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 atrybutów (zakres liczbowy nieciągły [1,10]):

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

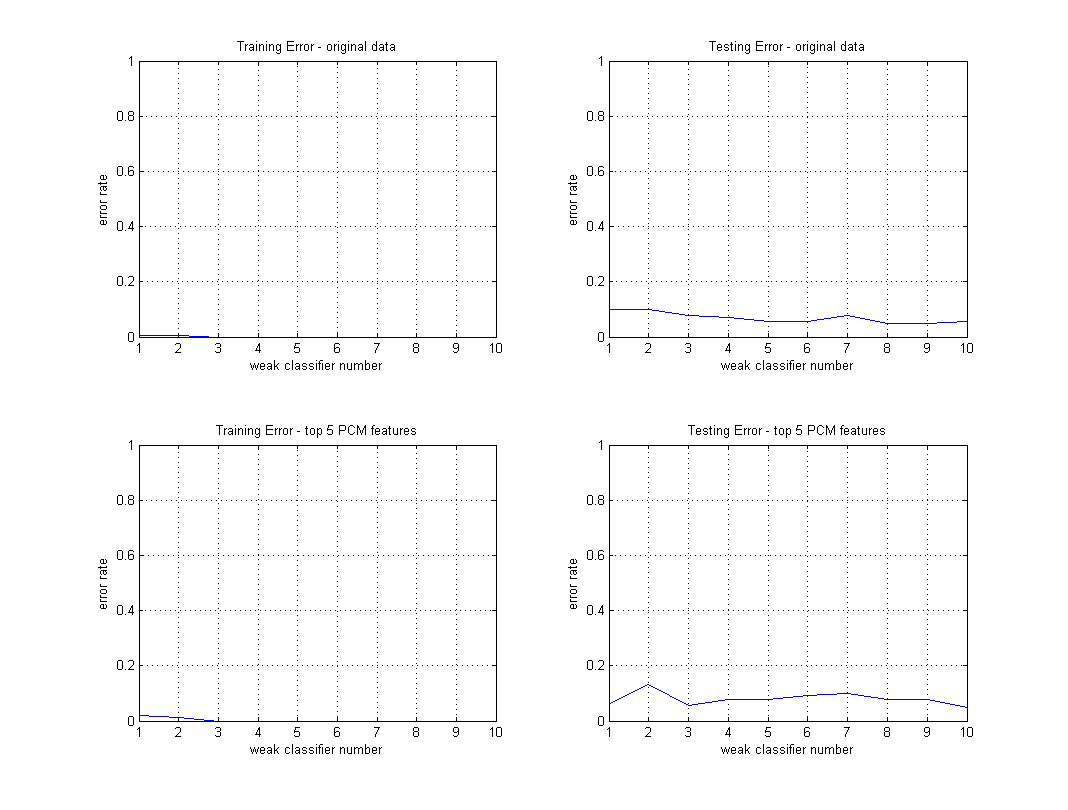
Pozwoliliśmy sobie nie tłumaczyć ich znaczeń ze względu na zawiłości medycznej terminologii. Cechy te zostały obliczone na podstawie cyfrowych zdjęć rentgenowskich (FNA). Opisują one właściwości struktury zwyrodnieniowej (rakowej).

Zbiór został podzielony na 4 podzbiory:

* TR75 – 426 elementów - treningowy
* TR50 – 283 elementów – treningowy, bez informacji o klasach
* TR25 – 143 elementów - treningowy
* TE25 – 143 elementów - testowy

