UNIVERSITY AT BUFFALO

CSE 601

Data Mining and Bioinformatics

Fall 2019

Project 1

Dimensionality Reduction & Association Analysis

Amit Anilkumar Menon (amitanil@buffalo.edu)

Linus Castelino (linuscas@buffalo.edu)

Deepak Ranjan (dranjan@buffalo.edu)

Principle Component Analysis (PCA)

- It is a method where data are reduced from higher dimension to lower dimension.
- It's the transformation of the correlated variables into linearly uncorrelated variables.
- All PCA are orthogonal to each other and are defined based on the variance in the data.
- First PCA comes are the one with highest variance and so on.

PCA Implementation:

Preprocessing of data i.e. extracting data from the file:

- Read input file and transform the file data into array object for processing.
- Separate data into feature and labels.

PCA Steps:

- Get the mean of the data along each dimension
- Adjust the data around the mean using below:

  X = X – mean(X)

- Find co-variance matrix for new adjusted data
- Calculate the eigen-vector and eigen-value of the above calculated co-variance matrix.
- Pick the top two eigen-value and those two values becomes PCA.
- Plot the data in these two reduced dimensions.
- Get the new data set.

Scatter Plots for different datasets provided:

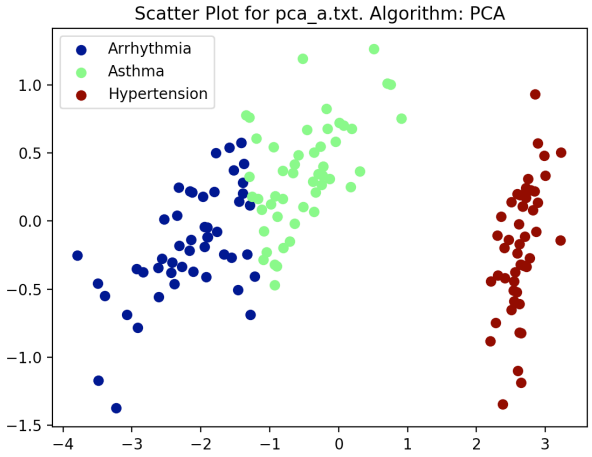


Figure 1: pca_a.txt Scatter Plot

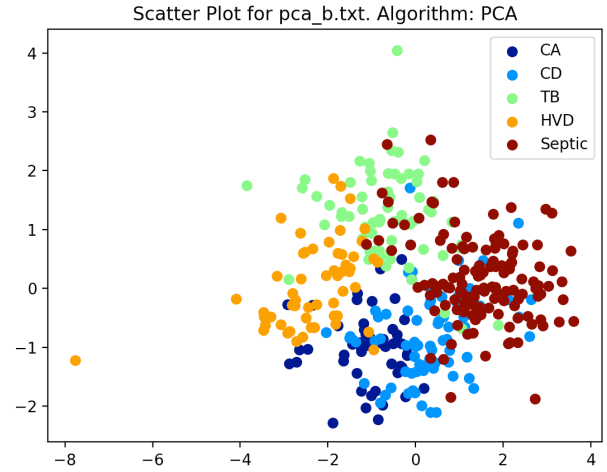


Figure 2: pca_b.txt Scatter Plot

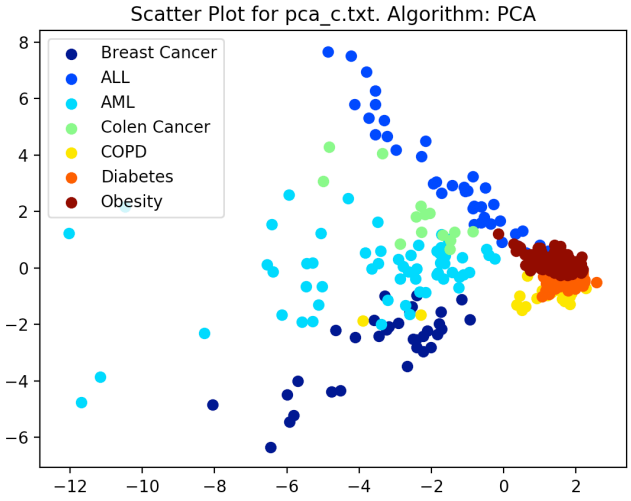
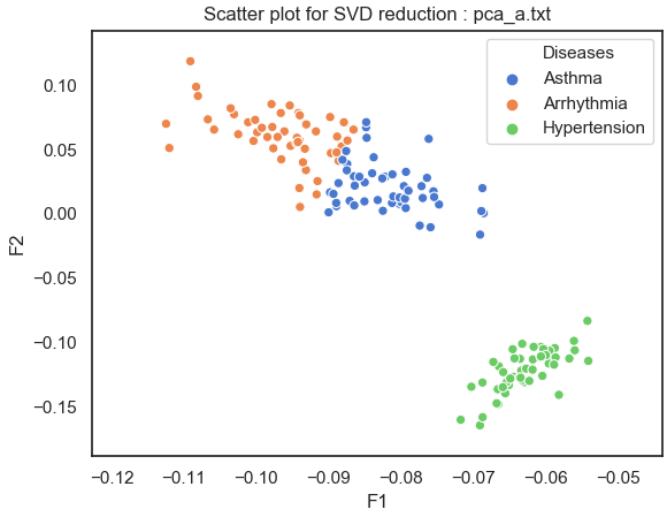


