Uczenie Maszynowe

Drzewo decyzyjne w zadaniu klasyfikacji danych z brakującymi wartościami atrybutów.
Dokumentacja Wstępna
cm!

Marcin Latawiec (311031)
Jan Sosnowski (311094)

Politechnika Warszawska

Wydział Elektroniki i Technik Informacyjnych
Politechnika Warszawska
Warszawa
28 stycznia 2023

Spis treści

1 Wstęp

Celem projektu jest zbadanie problemu drzew decyzyjnych w zadaniach klasyfikacji danych z brakującymi wartościami atrybutów. Rozpatrzymy różne modele klasyfikacji i wskażemy ich wady i zalety. Zbadamy jak wybrane zbiory danych są przetwarzane przez każdy z nich. Zbadamy działanie i algorytmy drzew decyzyjnych zarówno z pełnymi danymi jak i z brakami w atrybutach. Przetestujemy i porównamy nasze rozwiązania. Na podstawie uzyskanych wyników postaramy się wskazać rozwiązanie najlepsze z możliwych oraz stwierdzić. Program będziemy pisać w języku Python przy użyciu jak najmniejszej liczby dodatkowych bibliotek.

2 Problem braku danych

W przypadku napotkania braku atrybutu należy wykorzystać co najmniej 1 z poniższych metod:

- Ignoruj pustą wartość
- Wygeneruj losową wartość
- Przypisz do losowego kolejnego węzła na podstawie prawdopodobieństwa zgodnego z rozkładem Bernoulliego Categorical distribution
- Przypisz do największej podgrupy
- Porównaj prawdopodobieństwo jaka będzie klasa gdyby była to dowolna wartość
- Brak atrybutu to też atrybut

3 Typowe algorytmy

Istnieją rozwiązania które skutecznie pozwalają na klasyfikacje. Algorytmy, które uważane są za efektywne przy pracy z takimi zbiorami danych i których użyjemy do budowy węzłów w naszym modelu to:

- ID3
- C4.5
- CRT

Krótki opis powyższych algorytmów zamieszczamy w kolejnej sekcji.

4 Opis algorytmów

4.1 C5.0

Działanie algorytmu *C5.0* polega na podziale próby na podstawie zmiennej oferującej największy zysk informacyjny. Każda podpróba zdefiniowana w wyniku pierwszego podziału jest ponownie dzielona, zwykle na podstawie innej zmiennej, a proces powtarzany jest do momentu, aż podprób nie da się już dalej podzielić. Po podziale podpróby na najniższym poziomie są ponownie analizowane, a te z nich, które nie przyczyniają się istotnie do budowania wartości modelu, są usuwane lub przycinane.
Węzeł *C5.0* może przewidywać tylko zmienną jakościową. Podczas analizowania danych ze zmiennymi jakościowymi (nominalnymi lub porządkowymi) węzeł z większym prawdopodobieństwem będzie grupował kategorie niż węzeł C5.0 w wersjach wcześniejszych niż 11.0.
Węzeł *C5.0* może generować dwa rodzaje modeli. Drzewo decyzyjne jest prostym opisem podziałów znalezionych przez algorytm. Każdy węzeł końcowy („liść") opisuje konkretny podzbiór danych uczących, a każda obserwacja w danych uczących należy do dokładnie jednego węzła końcowego w drzewie. Innymi słowy dla każdego konkretnego rekordu danych odzwierciedlonego w drzewie decyzyjnym możliwa jest dokładnie jedna predykcja.