# Ekstrakcja i selekcja cech na potrzeby klasyfikacji

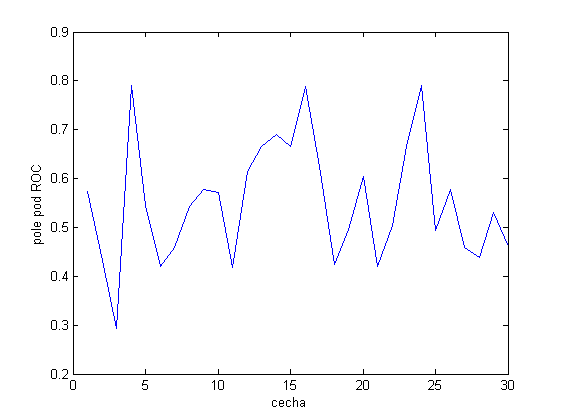
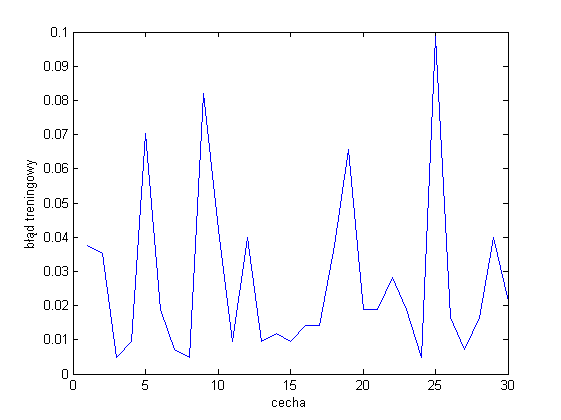
Na zbiorze TR75 wyznaczono macierz przekształcenia liniowego do nowych cech metodą PCA. Następnie wyuczono klasyfikator AdaBoost na nowych(wykorzystujących pierwsze 5 komponentów głownych) i pierwotnych cechach oraz porównano skuteczność klasyfikacji dla obydwu zbiorów. Wyniki przedstawiono na poniższych wykresach:



Ostatecznie (w 10 iteracji) osiągnięto:

* dla klasyfikacji na pierwotnych cechach:
  + błąd 0.0559
  + pole pod krzywą ROC: 0.8704
* dla klasyfikacji na przekształconych danych:
  + błąd: 0.0490
  + pole pod krzywą ROC: 0.7358

pkt c:

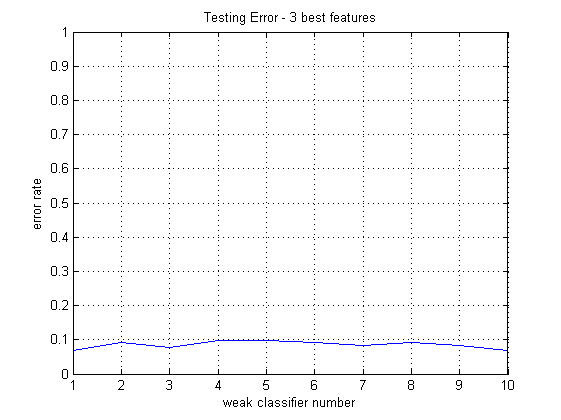


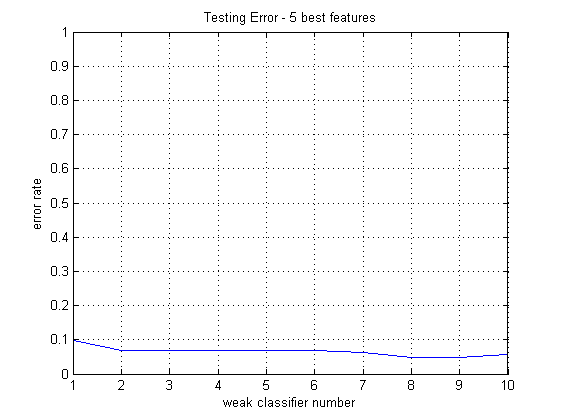
pole pod roc:

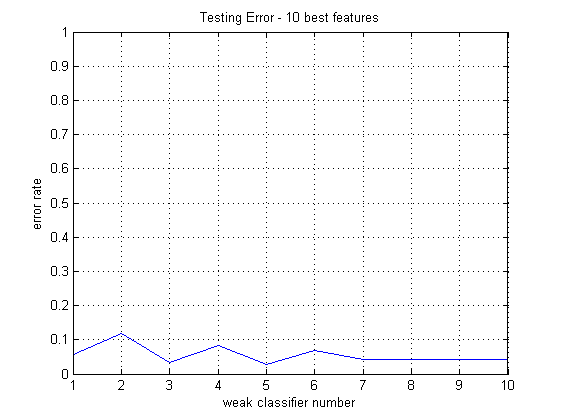
|  |  |
| --- | --- |
| Liczba cech |  |
| 3 | 0.7759 |
| 5 | 0.7528 |
| 10 | 0.7129 |
| 21 | 0.7528 |

