Dokumentacja wstępna

Autorzy: Hubert Gołębiowski, Jakub Rozkosz

Opis zadania:

Zmodyfikowany las losowy w zadaniu klasyfikacji. Podczas budowy każdego z drzew testy (z pełnego zbioru testów) wybieramy za pomocą turnieju, czyli wybieramy losowo 2 testy, liczymy ich jakość, stosujemy ten z wyższą jakością. Do wstępnych testów polecam zbiór danych z zadaniem klasyfikacji grzybów, na którym powinno dać się uzyskać wyniki w okolicy 100%. Do dalszych testów należy znaleźć i pobrać inny zbiór danych (na na wskazanej stronie lub w Kaggle).

Opis algorytmów:

Tworzenie drzewa decyzyjnego:

1. Losujemy dwa testy - atrybuty (ze zbioru atrybutów, których jeszcze nie wykorzystaliśmy).
2. Liczymy jakość obu wylosowanych atrybutów. Miarią jakości jest zdobycz informacyjna *InfGain* na zbiorze (zobacz *Obliczanie zdobyczy informacyjnej* poniżej).
3. Wybieramy lepszy atrybut i tworzymy dla niego węzeł decyzyjny, a następnie dzielimy zestaw danych treningowych na mniejsze podzbiory, z których każdy odnosi się do jednego z możliwych wyników węzła.
4. Powtarzamy kroki 1-3 dla każdego podzbioru, tworząc kolejne węzły decyzyjne wzdłuż każdego z atrybutów, aż do osiągnięcia warunku zakończenia. Warunkami zakończenia są:
    - brak atrybutów do podziału
    - wszystkie próbki w danym węźle należą do tej samej klasy
5. Tworzymy liść. Będzie to najliczniejsza klasa w pozostałym zbiorze.

Tworzenie lasu losowego:

Powyższy algorytm tworzenia drzewa powtarzamy *n* razy (n - parametr), aby utworzyć las losowy. Każde z drzew budowane jest na podzbiorze dostępnych danych - podzbiorze uczącym. Predykcja następuje przez agregację wyników - wybieramy najczęstszą klasę.

Obliczanie zdobyczy informacyjnej InfGain:

$$InfGain(d, U) = I(U) - Inf(d, U),$$ gdzie:

$$I(U) = -\sum_i f_i ln(f_i) \qquad Inf(d, U) = \sum_j \frac{|U_j|}{|U|} I(U_j)$$

d - wybrany atrybut

j - wartość atrybutu *d*

U - zestaw danych treningowych

f_i - częstość *i*-tej klasy

U_j - zestaw danych podzielony przez wartość *j* atrybutu *d*

Aby zobrazować przykład weźmy poniższy zestaw treningowy U:

x	*outlook*	*temperature*	*humidity*	*wind*	*play*
1	*sunny*	*hot*	*high*	*normal*	*no*
2	*sunny*	*hot*	*high*	*high*	*no*
3	*overcast*	*hot*	*high*	*normal*	*yes*
4	*rainy*	*mild*	*high*	*normal*	*yes*
5	*rainy*	*cold*	*normal*	*normal*	*yes*
6	*rainy*	*cold*	*normal*	*high*	*no*
7	*overcast*	*cold*	*normal*	*high*	*yes*
8	*sunny*	*mild*	*high*	*normal*	*no*
9	*sunny*	*cold*	*normal*	*normal*	*yes*
10	*rainy*	*mild*	*normal*	*normal*	*yes*
11	*sunny*	*mild*	*normal*	*high*	*yes*
12	*overcast*	*mild*	*high*	*high*	*yes*
13	*overcast*	*hot*	*normal*	*normal*	*yes*
14	*rainy*	*mild*	*high*	*high*	*no*

