Projekt 3 – data realizacji: 18/01/2010

Metody ekstrakcji i selekcji cech

Cel projektu

Zapoznanie się z metodami wyboru najistotniejszych cech dla problemu, oraz generowania niewielkiej liczby reprezentatywnych cech na potrzeby wizualizacji danych.

Dane wejściowe

Zbiory danych – należy wybrać 1 zbiór dwuklasowy ze zbiorów dostępnych na stronie http://archive.ics.uci.edu/ml/., np. jeden z wykorzystanych w ramach Projektu 2.

Opis projektu

1. Przygotowanie wybranego zbioru danych: podzielić na część treningową (75%) i testową (25%), oznaczane niżej jako Tr75 i Te25.

2. Ekstrakcja i selekcja cech na potrzeby klasyfikacji:
   a. na zbiorze Tr75 należy wyznaczyć macierz przekształcenia liniowego do nowych cech metodą PCA. Następnie przy pomocy tej macierzy należy przekształcić przykłady ze zbiorów Tr75 i Te25 do nowej przestrzeni.
   b. wyucz klasyfikator AdaBoost na przekształconym zbiorze Tr75, wykorzystując pierwsze 5 komponentów głównych. Dla przykładów ze zbioru Te25, porównaj skuteczność klasyfikacji (błąd + pole pod krzywą ROC) tak wyuczonym klasyfikatorem z klasyfikatorem AdaBoost uczonym na pierwotnych cechach.
   c. Na zbiorze Tr75 oblicz wartość pola pod krzywą ROC dla każdej cechy osobno:
   (patrz http://www.icsr.agh.edu.pl/~mro/proj3/FeatSel1.ps).
   Sporządź ranking cech od tej o największym polu pod krzywą ROC do tej o najmniejszym. Wyucz klasyfikator AdaBoost dla podzbioru 3, 5, 10 i 20 najlepszych cech i porównaj z wynikami dla klasyfikatora wykorzystującego wszystkie cechy.

3. Ekstrakcja cech na potrzeby wizualizacji danych wielowymiarowych
   a. Dokonaj klasteryzacji aglomeratywnej zbioru danych, oraz jego klasyfikacji metodą AdaBoost, wykorzystując oryginalne cechy.

b. Dokonaj transformacji zbioru metodą PCA. Wykorzystaj pierwsze 2 oraz pierwsze 3 komponenty główne do zwizualizowania w 2D/3D zbioru danych.

c. Wykorzystaj metodę MDS z kryterium Sammona by uzyskać nieliniową wizualizację 2D/3D zbioru danych.

d. Przeanalizuj wynik klasteryzacji na wizualizacjach PCA i MDS. Zaznacz punkty z każdego klastra osobnym kolorem. Zaobserwuj, czy uzyskane klastry są widoczne, i czy przykłady w ramach klastrów tworzą wyróżnione skupiska punktów.

e. Przeanalizuj zagadnienie klasyfikacyjne metodami PCA i MDS, dokonując wizualizacji wszystkich przykładów ze zbioru. Wykorzystaj różne kolory do zaznaczenia różnych klas, oraz do zaznaczenia czy przykład jest ze zbioru treningowego czy testowego. Sprawdź na ile przykłady testowe danej klasy zajmują podobne rejony przestrzeni co przykłady treningowe.

f. Przeanalizuje wynik klasyfikacji na zbiorze Te25 przy użyciu wizualizacji PCA i MDS, wyróżniając klasy różnymi kolorami. Zaobserwuj, czy klasy definiujące zagadnienie klasyfikacyjne są geometrycznie wyróżnione. W ramach klas, rożnymi kolorami zaznacz punkty dla których klasyfikator dał poprawną/błędną odpowiedź, i przeanalizuj jak geometria tych przykładów współgra z poprawnością klasyfikacji.

Przydatne funkcje

- Implementacja drzew decyzyjnych - funkcje treefit, treeval, treeprune lub funkcja classregtree i inne podane w helpie do niej
- *linkage, cluster, pdist* – klasteryzacja aglomeratywna
- *mdscale* – MDS
- *princomp* – PCA
- *scatter, scatter3* – wizualizacja punktów różnymi kolorami