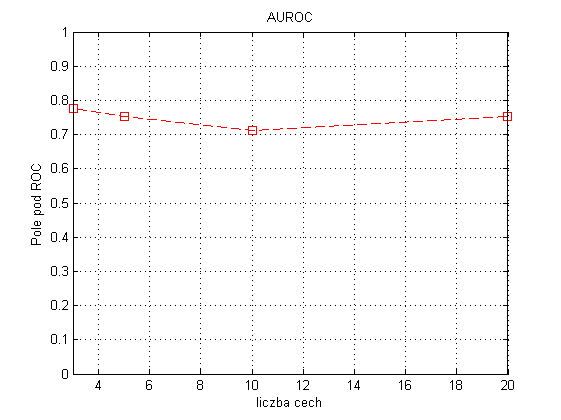
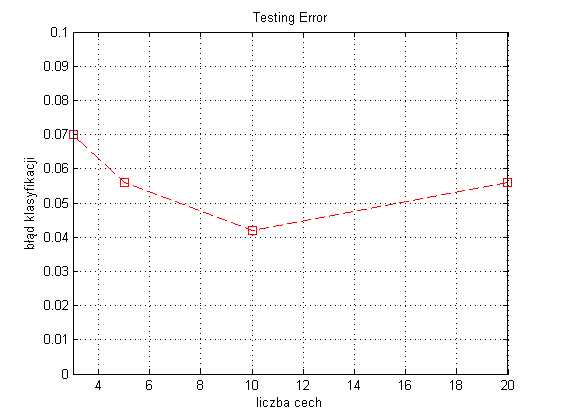
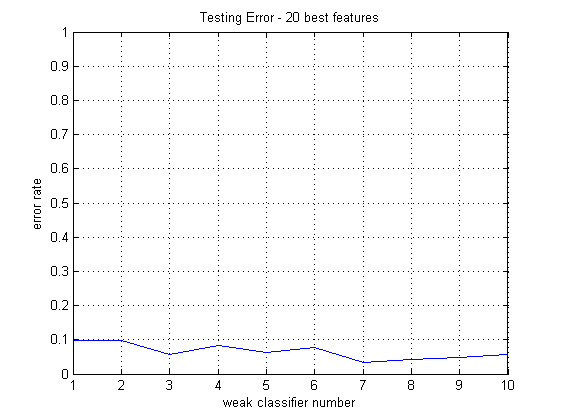
bład:

|  |  |
| --- | --- |
| Liczba cech |  |
| 3 | 0.0699 |
| 5 | 0.0559 |
| 10 | 0.0420 |
| 21 | 0.0559 |









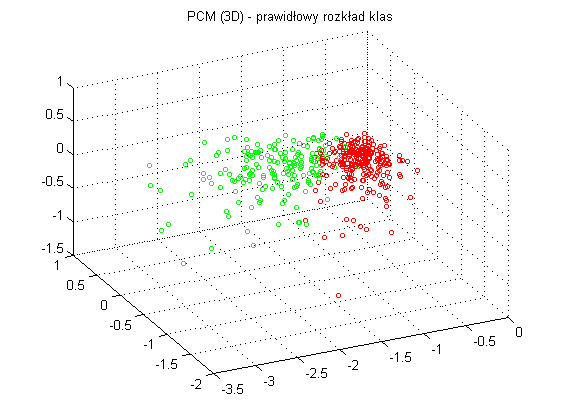
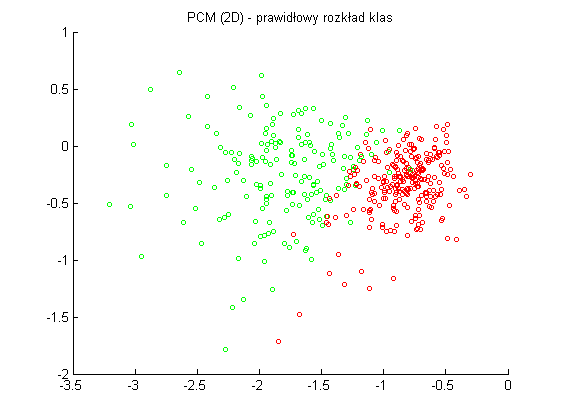
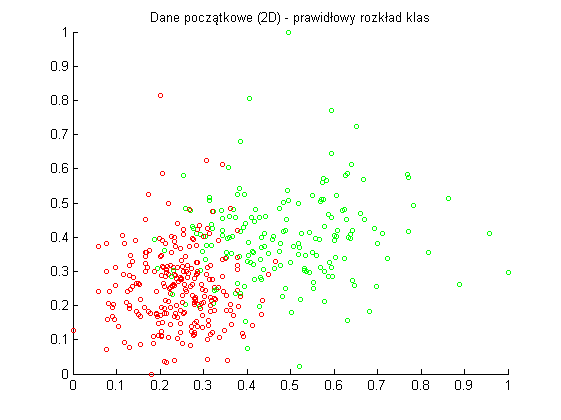
# Ekstrakcja cech na potrzeby wizualizacji danych wielowymiarowych