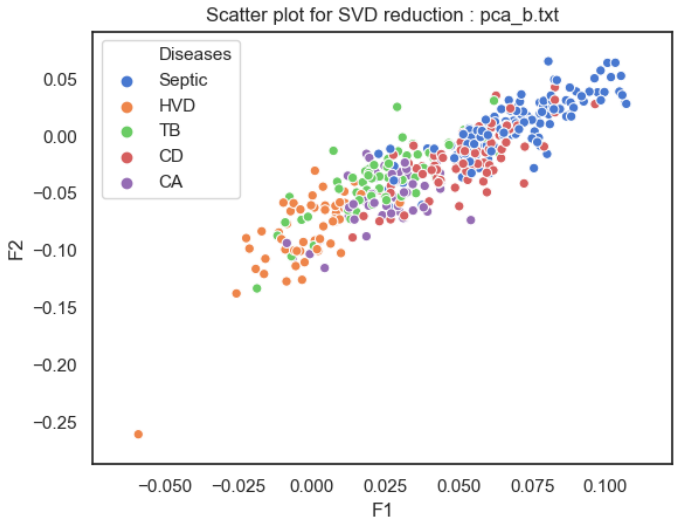
Figure 3: pca_c.txt Scatter Plot

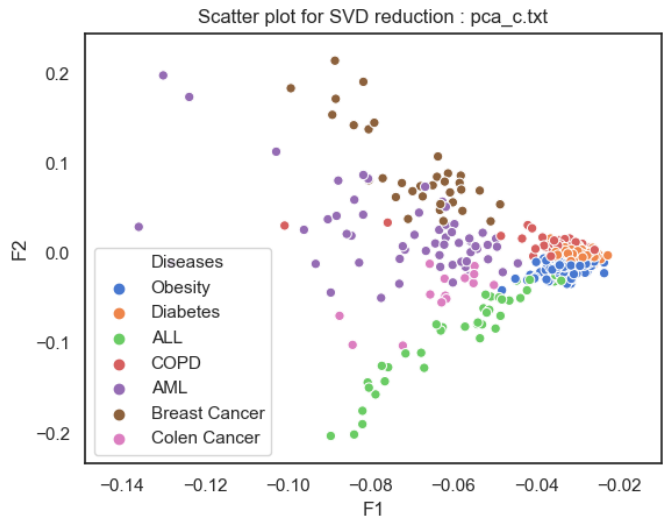
SINGULAR VALUE DECOMPOSITION

- Singular value decomposition in linear algebraic terms, is the process of decomposing a matrix into 3 matrices $A = UDV^T$.
- In dimensionality reduction, U becomes a low rank approximation of the data matrix A.

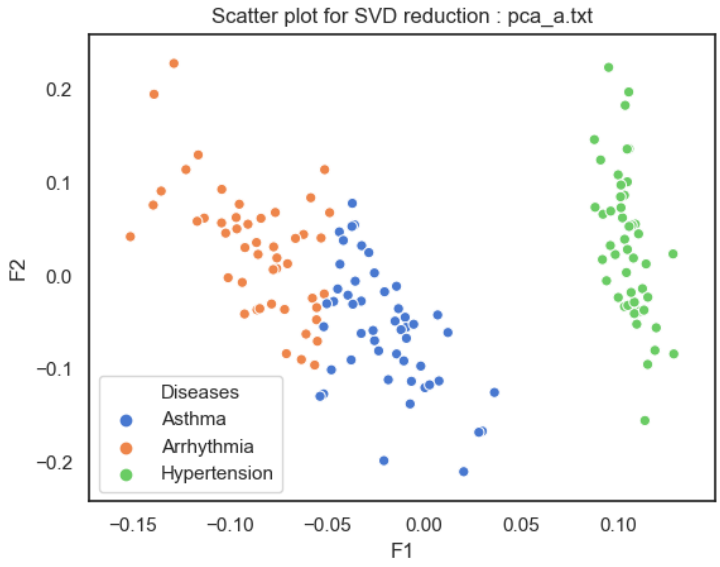
Scatter plot results for SVD:





