Indukcyjne Metody Analizy Danych

laboratorium

Ćwiczenie 2. Indukcja drzew decyzyjnych C4.5 w R

opracował: P.B. Myszkowski * data aktualizacji: 17.03.2018

Cel ćwiczenia

Zapoznanie się z indukcją drzew decyzyjnych na przykładzie C4.5 na platformie w R.

Realizacja ćwiczenia

•Zapoznanie się z metodą reprezentacji wiedzy, jaką są drzewa decyzyjne
•Zapoznanie się z klasyczną implementacją indukcji drzew decyzyjnych – C4.5
•Zapoznanie się z platformą R
•Przebadanie działania metody C4.5 na wybranych zbiorach
•Przebadanie wpływu parametrów na skuteczność/efektywność metody (rozmiar drzewa a miary jakości klasyfikatora Acc,Rec,Prec,Fsc)
•Sporządzenie sprawozdania z ćwiczenia

Informacje pomocnicze

Proces pozyskiwania wiedzy z baz danych (KDD, *knowledge discovery in databases*) jest jednym z ważniejszych zastosowań metod sztucznej inteligencji. W tym procesie najbardziej interesuje nas etap data mining (drążenie danych), które zawężamy do zadania klasyfikacji. Dodatkowo skupiamy się tylko na jednej konkretnej metodzie, jaką jest indukcja drzew decyzyjnych C4.5.

Zadanie polega na zbudowaniu modelu (klasyfikatora), który na podstawie zbiorów danych (zbiory: wine, glass, pima diabeters + iris do wstępnej weryfikacji) próbuje stworzyć skuteczny klasyfikator oparty przy pomocy C4.5. Dodatkowo interesuje nas przebadanie wpływu parametrów algorytmu uczenia się klasyfikatora C4.5 na końcowy wynik.

Pośrednim parametrem jest kroswalidacja i jej rodzaj (zwykła i stratyfikowana)

Ocena ćwiczenia (max 10pkt)

1pkt	Przeanalizowanie danych badawczych
1pkt	Zapoznanie się z platformą R
1pkt	Krótki opis działania algorytmu C4.5
2pkt	Zbadanie działania algorytmu C4.5 na 3 zbiorach
2pkt	Porównanie działania algorytmu przy różnych danych i wartościach parametrów – tabelki, wnioski
1pkt	Porównanie wyników działania algorytmu – graficzne przestawienie uzyskanych wyników. Wykresy i rysowanie drzew.
1pkt	Przebadanie wpływu rozmiaru krosswalidacji „zwykłej” i stratyfikowanej
1pkt	Porównanie najlepszych wyników do najlepszych wyników działania bayesaNaiwnego

Pytania pomocnicze

1. Jakie parametry ma C4.5? Jakie mają wartości są „optymalne”?
2. Jak rozmiar/typ krosswalidacji wpływa na jakość końcowego klasyfikatora?
3. Czy C4.5 potrzebuje dyskretyzacji?
4. Czy drzewo może być zbyt małe lub za duże?
5. Jaką rolę pełni przycinanie drzewa i czy zawsze jest potrzebne?

Literatura

1.Wykłady do przedmiotu autorstwa prof. H.Kwaśnickiej
2.Cichosz P. ”Systemy uczące się”, WNT Warszawa
3.Eksploracja danych (seria wykładów) http://wazniak.mimuw.edu.pl/index.php?title=Eksploracja_danych
[Szczególnie interesują nas wykłady 7-9]
4.Klasyfikacja: http://wazniak.mimuw.edu.pl/index.php?title=Sztuczna_inteligencja/SI_Modu%C5%82_10_-_Zadanie_i_metody_klasyfikacji
5.Zasoby Internetu: uczenie maszynowe (machine learning), data mining, klasyfikacja

Linki odnośnie samego R

http://books.goalkicker.com/RBook/ →
http://www.biecek.pl/R/ → start z R
https://cran.r-project.org/web/packages/C50/C50.pdf
https://www.rdocumentation.org/