a.

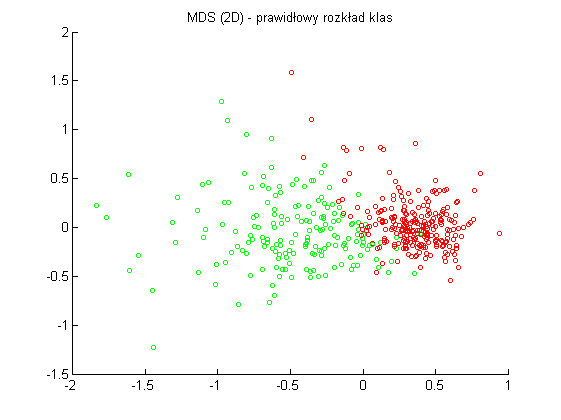
Klasteryzacja metodą ‘*complete’* -192 rozpoznanych na 569 (34%), błąd (66%).

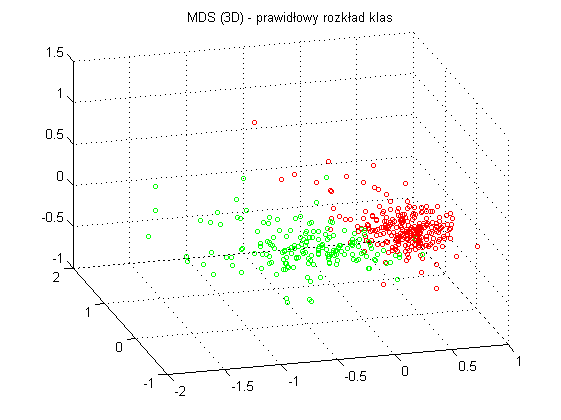
AdaBoost – błąd: 5,59 %

b.



c.