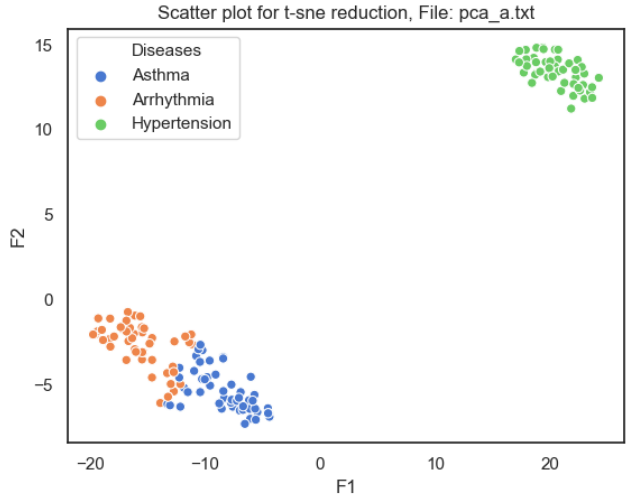


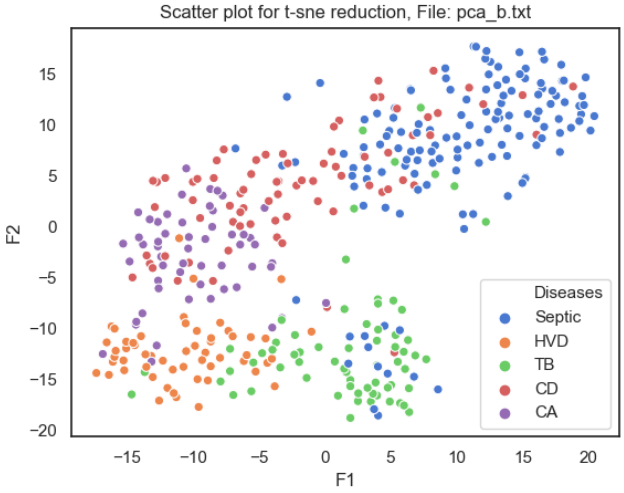
- Since the value of D obtained in the result of S.V.D. are in sorted (descending) order, the inference is that the first 2 columns of U gives the strongest approximation of the original matrix.
- Secondly PCA gives the results centered around the mean, so if we give data centered around the mean we get results similar to PCA as shown.

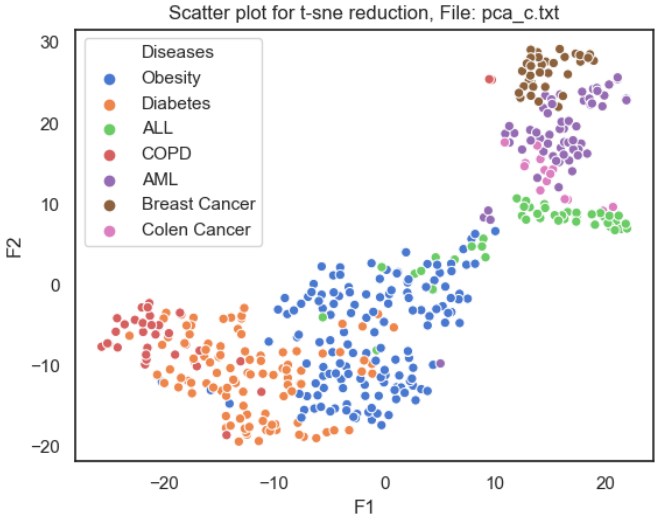


T-SNE Results:

- T-SNE method uses a probability distribution using the relationships between all the data points.
- It then projects it on lower dimensional space using a ‘t’ distribution that resembles the original one as closely as possible.


Scatter plot for t-sne reduction, File: pca_a.txt
Diseases
Asthma
Arrhythmia
Hypertension
F1
F2


Scatter plot for t-sne reduction, File: pca_b.txt
Diseases
Septic
HVD
TB
CD
CA
F1
F2


Scatter plot for t-sne reduction, File: pca_c.txt
Diseases
Obesity
Diabetes
ALL
COPD
AML
Breast Cancer
Colen Cancer
F1
F2