Assignment 6

Harshal Rohit 19111036
Raunak Kumar 19111069
Debdan Bhandari 19111026

November 2019

1 Introduction

In this assignment, An implementation of diffusion based dimentionality reduction technique has been performed followed by an unsupervised classification to classify grid conditions into two categories, normal and stressed. The data set and other resources has been obtained from the paper "Methodology for a Security/Dependability Adaptive Protection Scheme Based on Data Mining by Bernabeu, Thorp, Centeno" whereas the diffusion based reduction technique is given in "An Introduction to Diffusion Maps by J. de la Porte, B. M. Herbst, W. Hereman , S. J. van der Walt".

2 Dataset

The dataset containes a total of 4150 different system operating points among which, 2514 are labeled as one, denoting stressed condition and 1636 were classified as zero, denoting normal condition. Every data point has 132 features.

3 Methodology

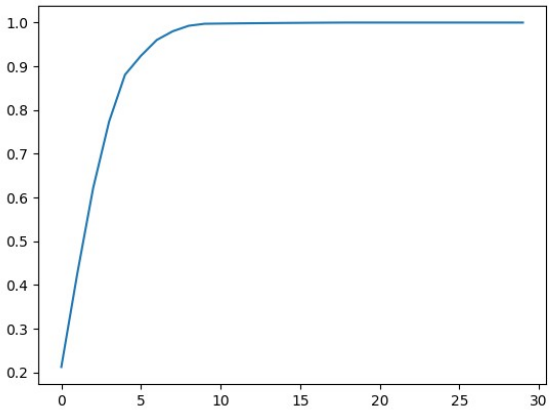
3.1 Diffusion map

Once the training data has been separated, an RBF kernel was applied on the training data to find the pairwise gaussian distance and thus the connectivity matrix(K) was formed which was of size 4150x4150. Then a diagonal matrix D containing the row-wise sum of the connectivity matrix was formed. The diffusion matrix(P) was then formed via the operation $D^{-1}K$. This step normalizes the matrix rows. In the diffusion step, P was multiplied with itseld n number of times where n is a hyperparameter denoting allowed jumps. The suitable value of n and variance of rbf kernel was estimated using a grid search over possible values ranging from 0.01 to 20 for variance and 5 to 30 for n. Then, the matrix P' was formed via the equation $D^{1/2}PD^{-1/2}$ and its eigen-decomposition gave the eigen vectors as S and eigen values as E. The left singular values of P was

obtained by having $D^{-1/2}S$. The dominant m eigen values and vectors were chosen through a scree plot. Upon multiplication of $D^{-1/2}S$ and E, it gave the low dimensional representation of the diffusion matrix. The optimal values of variance is set at 1 and the number of diffusion iterations is set at 10.

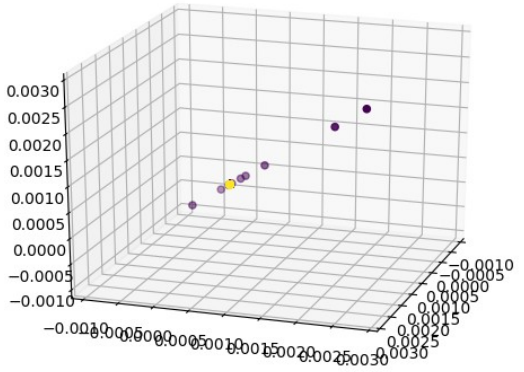
3.2 Dimension reduction

By taking the percentage coverage of the eigen values obtained from the above process, a scree plot was created to obtain the appropriate value of the statistical dimension. The scree plot is shown as below.



3.3 Clustering

For the clustering, K means++ algorithm was applied to find the suitable clusters. Since it is known beforehand that there are 2 classes, K means was applied with 2 means. To see whether any cluster can be seen after the dimensionality reduction, 3D plots are created by taking the dominant 3 eigen vectors and projecting the 4150 dimensional points onto a 3D space. The results are shown as follows.



The points in the projected space are as close as e^{-6} and hence appear to be packed together in the plot. Since the distance in the projection space are euclidean in D^{-1} metric and not in standard identity metric, K means, which uses euclidean distance doesn't appear to be giving proper results. On the other hand, a supervised learning like K nearest neighbors that uses the D^{-1} metric achieves over 80% accuracy, showing that diffusion and subsequent projection indeed reduces the dimension without losing relevant data.

4 Conclusion

The Diffusion process appears to be appropriate to estimate the non-linearity of the data points since the scree plot shows sharp improvement in the number of eigen values required to represent the original matrix and the matrix after diffusion. However, the data don't seem to separate even after that when seen from the standard euclidean perspective. The clustering on such data would produce a single supercluster containing all the data points and the other cluster would contain as few as one data point in it. However, use of D^{-1} metric to calculate distances do give better results.