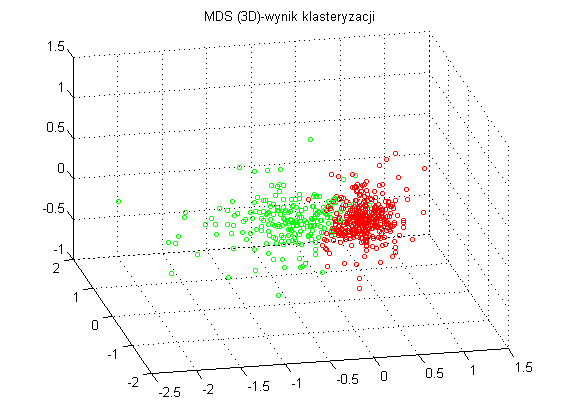
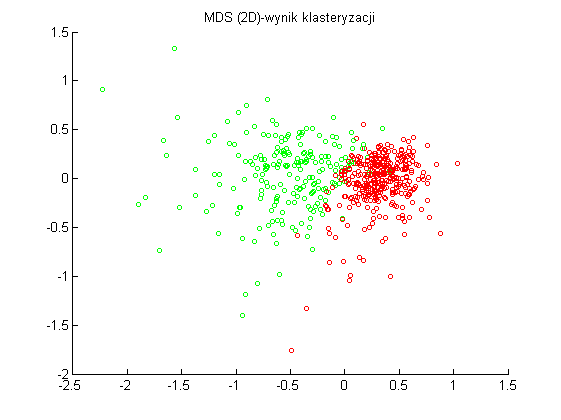
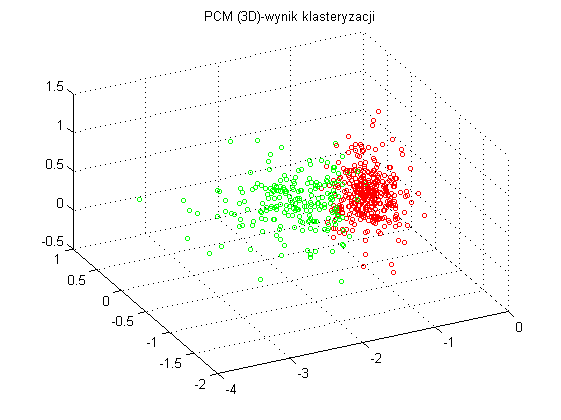
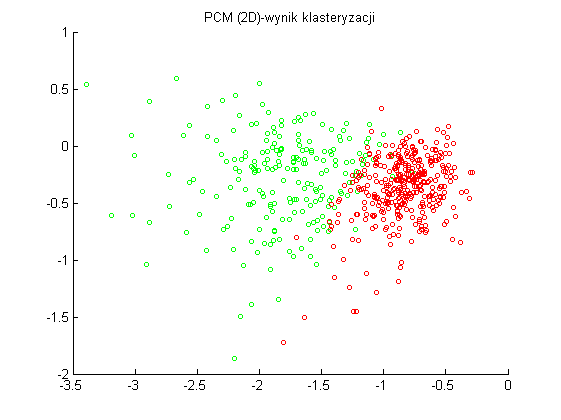




d.

