Data Mining and Bioinformatics

Project 1 Report: Dimensionality Reduction and Association Rule Mining

Team 23:

Arun Sharma (arunshar)

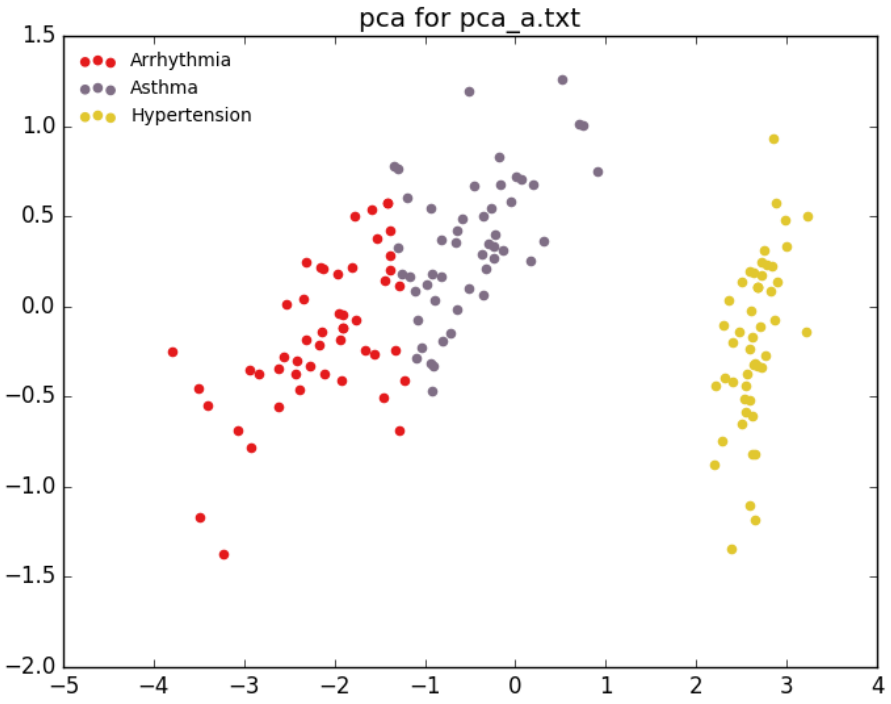
Prashanth Chandrashekar (pc74)

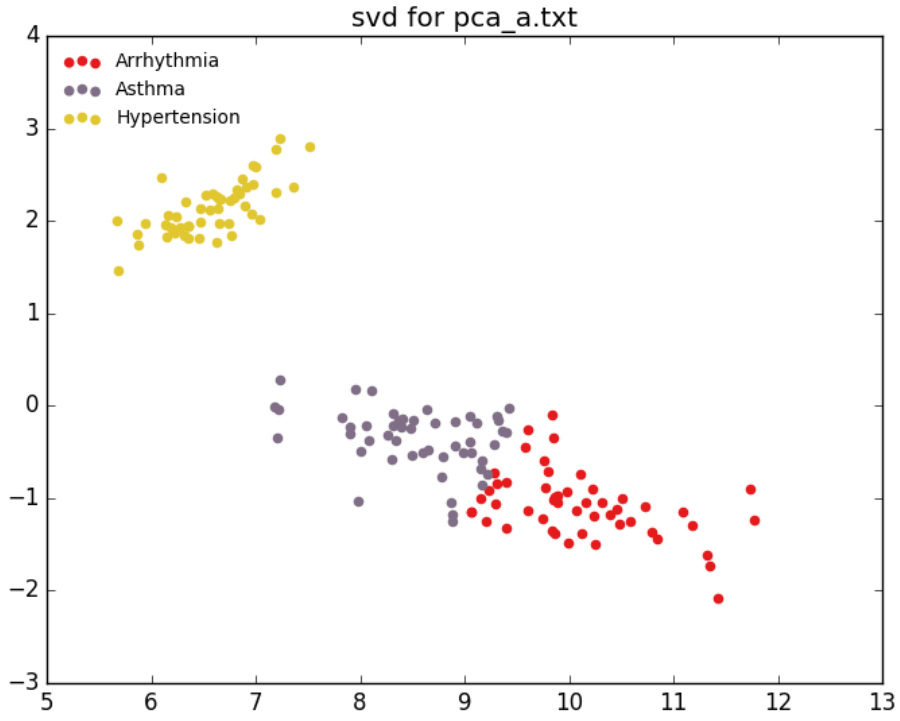
Vipin Kumar (vkumar25)

