Matuszek Albert 16.01.2010

Krupczak Łukasz

Sprawozdanie z laboratorium z przedmiotu

Metody Rozpoznawania Obrazów

Temat: **Metody ekstrakcji i selekcji cech**

# Przygotowanie zbioru danych

Jako zbiór danych przyjęto zbiór Cancer ze strony <http://archive.ics.uci.edu/ml/.> Jest to zbiór danych składający się z 569 elementów opisujący raka, który pozwala sprawnie klasyfikować typ raka jako złośliwy lub zwykły.

Opis charakterystyki stanowi 9 cech na które się składa 30 atrybutów (zakres liczbowy nieciągły [0,100]), cechy:

* Clump Thickness
* Uniformity of Cell Size
* Uniformity of Cell Shape
* Marginal Adhesion
* Single Epithelial Cell Size
* Bare Nuclei
* Bland Chromatin
* Normal Nucleoli
* Mitoses

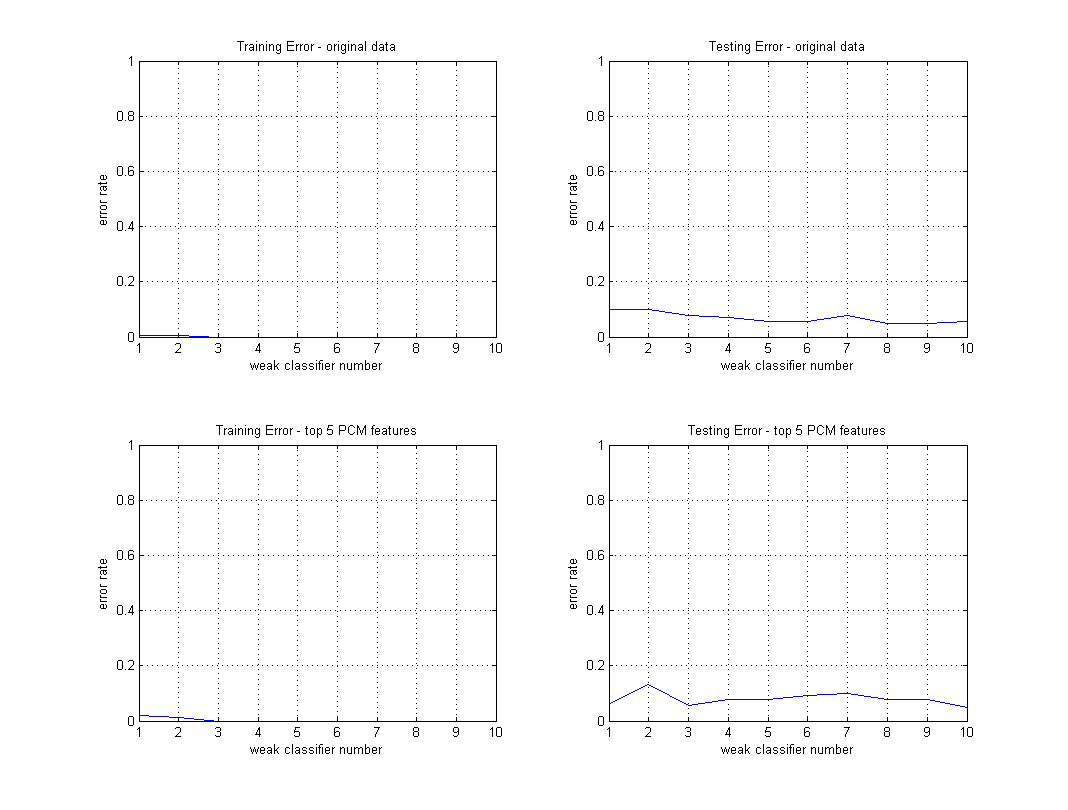
Pozwoliliśmy sobie nie tłumaczyć ich znaczeń ze względu na zawiłości medycznej terminologii. Cechy te zostały obliczone na podstawie cyfrowych zdjęć rentgenowskich (FNA). Opisują one właściwości struktury zwyrodnieniowej (rakowej).