pcm – 368 hits, 64,67% rozp

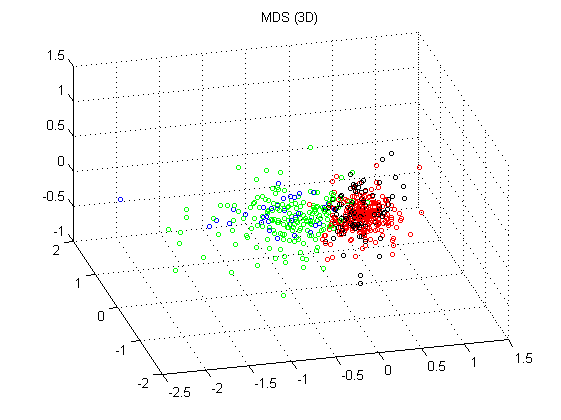
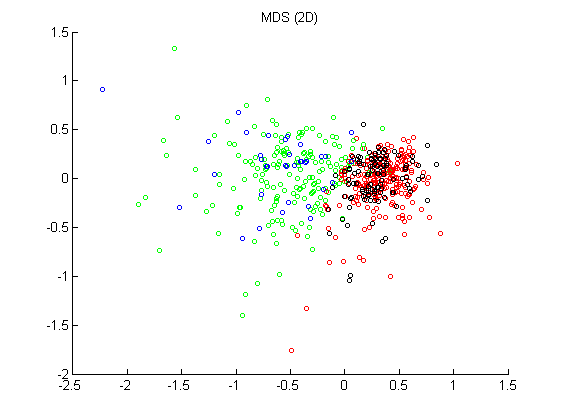
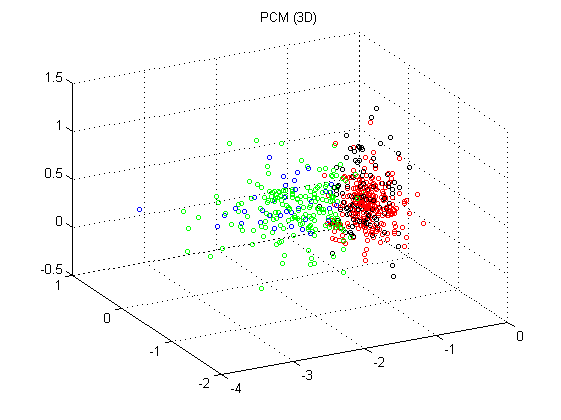
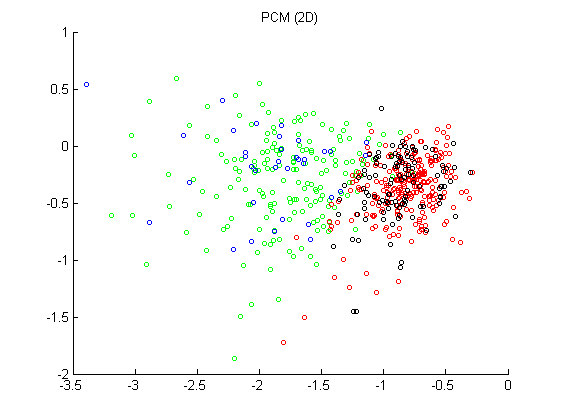
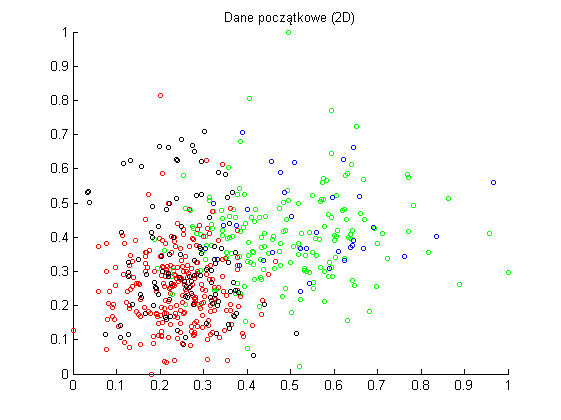
mds – 500hits, 87,87% rozp



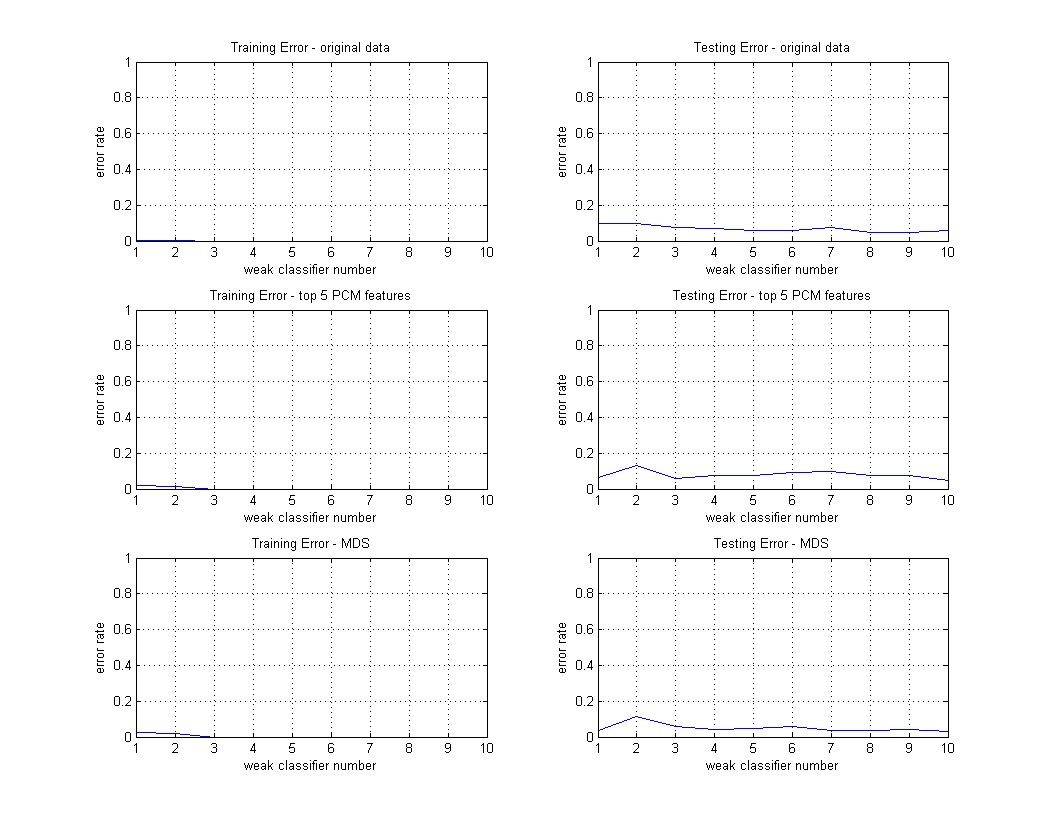
e.