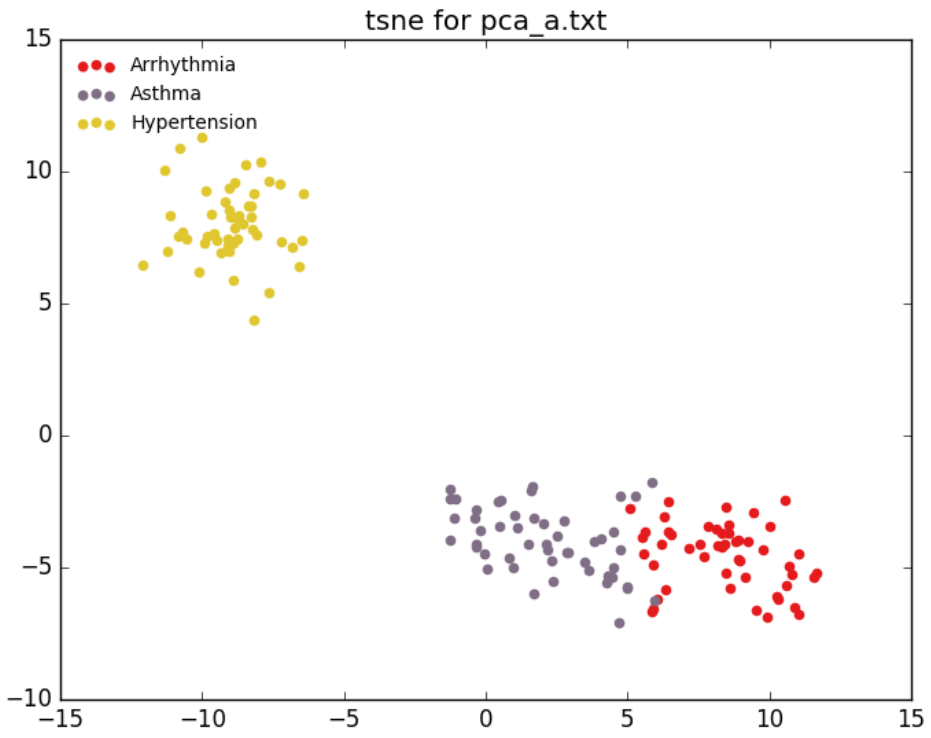
Dimensionality reduction using PCA, SVD and tSNE

The following are the plots obtained for the three given datasets using the three algorithms:

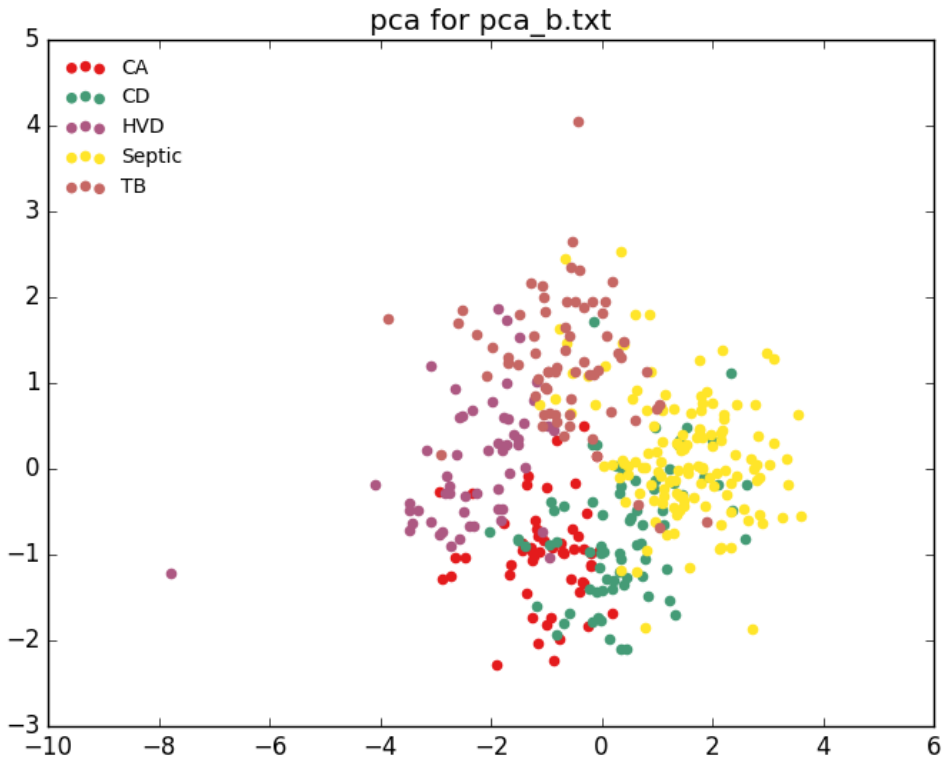
1) PCA, SVD and tSNE for pca_a.txt:

