Indukcyjne Metody Analizy Danych
laboratorium

Ćwiczenie 4. Algorytm klasyfikacji *k*-najbliższych sąsiadów

opracował: P.B.Myszkowski * data aktualizacji: 27.04.2018

Cel ćwiczenia
Zapoznanie się z metodą klasyfikacji *k*-najbliższych sąsiadów (knn, *k-nearest neighbours*) na przy samodzielnej implementacji. Do wyboru studenta narzędzie realizacji: R lub python.

Realizacja ćwiczenia

•Zapoznanie się z uczeniem „leniwym” w zastosowaniu do zadania klasyfikacji
•Wybór 5 zbiorów danych (powinny być te same co na poprzednich zajęciach + iris)
•Przebadanie działania metody *KNN* na wyżej wymienionych zbiorach
•Przebadanie wpływu zbioru/atrybutów/wartości na skuteczność/efektywność metody
•Zbadanie wpływy postaci miary odległości, sposobów głosowania i liczby *k*
•Porównanie z wynikami klasyfikatorów z poprzednich zadań
•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 klasyfikacji *knn*, jako przykładu metody uczenia leniwego.

Zadanie polega na implementacji metody *knn*, który na podstawie zbiorów danych zbuduje „model” klasyfikacji. Ćwiczenie zakłada pracę z 5 zbiorami.

Zakłada się przebadania kilku wartości *k* (sąsiadów), 3 różne sposoby głosowania (np. większościowe równoprawne, ważone odległością itp. – można inne) oraz minimum 2 sposoby definicji miary odległości (np. Euklides, manhattan, minkowski, Mahalanobis’a, Czybyszewa)

Kroswalidacja (walidacja krzyżowa, crossvalidation) jest to sposób na podział danych na zbiory uczące i testowe w taki sposób, aby zminimalizować ich wpływ na obserwowane wyniki. Zwykle stosuje się kroswalidacje 2, 3 lub 5, rzadziej 10. To dla przypomnienia – tym razem też się przyda. Proszę o rozważenie użycia kroswalidacji stratyfikowanej.

Przydatne linki

•materiały z wykładu prof.Kwaśnickiej
•dane warto użyć z poprzedniego ćwiczenia (UCL Repository) - http://archive.ics.uci.edu/ml/
Trzeba pamiętać o dostosowaniu danych do algorytmu.
• http://wazniak.mimuw.edu.pl/index.php?title=ED-4.2-m09-1.0-toc
•http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.31.1422
•(ostatecznie) http://pl.wikipedia.org/wiki/Algorytm_k_najbli%C5%BCszych_s%C4%85siad%C3%B3w

Ocena ćwiczenia (max 10pkt)

2pkt	Implementacja algorytmu *knn*
1pkt	Krótki opis działania algorytmu *knn* wraz z sposobami głosowania (min 3) i różnych miar
2pkt	Zbadanie wypływu na skuteczność *knn* sposobu liczenia odległości – dla 5 zbiorów
2pkt	Zbadanie wpływu sposobu głosowania w *knn* (5 zbiorów)
2pkt	Porównanie działania algorytmu przy różnych podziałach danych (kroswalidacja!) – tabelki, wnioski
1pkt	Porównanie działania algorytmu – graficzne przestawienie uzyskanych wyników (porównanie z innymi, wcześniej przebadanymi algorytmami) – bayes, C4.5

Literatura

1.Materiały prof. Kwaśnickiej

2.Cichosz P. ”Systemy uczące się”, WNT Warszawa

3.Zasoby Internetu: uczenie maszynowe (*machine learning*), *data mining*

W razie pytań, uwag czy komentarzy proszę o kontakt z prowadzącym zajęcia.
P. Myszkowski