4.2 CRT

Działanie algorytmu CRT rozpoczyna się od analizy zmiennych wejściowych w poszukiwaniu najlepszych podziałów, przy czym jakość podziału mierzona jest ograniczeniem wskaźnika zanieczyszczenia uzyskanego wskutek podziału. W wyniku podziału powstają dwie podgrupy, z których każda jest następnie dzielona na następne dwie podgrupy i tak dalej, aż do spełnienia kryterium zatrzymania. Wszystkie podziały są binarne (tylko na dwie podgrupy).
W przypadku algorytmu CRT możliwe jest najpierw zbudowanie dużego drzewa, a następnie przycięcie go z zastosowaniem algorytmu analizy kosztu i złożoności, który koryguje oszacowanie ryzyka na podstawie liczby węzłów końcowych. Ta metoda, która umożliwia rozrost drzewa przed przycięciem go na podstawie bardziej złożonych kryteriów, pozwala na uzyskanie mniejszych drzew, które lepiej poddają się walidacji krzyżowej. Zwiększenie ryzyka węzłów końcowych co do zasady zmniejsza ryzyko błędu w odniesieniu do bieżących danych (tj. danych uczących), ale faktyczne ryzyko może być wyższe, gdy model zostanie uogólniony dla danych nieznanych wcześniej. Wyobraźmy sobie skrajny przypadek, w którym dla każdego rekordu w zbiorze uczącym istnieje osobny węzeł końcowy. Oszacowanie ryzyka wyniosłoby 0%, ponieważ każdy rekord ma swój węzeł, ale ryzyko błędnej klasyfikacji na danych nieznanych (testowych) niemal na pewno byłoby większe od 0. Miara kosztu i kompletności jest próbą skompensowania tego zjawiska.
Mocną stroną tego algorytmu jest odporność na brak danych i możliwość operowania na dużej ilości zmiennych.

4.3 ID3

Algorytm ID3 zakłada start z domyślnym węzłem Root - S. Przy każdej iteracji algorytmu sprawdzana jest Entropia $H(S)$ i funkcja maksymalizująca przyrost informacji w każdym rozgałęzieniu (*ang. information gain*) - $IG(S)$. Następnie wybierany jest atrybut z najmniejszą entropią albo największą wartością information gain. Węzeł dzieli się dalej poprzez wybrany atrybut by uzyskać podzbiory danych. Na przykład węzły-child mogą sprawdzać czy średnia wieku danej populacji jest poniżej 50, między 50 a 100 i powyżej 100. Kontynuacja algorytmu polega na kolejnej iteracji na określonym podzbiorze, rozpatrując tylko atrybuty nie wybrane wcześniej.

5 Proponowane zbiory danych

Przy testowaniu wybranych algorytmów proponujemy zastosowanie następujących zbiorów danych:

- próba kontrolna - dane z wszystkimi atrybutami
- zbiór_1 - dane gdzie brakuje 1 atrybutów
- zbiór_2 - dane gzie brakuje 2 atrybutów
- zbiór_3
- zbiór_log10
- zbiór_log2
- zbiór_sqrt
- zbiór_1/2 - dane gdzie brakuje połowy atrybutów

Każdą wygenerowaną próbę kontrolną można wycinać i doprowadzać do konkretnych zbiorów.
Proporcje danych uczących do testujących:

- 1:9
- 5:5
- 8:2
- 9:1
- 99:1

6 Plan Rozwiązania

Proponujemy określony tryb wykorzystania kolejnych algorytmów:

- 1) wygeneruj próbę kontrolną
- 2) dla każdego algorytmu zastosuj każdą metodę radzenia sobie z brakiem atrybutu na każdym zbiorze danych dla każdej proporcji danych uczących:testowych i zapisz skuteczność
- 3) zapisz/wygeneruj podsumowanie dla danej próby kontrolnej
- 4) powtarzaj dla różnych prób kontrolnych (różna liczność zbioru i atrybutów, różna losowość wartości atrybutów

7 Problem nadmiernego dopasowania w przypadku korzystania z drzew decyzyjnych

W przypadku korzystania drzew decyzyjnych należy zdawać sobie sprawę że nadmierne dopasowanie i zbyt duże ograniczanie przestrzeni na podstawie której będą budowane węzły może doprowadzić do nie pożądanego działania drzewa.

8 Zastosowanie regresji liniowej w przypadku ilościowych zbiorów danych

W przypadku liczbowego zbioru danych i brakującej 1 wartości możemy zastosować regresje liniową do wyznaczenia/oszacowania tej wartości.

9 Wyniki

Miarą skuteczności danego algorytmu w praktyce będzie wartość $\alpha = \frac{p}{n}$, gdzie
p - ilość poprawnie sklasyfikowanych przykładów.
n - liczba danych
Im wyższy parametr α tym skuteczniejszy model.

10 Podsumowanie

Mamy nadzieję osiągnąć jak najwyższą skuteczność klasyfikacji oraz potrafić ocenić kiedy brak wartości atrybutu jest mało znaczącym problemem.