atrybutów w zupełności wystarczy do klasyfikowania grzybów czy są trujące, czy też nie. Może to też oznaczać korelację pozostałych atrybutów z pierwszymi pięcioma zbiorami danych, czyli brak sprzeczności w zbiorze badanym.

Zbiór mushroom udało mi się również dobrze klasyfikować nawet dla pojedynczego atrybutu w kolumnie 5 – jest to atrybut odor.

W tabeli poniżej zapisałem uzyskaną dokładność:

| Minimum | Średnia | Maksimum | Odchylenie std |
|---|---|---|---|
| 0,9806 | 0,9854 | 0,9892 | 0,002 |

Uśredniona macierz pomyłek – próba testowa zawierała 3250 rekordów

| Prawdziwie pozytywne | Prawdziwie negatywne | Fałszywie pozytywne | Fałszywie negatywne |
|---|---|---|---|
| 1685,32 | 1517,09 | 47,59 | 0 |

Istnieją również takie podzbiory atrybutów dla których przewidywanie działa znacznie słabiej. Przykładowo dla atrybutów w kolumnach od 6 do 8 zbudowane drzewo osiąga mniejszą dokładność:

| Minimum | Średnia | Maksimum | Odchylenie std |
|---|---|---|---|
| 0,744 | 0,76 | 0,776 | 0,005 |

Uśredniona macierz pomyłek – próba testowa zawierała 3250 rekordów

| Prawdziwie pozytywne | Prawdziwie negatywne | Fałszywie pozytywne | Fałszywie negatywne |
|---|---|---|---|
| 1623,81 | 846,5 | 721,46 | 58,23 |

Można stwierdzić, że nie wszystkie atrybuty w zbiorze mushroom są w stanie dokonywać równie skutecznej predykcji. Można zauważyć znaczny spadek dokładności, zatem wybrane kolumny nie determinują już jednoznacznie czy grzyb jest trujący, czy też jest jadalny.

Analogicznie spróbowałem znaleźć kombinację atrybutów, która lepiej może klasyfikować zbiór breast cancer. Niestety nie znalazłem znacząco lepszego podzbioru niż cały zbiór. Udało się znaleźć takie podzbiory, gdzie była osiągana wyższa dokładność, ale nie była to duża różnica.

W tabeli przedstawiam uzyskaną dokładność:

| Minimum | Średnia | Maksimum | Odchylenie std |
|---|---|---|---|
| 0,617 | 0,691 | 0,765 | 0,03 |

Macierz pomyłek – próba testowa ma 115 elementów:

| Prawdziwie pozytywne | Prawdziwie negatywne | Fałszywie pozytywne | Fałszywie negatywne |
|---|---|---|---|
| 70,64 | 8,78 | 25,91 | 9,67 |

Sprawdziłem, jeszcze jaki wpływ będzie miało ograniczenie liczby wierszy do małej ilości danych uczących jak i trenujących – ograniczyłem rekordy do 5% dla zbioru mushroom, 10% dla breast cancer, które są wybierane losowo z całego zbioru (dane trenujące pozostały w stosunku 3:2 do danych testujących:

-Mushroom

Dokładność:

| Minimum | Średnia | Maksimum | Odchylenie std |
|---|---|---|---|
| 0,963 | 0,99 | 1 | 0,007 |

Macierz pomyłek:

Macierz pomyłek – próba testowa ma 163 elementy:

| Prawdziwie pozytywne | Prawdziwie negatywne | Fałszywie pozytywne | Fałszywie negatywne |
|---|---|---|---|
| 82,87 | 78,81 | 1,06 | 0,26 |

Można zauważyć, że pomimo ograniczenia ilości danych klasyfikator radzi sobie dalej bardzo dobrze.

-Breast cancer

Dokładność:

| Minimum | Średnia | Maksimum | Odchylenie std |
|---|---|---|---|
| 0,25 | 0.62 | 0,917 | 0,14 |

Macierz pomyłek:

Macierz pomyłek – próba testowa ma 12 elementów:

| Prawdziwie pozytywne | Prawdziwie negatywne | Fałszywie pozytywne | Fałszywie negatywne |
|---|---|---|---|
| 6,39 | 1,05 | 2,60 | 1,96 |

Dla klasyfikatora breast cancer wyniki są już mocno rozrzucone, bardziej losowe. Dla najlepszego doboru dobiera osiągnął dokładność bliską 1, w najgorszym przypadku dobrze wskazał co czwarty raz.

## Wnioski

Z przeprowadzonych badań można stwierdzić, że drzewo ID3 jest dobrym klasyfikatorem dla zbiorów danych, w których łatwo można przydzielić elementom odpowiadające klasy. Dla zbioru mushroom osiągnięta została wysoka dokładność. Może to wynikać z faktu, że pewna grupa atrybutów może dokładnie odpowiedzieć na pytanie czy grzyb jest trujący czy nie. Ta grupa atrybutów jednoznacznie determinuje czy grzyb jest jadalny czy trujący. Natomiast dla zbioru breast cancer trudno uzyskać pojedynczym drzewem wyższą dokładność przewidywania. Istnieją elementy w zbiorze o podobnym zestawie atrybutów, a innej klasie, która powinna zostać przewidziana.

Branie tylko części danych ze zbiorów nie powoduje w przypadku zbioru mushroom znaczących zmian w wynikach, natomiast dla zbioru breast cancer, zapewne z racji mniejszej ogólnej liczby rekordów, znaczne obniżenie liczby rekordów.