Zbiór został podzielony na 2 podzbiory:

* Tr75 – 426 elementów - treningowy
* Te25 – 143 elementów - testowy

# Ekstrakcja i selekcja cech na potrzeby klasyfikacji

* 1. Na zbiorze TR75 wyznaczono macierz przekształcenia liniowego do nowych cech metodą PCA, a następnie przy jej pomocy przekształcono zbiory Tr75 i Te25 do nowej przestrzeni.
  2. Wyuczono klasyfikator AdaBoost na nowych(wykorzystujących pierwsze 5 komponentów głównych) i pierwotnych cechach oraz porównano skuteczność klasyfikacji dla obydwu zbiorów. Wyniki przedstawiono na poniższych wykresach:

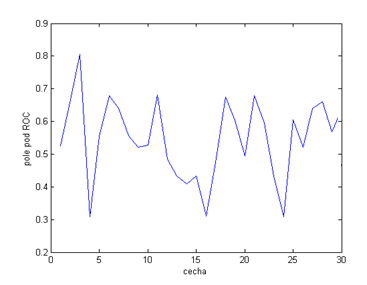
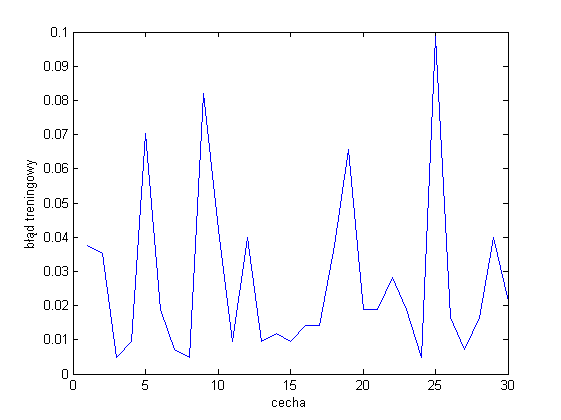


Ostatecznie po 10 iteracjach osiągnięto następujące wyniki:

* dla klasyfikacji na pierwotnych cechach:
  + błąd: 0.0559
  + pole pod krzywą ROC: 0.1296
* dla klasyfikacji na przekształconych danych:
  + błąd: 0.0490
  + pole pod krzywą ROC: 0.2642

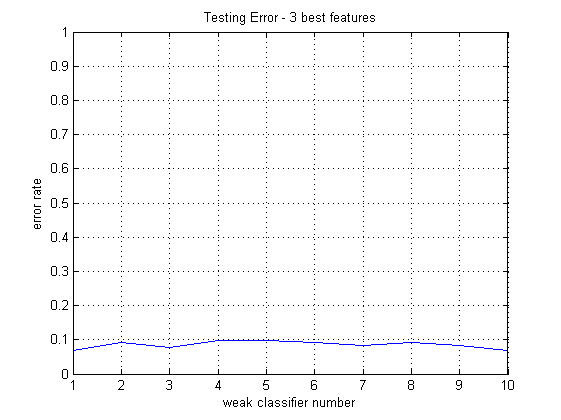
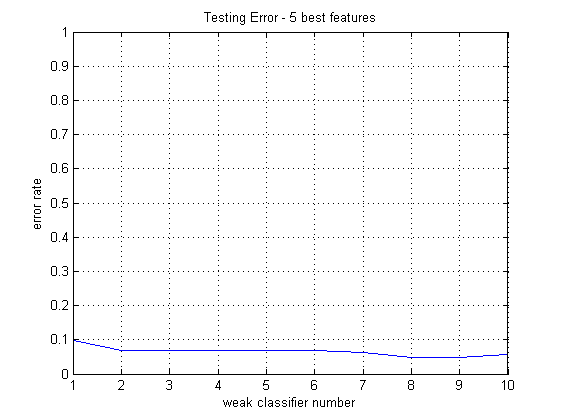
Jak wynika z powyższych wyników wykorzystując przekształcony zbiór uzyskano spadek błędu a także zwiększyło się pole pod krzywą ROC, co świadczy o poprawie jakości klasyfikacji.