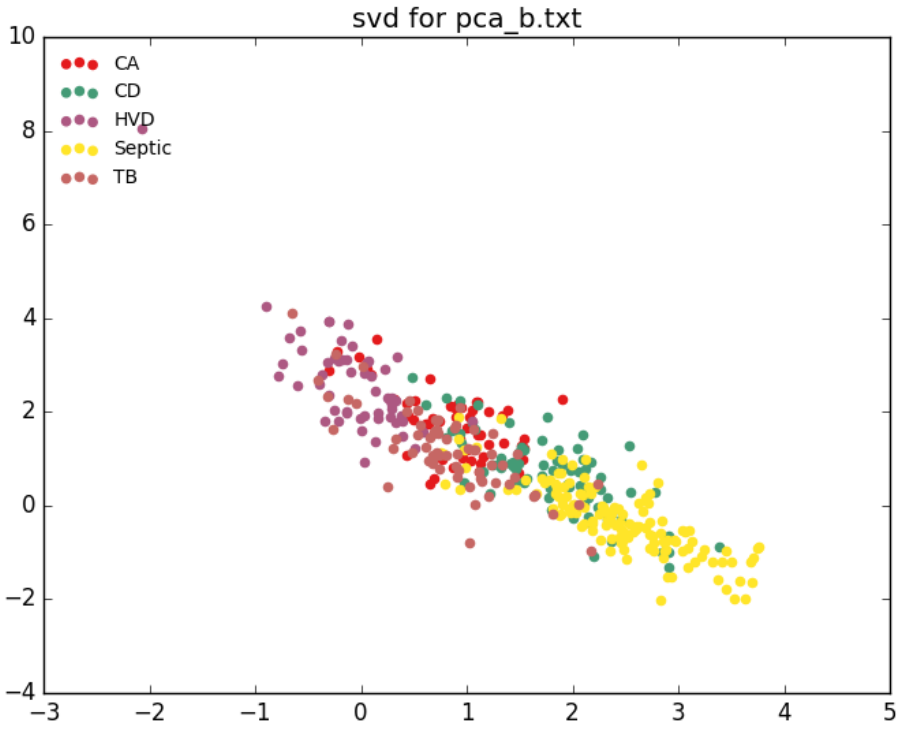


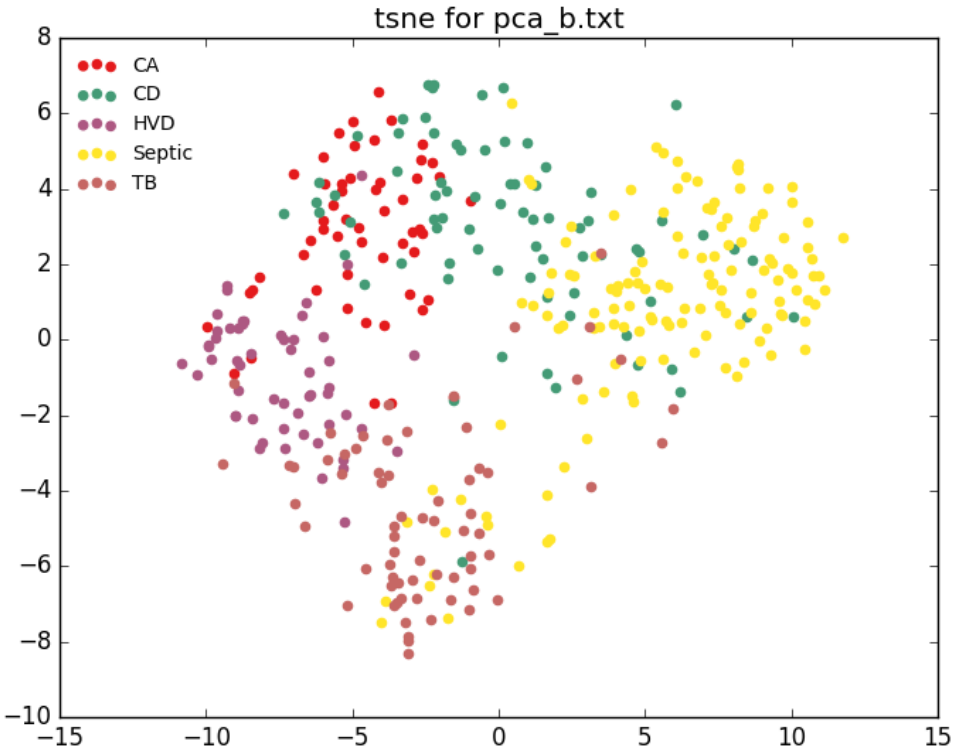




