Porównanie wizualizacji PCA i LDA

Aliaksandr Karolik, Łukasz Knigawka, Bazyli Reps

7 marca 2020

1 Opis problemu

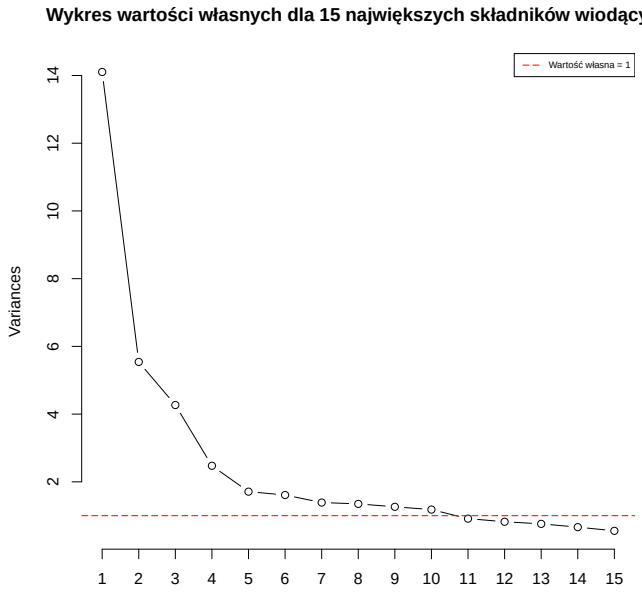
Celem niniejszej pracy jest porównianie kompresji wymiarów danych za pomocą metod PCA oraz LDA.
Zamierzamy przedstawić przede wszystkim analizę redukcji wymiarowości dla obu metod, a następnie porównać je pod względem czasu wykonania redukcji a także uzyskanego stopnia kompresji.
Skrypty potrzebne do wykonania analiz napisane zostały w języku R.

2 Opis danych

Analizy przeprowadzone zostały na danych pobranych ze strony UCI Machine Learning Repository: `http://archive.ics.uci.edu/ml/datasets/SPECTF+Heart`
Przedstawiają one badania wyniki badań serca dla poszczególnych ROI (Region Of Interest) w zależności od stanu badanej osoby (w spoczynku lub podczs wysiłku).
Dane zawierają 44 kolumny zawierające dane dla 22 ROI w obu stanach oraz jedną kolumnę opisującą która przyjmuje dwie wartości: 'normal' oraz 'abnormal'. Zestaw zawiera 265 rekordów.

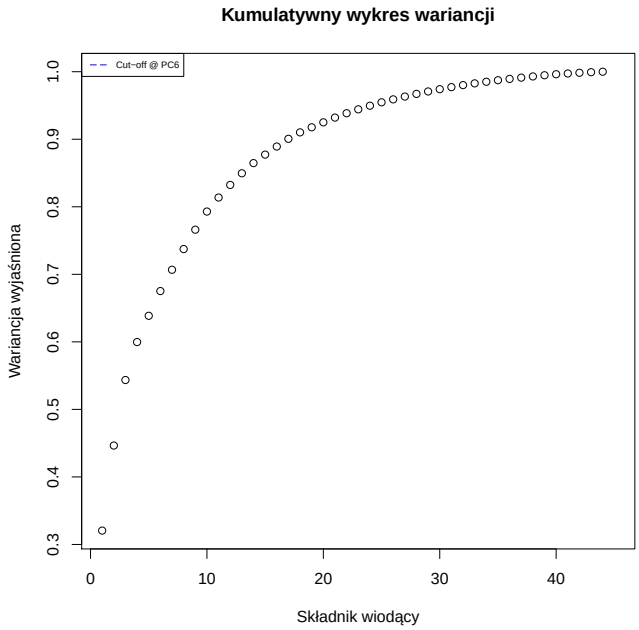
3 Analiza PCA

Dane zostały zredukowane do dwóch wymiarów. Na początku zostały wyznaczone wartości własne.



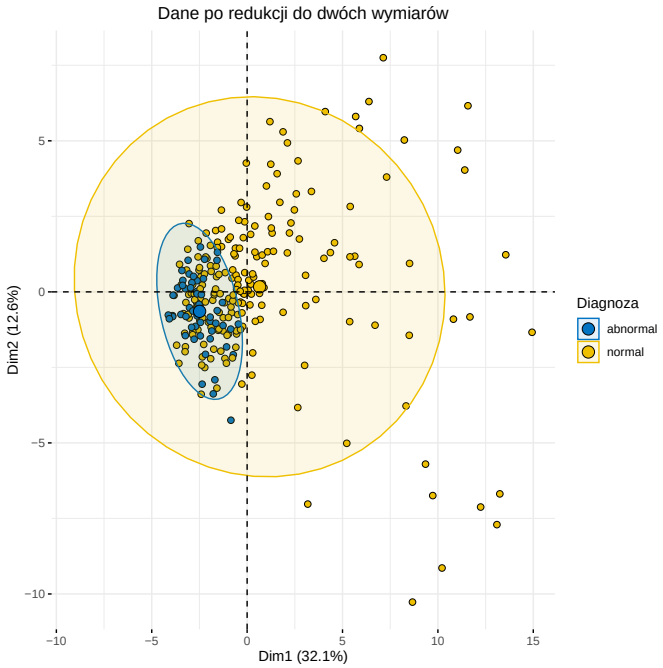
Rysunek 1: Istnieje dziesięć wartości własnych większych od 1. Pozostałe można swobodnie usunąć.

Następnie zbadana została wariancja wyjaśniona w zależności od ilości wziętych składników wiodących.



Rysunek 2: Największe dwa składniki wiodące wyjaśniają około 45% wariancji danych

W dalszej kolejności zwizualizowane zostały dane po redukcji do dwóch wymiarów. Na wykresie zaznaczone zostały klasy obiektów.



Rysunek 3: Wyraźnie widoczne jest zgrupowanie rekordów z diagnozowanych jako 'abnormal'. Niestety charakter danych, a co za tym idzie mniejsza od 60% wyjaśniona wariancja powodują, że pomimo redukcji obie klasy nie są wyraźnie od siebie odseparowane.