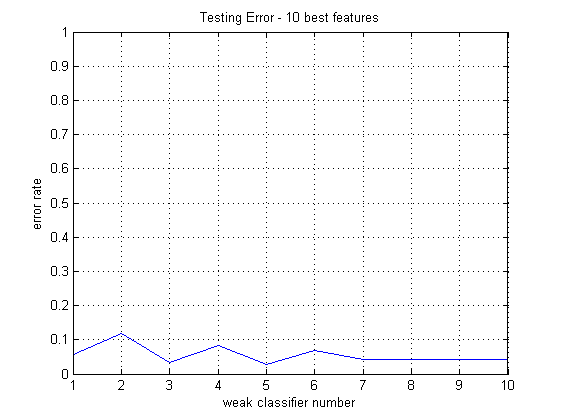
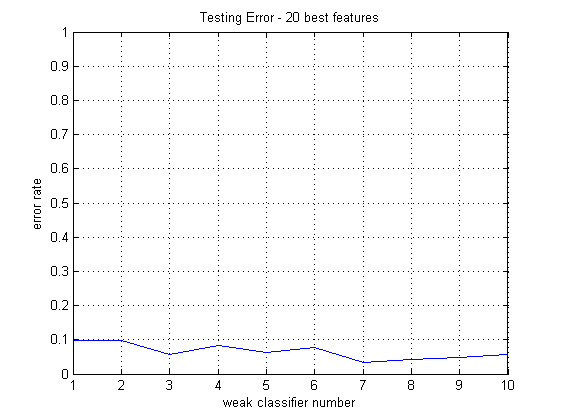
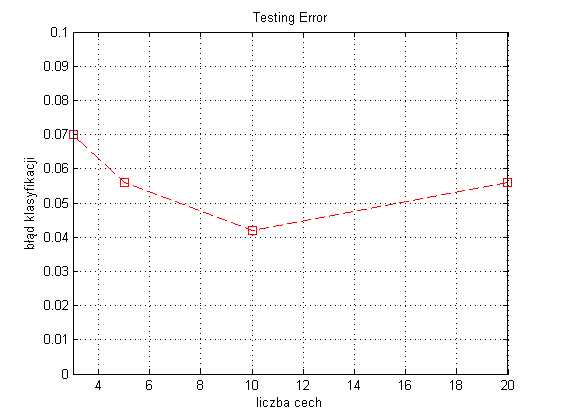
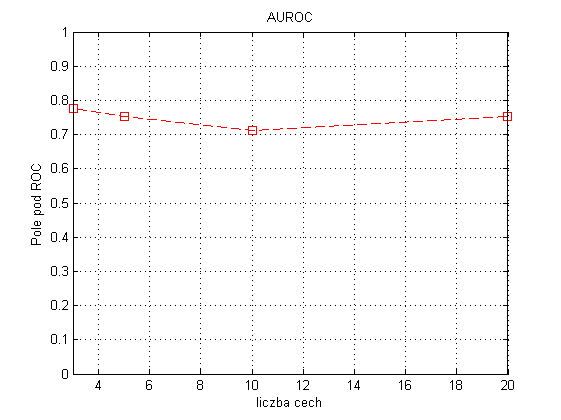
* 1. Wykorzystując każdą z cech z osobna dla zbioru Tr75 wyznaczono pole pod krzywą ROC oraz błąd treningowy:



Jak widać niektóre cechy mają większy wpływ na klasyfikacje od innych (np. odp. 25 i 24).