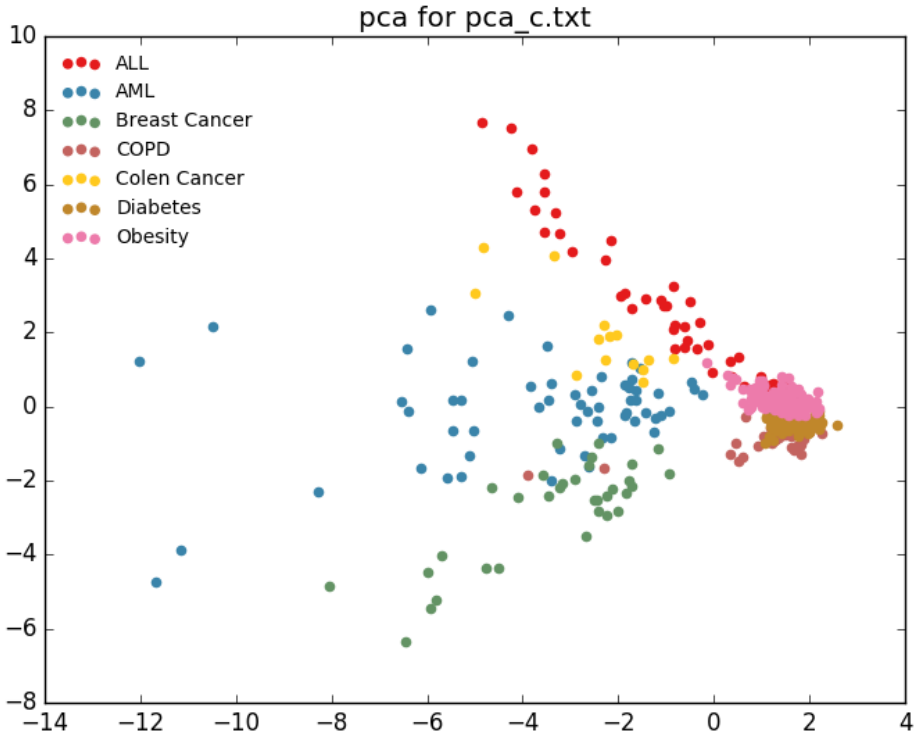
2) PCA, SVD and tSNE for pca_b.txt:

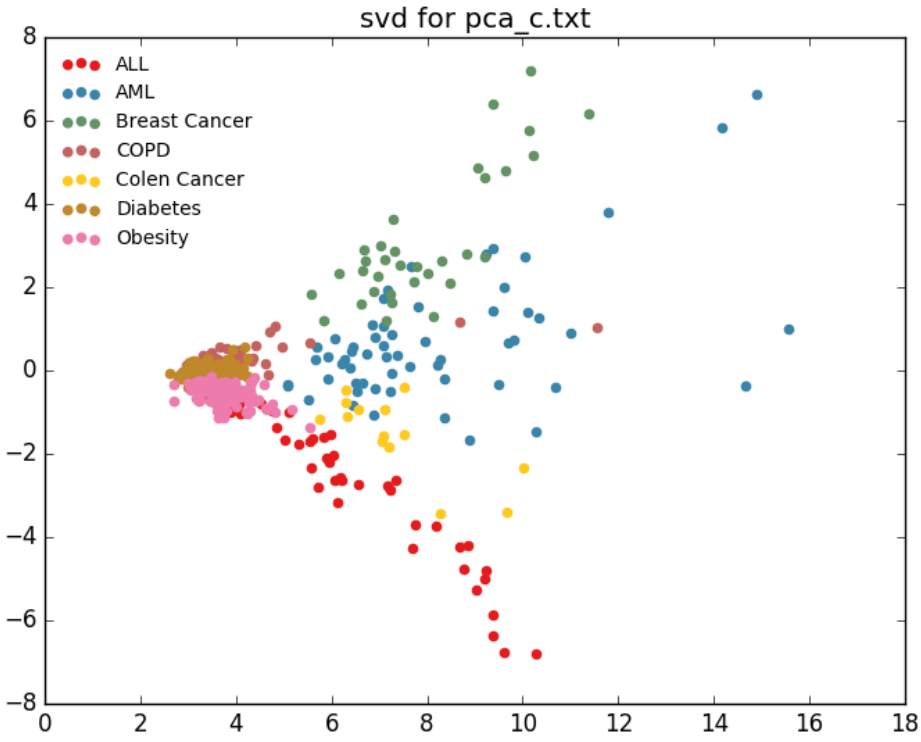


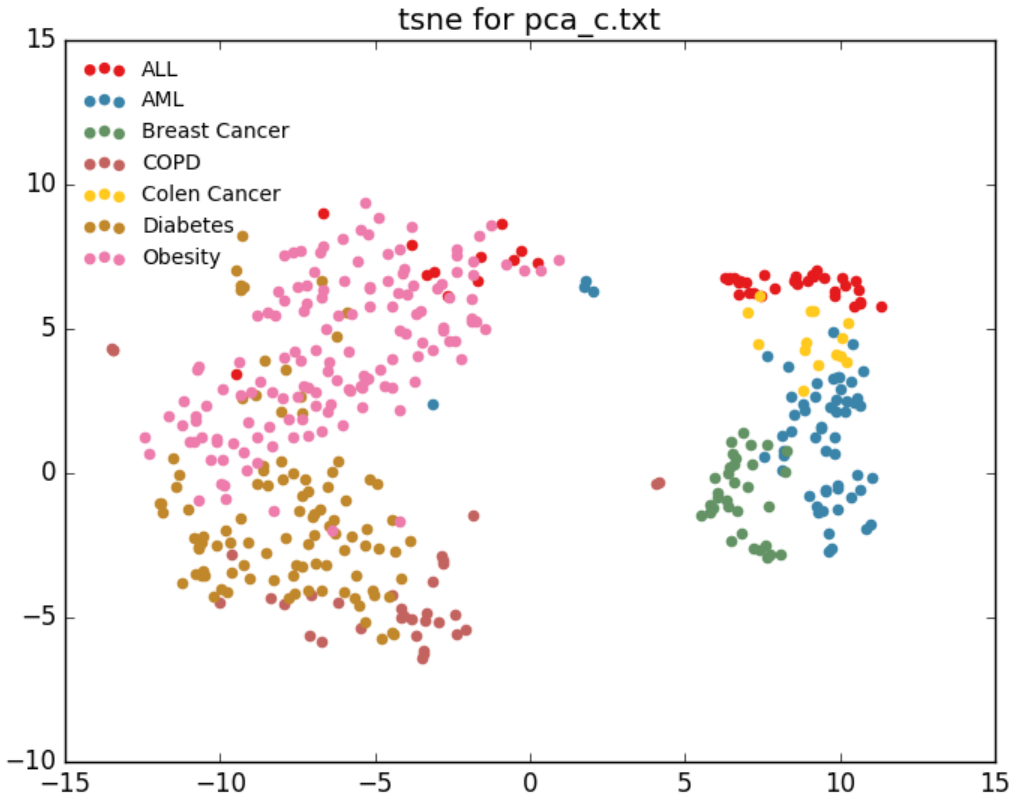

svd for pca_b.txt
CA
CD
HVD
Septic
TB
10
8
6
4
2
0
−2
−4
−3
−2
−1
0
1
2
3
4
5


tsne for pca_b.txt
CA
CD
HVD
Septic
TB
8
6
4
2
0
−2
−4
−6
−8
−10
−15
−10
−5
0
5
10
15

3) PCA, SVD and tSNE for pca_c.txt:







Brief Discussion of PCA Implementation:

- The input dataset contains an extra column containing the type of the disease, which is used to label the data points and will be ignored while implementing PCA.
- For each attribute, we find the mean and subtract it from all the attribute so that our data is now zero centered, i.e. has a mean of zero. We now compute the covariance matrix of the given data.
- We now obtain the Eigen values and Eigen vectors of this covariance matrix and then get the top ‘d’ Eigen values sorted in descending order and their corresponding Eigen vectors (d is the number of dimension we are trying to reduce the original data to, viz. 2 in our case).
- To obtain the reduced d-dimensions we need to perform a dot product of the original data with the Eigen vectors corresponding to the top ‘d’ Eigen values.
- This d-dimensional data is plotted with labels from the last column of the original data.

Results and observations:

- Using PCA we can orthogonally transform the given high dimensional (possibly correlated) data into uncorrelated **Principal Components**. The top principal components have the most variance, and hence by choosing the top 2 Eigen vectors (based on the biggest 2 Eigen values), we are able to map it to a 2 dimensional space with the 2 uncorrelated variables.

- In pca_a dataset, we can see that instead of 4 attributes we have projected the data on a 2D plane which has made visualization easier and we can make some useful inference if we have the knowledge of the domain. For eg: if Arrhythmia and Asthma can occur in children, and Hypertension doesn't, it seems the x coordinate is related to age of the person, similarly the other coordinate could be body weight, etc.