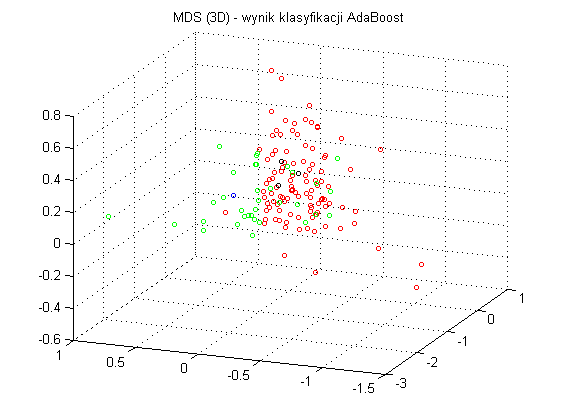
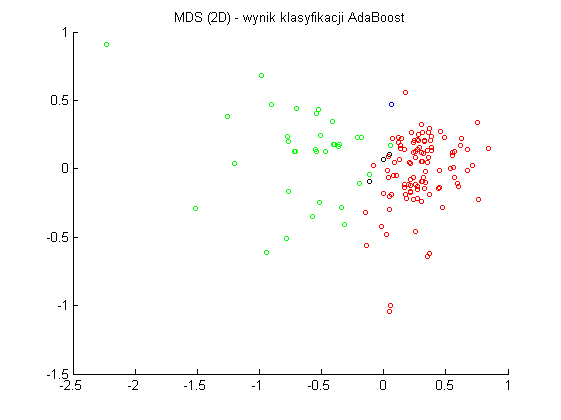
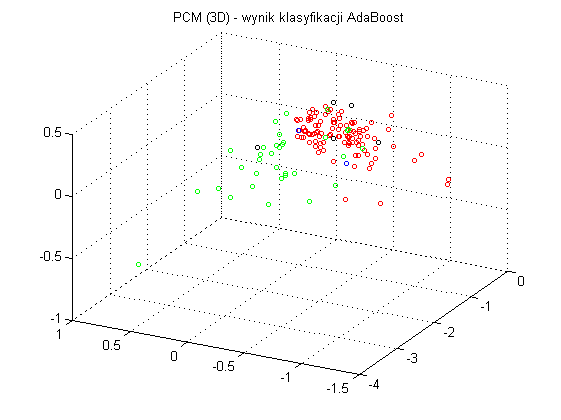
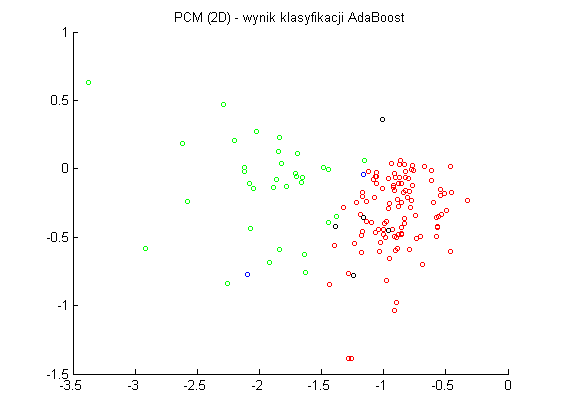
ziel – klasa 1 tr, nieb – klasa 1 te

cze – klasa 2 tr, czar – klasa 2 te



f.





# Wnioski