Obliczmy zdobycz informacyjną dla atrybutu *wind*:

$$I(U) = -\frac{5}{14}ln(\frac{5}{14}) - \frac{9}{14}ln(\frac{9}{14}) \approx 0.651$$

$$I(U_{high}) = -\frac{3}{6}ln(\frac{3}{6}) - \frac{3}{6}ln(\frac{3}{6}) \approx 0.693$$

$$I(U_{normal}) = -\frac{2}{8}ln(\frac{2}{8}) - \frac{6}{8}ln(\frac{6}{8}) \approx 0.562$$

$$Inf(wind, U) = \frac{6}{14} * 0.693 + \frac{8}{14} * 0.562 = 0.618$$

$$InfGain(wind, U) = 0.651 - 0.618 = 0.033$$

Plan eksperymentów:

Dla porównania jakości naszej wersji algorytmu, skorzystamy z klasycznej implementacji lasu losowego znalezionej w Internecie.
Algorytmy zamierzamy przetestować na dwóch zbiorach: agaricus-lepiota i heart failure prediction dataset. Plan testów dla każdego ze zbiorów wygląda następująco:

1. Podział zbioru danych na zbiór testowy i treningowy. Zbiór danych podzielimy w sposób losowy w proporcji 1:4.
2. Stworzenie obywdu lasów losowych (nasz z treści zadania oraz klasyczny) na zbiorze treningowym i przetestowanie ich na zbiorze testowym.
3. Ze względu na losowość algorytmu, krok 2 powtórzymy *k* razy, aby przetestować jak ta losowość wpływa na wyniki.
4. Krok 2-3 powtórzymy dla różnych wartości liczby drzew decyzyjnych, z których będą się składały lasy losowe, aby zbadać wpływ tego parametru na jakość modelu.
5. Kroki 2-4 powtórzymy 3 razy, aby przetestować algorytmy na różnych zbiorach.

Ostatecznie otrzymamy po *k* wyników dla 3 różnych podziałów naszego zbioru. Wartości *m* i *k* są do ustalenia. Wyniki zbierzemy w tabeli i porównamy ich wartości, wartości średnie oraz odchylenia standardowe. Do oceny wykorzystamy następujące miary jakości:

- Tabela pomyłek (TP, TN, FP, FN)
- Dokładność - stosunek poprawnie sklasyfikowanych próbek do całości
- Precyzja - stosunek poprawnie sklasyfikowanych pozytywnych próbek do wszystkich próbek sklasyfikowanych jako pozytywne. Precyzja = TP / (TP + FP)

Wykorzystywane zbiory danych:

- Agaricus-lepiota - docelowo ok. 100% skuteczności
  Zbiór danych zawierających informację o różnych cechach grzybów z rodziny Agaricus i Lepiota (kształt kapelusza, powierzchnię kapelusza, kolor kapelusza, czy grzyb ma "obrączkę" (ang. ring), kolor zarodników, siedlisko i wiele innych). Każdy grzyb zawiera 22 takie atrybuty. Celem jest stworzenie modelu predykcyjnego, który na podstawie cech grzyba, będzie określał czy jest on jadalny czy trujący. Zbiór zawiera 4208 przykładów pozytywnych (jadalne) oraz 3916 negatywnych (trujące)
- Heart Failure Prediction Dataset (ze strony Kaggle)
  Zbiór danych medycznych dotyczących pacjentów z niewydolnością serca. Zawiera on 11 kolumn opisujących różne cechy pacjentów, takie jak wiek, płeć, poziom kreatyniny, poziom sodu we krwi, ciśnienie krwi, palenie tytoniu czy anemia. Ostatnia kolumna zawiera informację, czy pacjent zmarł. Zbiór danych zawiera 918 obserwacji - 508 pozytywnych oraz 410 negatywnych. Celem jest stworzenie modelu predykcyjnego, który na podstawie cech pacjentów będzie przewidywał, czy pacjent ma chorobę serca.

  Jeżeli nasza implementacja na zbiorze “Heart Failure” okaże się mniej skuteczna niż algorytm w klasycznej postaci to postaramy się znaleźć inny zbiór danych (prawdopodobnie z większą liczbą atrybutów), na którym potencjalnie uzyskamy przewagę i powtórzymy testy.