Następnie sporządzono ranking 3, 5, 10 i 20 najlepszych cech oraz porównano je z wynikami z pkt. 2.2 gdzie wykorzystane były wszystkie cechy:



Ranking:

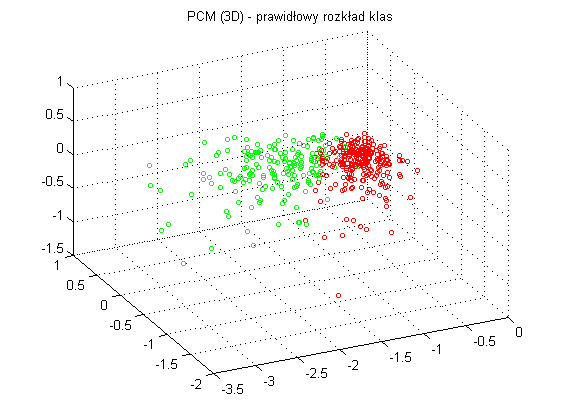
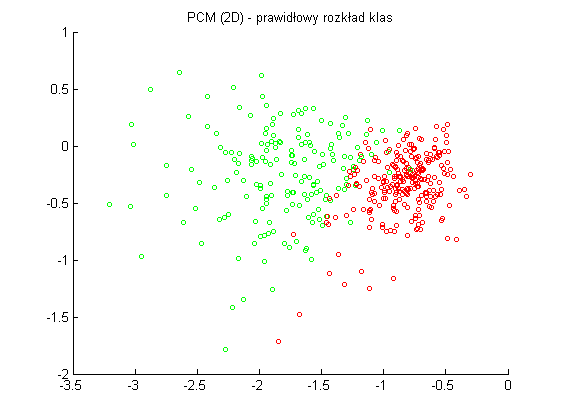
|  |  |  |
| --- | --- | --- |
| Ilość najlepszych cech: | ROC | błąd |
| 3 | 0.2241 | 0.0699 |
| 5 | 0.2472 | 0.0559 |
| 10 | 0.2879 | 0.0420 |
| 20 | 0.2472 | 0.0559 |
| 30 (wszystkie) | 0.1296 | 0.0559 |

Jak widać na powyższych wykresach oraz tabeli już dla 5 najlepszych cech uzyskuje się błąd klasyfikacji równy błędowi klasyfikacji przy wykorzystaniu wszystkich cech. Najlepszy wynik uzyskuje się dla 10 najlepszych cech (zarówno błąd jak i pole pod ROC). Wynika to z cechy zbioru polegającej na tym że pewien podzbiór cech ma decydujący wpływ na klasyfikacje, a reszta jest mało znacząca. Odwzorowuje to pole pod krzywą ROC dla klasyfikacji wykorzystującej wszystkie atrybuty, pozostałe 10 cech ma bardzo mały wpływ na wynik oraz dodatkowo zwiększa błąd klasyfikacji.

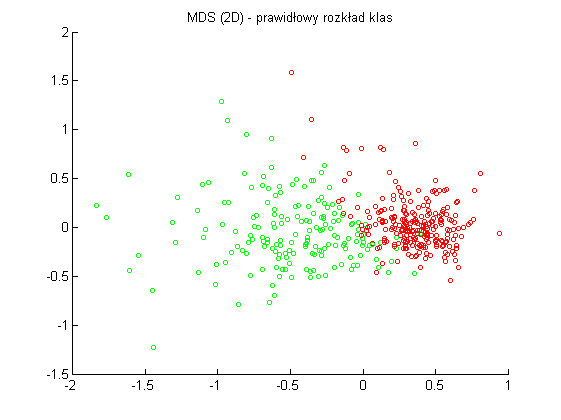
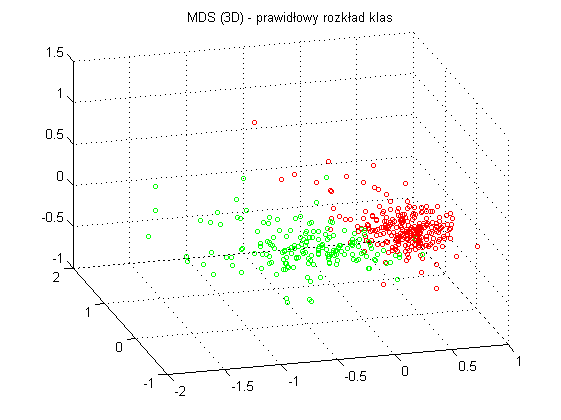
# Ekstrakcja cech na potrzeby wizualizacji danych wielowymiarowych

