WSI cw.3, 25Z

Autor: Yan Korzun

4 lutego 2026

1 Cel ćwiczenia

Celem ćwiczenia było zbadanie działania algorytmu drzewa decyzyjnego ID3 w zadaniu klasyfikacji binarnej danych medycznych. W ramach eksperymentu analizowano wpływ maksymalnej głębokości drzewa decyzyjnego na jakość klasyfikacji oraz oceniano, czy wszystkie wykorzystane cechy mają istotny związek z funkcją celu.

2 Przygotowanie danych

Zbiór danych zawierał zarówno cechy ciągłe, jak i dyskretne. Ponieważ algorytm ID3 operuje wyłącznie na atrybutach kategorycznych, cechy ciągłe zostały poddane dyskretyzacji. W tym celu zastosowano funkcję `pd.cut`, która dzieli zakres wartości danej cechy na przedziały o równej szerokości. Po przetworzeniu danych zbiór został podzielony na część treningową, walidacyjną oraz testową.

3 Metoda badawcza

Dla kolejnych wartości maksymalnej głębokości drzewa decyzyjnego od 1 do 8 trenowano model ID3 na zbiorze treningowym. Następnie oceniano jego skuteczność na zbiorze walidacyjnym. Jako miarę jakości klasyfikacji zastosowano współczynnik accuracy.

Na podstawie wyników walidacyjnych wybrano optymalną głębokość drzewa.

4 Wyniki eksperymentu

Uzyskane wartości accuracy dla zbioru walidacyjnego przedstawiono poniżej:

- Głębokość 1: 0.589
- Głębokość 2: 0.630

- Głębokość 3: 0.641
- Głębokość 4: 0.641
- Głębokość 5: 0.638
- Głębokość 6: 0.638
- Głębokość 7: 0.637
- Głębokość 8: 0.636

Najlepszy wynik walidacyjny uzyskano dla maksymalnej głębokości drzewa równej 4. Dla tej konfiguracji dokładność klasyfikacji na zbiorze testowym wyniosła 0.6251.

5 Analiza wyników

Zaobserwowano wyraźny wzrost jakości klasyfikacji wraz ze zwiększaniem maksymalnej głębokości drzewa do poziomu 4. Dalsze zwiększanie głębokości nie prowadziło do istotnej poprawy wyników, a skutkowało ich niewielkim pogorszeniem.

Sugeruje to, że po uwzględnieniu kilku najbardziej informatywnych cech dalsze podziały nie zmniejszają znacząco entropii. Może to wynikać z niskiej istotności części atrybutów, redundancji informacji pomiędzy cechami oraz zastosowanej dyskretyzacji o równej szerokości przedziałów.

Zbliżone wartości accuracy uzyskane dla zbioru walidacyjnego i testowego wskazują na brak silnego przeuczenia modelu.

6 Wnioski

Na podstawie przeprowadzonych badań sformułowano następujące wnioski:

1. Optymalna maksymalna głębokość drzewa decyzyjnego ID3 wynosi 4(dla tego zbioru danych).
2. Zwiększanie głębokości powyżej tej wartości nie poprawia jakości klasyfikacji.
3. Wyniki sugerują, że tylko część cech ma istotny wpływ na decyzje klasyfikacyjne.
4. Jakość klasyfikacji jest ograniczona zarówno przez charakter danych, jak i przez zastosowany algorytm oraz sposób dyskretyzacji.
5. Algorytm ID3 jest wrażliwy na sposób przygotowania danych wejściowych.