- Similarly, on pca_b, where we reduce the 16-feature data to that of a 2-dimensions, we can observe how CD with septic and CA with HVD are observed in patients having similar features in the reduced dimensions, as mentioned earlier, with domain specific knowledge, we could be able to infer useful data about the new dimensions.
- SVD is a factorization of a matrix, where we decompose it to 2 unitary matrices and a diagonal matrix. The original matrix can be reconstructed by multiplying these factors together again. Instead, if we consider only the top 'd' values of the diagonal matrix, and multiply it with the first unitary matrix, we get a new matrix with data mapped to 'd' dimensions.
- In pca_a, we can see that SVD and PCA have similar results, but SVD has better separation of data points for Arrhythmia and Asthma, as compared to PCA which has some overlaps.
- From the examples of pca_a, pca_b and pca_c, we can see that SVD and PCA are similar and the data points are related to each other in a very similar way, i.e. it appears as though the 2D spaces of SVD and PCA are slightly scaled and rotated versions of each other.
- t-Distributed Stochastic Neighbor Embedding is a dimensionality reduction technique, that models similar objects by nearby points and dissimilar objects by far away points. As this uses a non-convex objective function which is minimized by gradient descent initiated randomly, we get a different plot every time its invoked over the same dataset.
- For all three data sets, we ran the algorithm several times to get the plot that captures the best distribution visually to help us in inferring useful information. We can see that the plots for pca_a and pca_b have data distributed similar to the plots of PCA and SVD except that the coordinates seem to be scaled and rotated.
- For the pca_c dataset, we can clearly see that the visualization in tSNE is way better than that of PCA and SVD and we can clearly make inference on how the diseases are classified and spread across in these new dimensions.