|  |  |  |
| --- | --- | --- |
| Metoda | Poprawnie przydzielone | Trafień |
| Klasteryzacja metodą ‘*complete’* | 192 | 34% |
| Klasyfikacja metdą AdaBoost | 537 | 94%% |

* 1. Zbór przetransformowano wykorzystując metodę PCA oraz zwizualizowano podział na klasy dla 2 i 3 komponentów głównych:

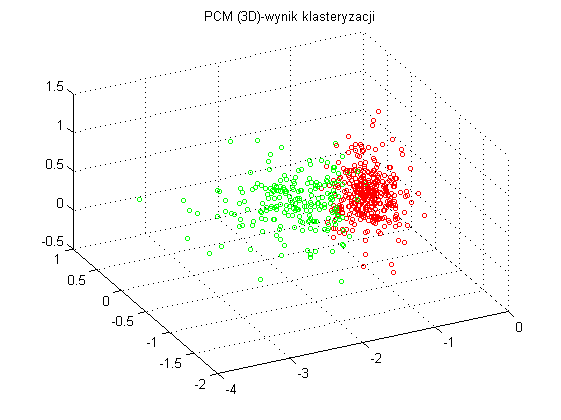


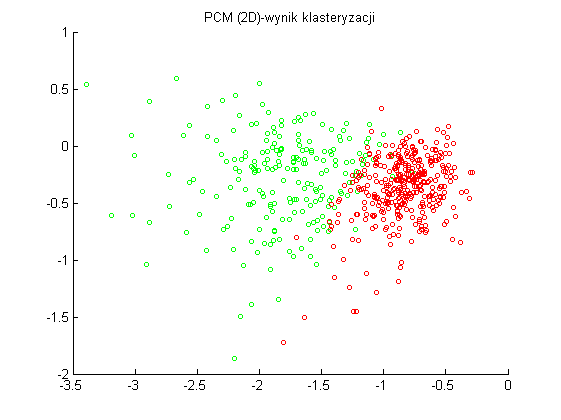
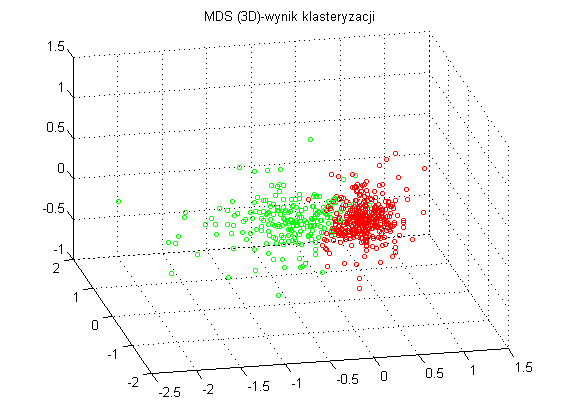
* 1. Nieliniowa wizualizacja 2D/3D (za pomocą metody MDS z kryterium Sammona) (kolory zgodne z oryginalnymi cechami):

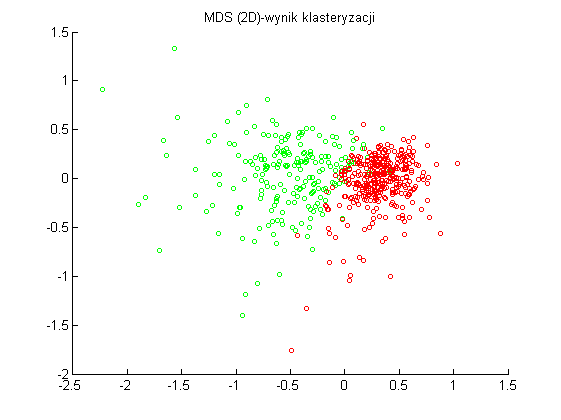


* 1. W obu przypadkach (PCM I MDS) klasyfikacja przebiegła poprawnie, a podział na klastry jest widoczny - klastry tworzą wyróżnione skupiska o jednolitym kolorze.

