Indukcyjne Metody Analizy Danych
laboratorium

Ćwiczenie 5. Zespoły klasyfikatorów

opracował: P.B.Myszkowski * data aktualizacji: 22.05.2018

Cel ćwiczenia
Zapoznanie się z trzema metodami tworzenia zespołów klasyfikatorów przy samodzielnej implementacji w R lub python

Realizacja ćwiczenia

• Zapoznanie się z metodami tworzenia zespołów klasyfikatorów: *bagging*, *boosting*
• Zapoznanie się z metodą *RandomForest*
• Wybór trzech zbiorów danych (mogą być te same co na poprzednich zajęciach)
• Wybór jednej metody klasyfikacji (**nie** *k-nn*) z wcześniejszych laboratorium – można użyć innej, niż były na zajęciach, ale trzeba zbadać właściwości „podstawowego" algorytmu.
• Przebadanie wpływu zbioru/atrybutów/wartości na skuteczność/efektywność metody
• 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. Przy tym zadaniu tworzenie zespołów klasyfikatorów skupia się na zwiększeniu skuteczności klasyfikacji zespołu. Pod uwagę należy wziąć nie tylko sposób tworzenia klasyfikatorów, ale także sposoby *łączenia* i *głosowania* w celu obliczenia wyniku dla (meta)klasyfikatora.

Zadanie polega poznaniu trzech metod tworzenia zbioru klasyfikatorów, który na podstawie zbiorów danych zbuduje „model" klasyfikatora. Ćwiczenie zakłada pracę z minimum trzema zbiorami danych.

Przy *RandomForest* mamy „narzucony" algorytm, ale można wybrać inny klasyfikator.

Tendencyjne pytania:
0. Dlaczego użycie *knn* przy zespole klasyfikatorów nie jest najlepszym pomysłem?
1. Czy jest potrzebna kroswalidacja?
2. Która metod jest „lepsza"?
3. Jak parametry metod wpływają na uzyskiwane wyniki? Pamiętamy, że używamy FSC, a ACC tylko jako pomocnicze.

Przydatne linki

•dane można użyć z poprzedniego ćwiczenia (UCL Repository) - http://archive.ics.uci.edu/ml/

• http://www.ke.tu-darmstadt.de/lehre/archiv/ws0405/mldm/ensembles.pdf

...
• http://google.com

Ocena ćwiczenia (max 10pkt)

1pkt	Implementacja obu metod tworzenia zbioru klasyfikatorów
1pkt	Krótki opis działania dwóch metod tworzenia zbioru klasyfikatorów
2pkt	Zbadanie wpływu parametrów na skuteczność metody *bagging*
2pkt	Zbadanie wpływu parametrów na skuteczność metody *boosting*
2pkt	Zbadanie wpływu parametrów na skuteczność metody *RandomForest*
1pkt	Porównanie działania metod – tabelki, wnioski
1pkt	Porównanie wyników działania metod – graficzne przestawienie uzyskanych wyników (porównanie z innymi, wcześniej przebadanymi algorytmami)

Literatura

1. materiały z wykładów prof. Kwaśnickiej

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

2. Koronacki J., Ćwik J., „Statystyczne systemy uczące się”, WNT Warszawa
=> szczególnie polecana!

3. Zasoby Internetu: uczenie maszynowe (*machine learning*), *data mining*, zespoły klasyfikatorów, *ensemble classifiers*, *boosting*, *baggin*, *random Forest*