|  |  |  |
| --- | --- | --- |
| Metoda | Trafień | Trafień [%] |
| PCM | 368 | 64,67% |
| MDS | 500 | 87,87% |

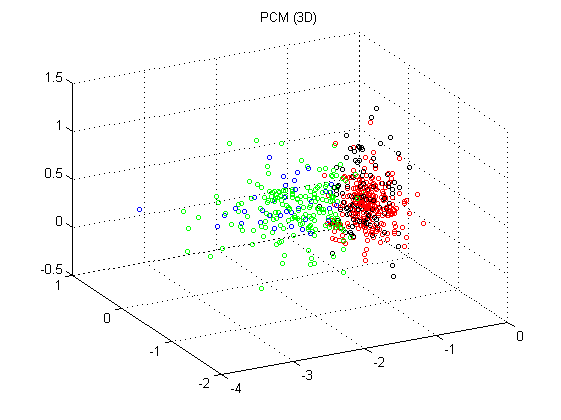
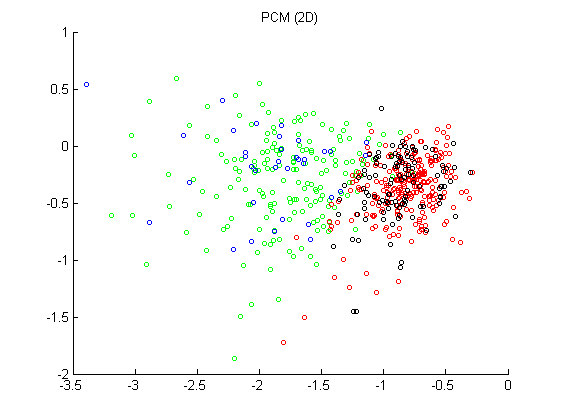


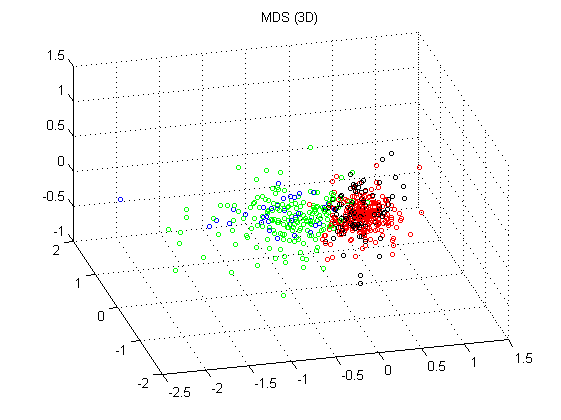


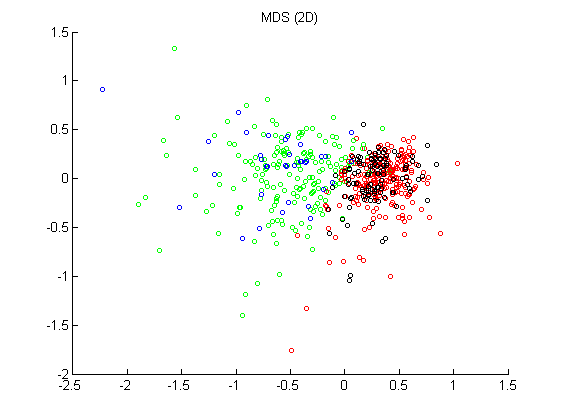


* 1. Analiza klasyfikacji metodami PCM i MDS z wykorzystaniem wszystkich przykładów ze zbioru.

Podział:

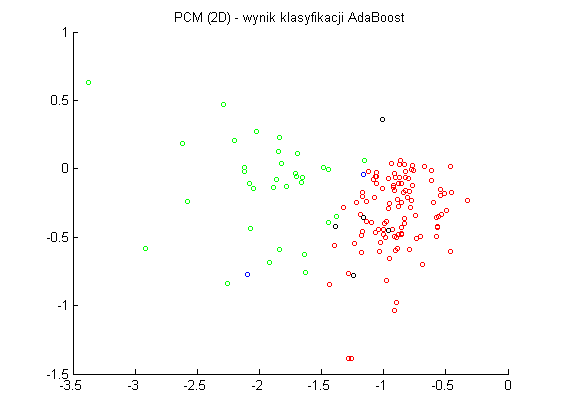
* Zbiór treningowy
  + Klasa 1 – zielony
  + Klasa 2 – czerwony
* Zbiór testowy
  + Klasa 1 – niebieski
  + Klasa 2 – czarny

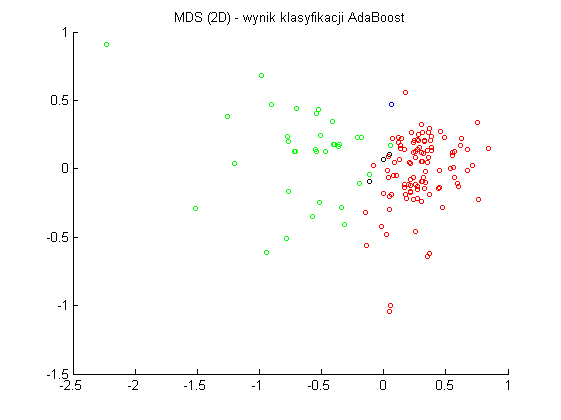
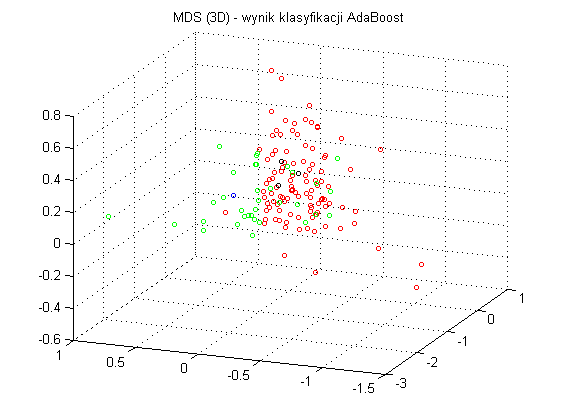


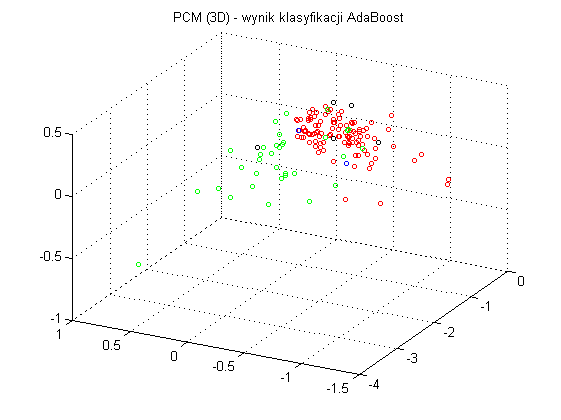


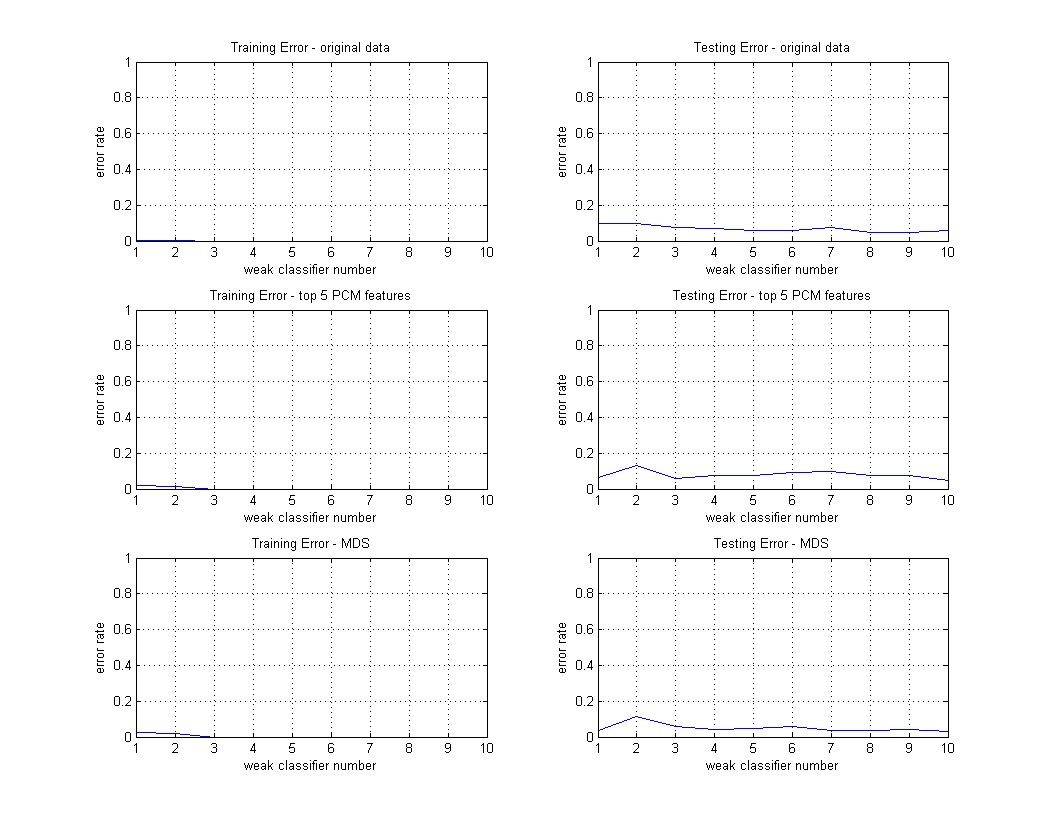
Dla obu metod zarówno przykłady treningowe jak i testowe z tych samych klas zajmują podobne regiony na wizualizacjach, co tłumaczy fakt, że niski błąd treningowy przenosi się na niski błąd testowy dla prezentowanych danych. Tego typu dane są łatwe do sklasyfikowania.

* 1. Poniżej widzimy wynik klasyfikacji zbioru Te25 przy użyciu wizualizacji PCM i MDS. Poszczególne klasy są zaznaczone różnymi kolorami. Na wizualizacji tych można zauważyć że poszczególne klasy są geometrycznie wyróżnione, a granica klas jest jednoznaczna i widoczna. Ilość błędnie sklasyfikowanych punktów jest znikoma i w obrębie bliskiego sąsiedztwa granicy klas, wynika to z niepewności klasyfikacji w warunkach brzegowych.

Punkty niebieskie są to źle zakwalifikowane obiekty należące do klastra 1 (zielonego) natomiast czarne to źle zakwalifikowane pkt. klastra nr 2.





Na poniższych wykresach błędów klasyfikacji widzimy, że metoda MDS daje wyniki lepsze pod PCM jedynie w nieznacznym stopniu (mała różnica błędów klasyfikacji). Ogólny błąd klasyfikacji dla obu metod jest bardzo niski, co mogliśmy zaobserwować na powyższych wizualizacjach(mała liczba pkt. niebieskich i czarnych). Na wizualizacji można także zauważyć różnice, w jakości klasyfikacji PCM i MDS (mniej pkt. niebieskich i czarnych dla MDS), co potwierdzają poniższe wykresy.