Assignment 3 - Unsupervised Learning and Dimensionality Reduction

Prithi Bhaskar
pbhaskar8@gatech.edu

1 Datasets- What makes them interesting?

Two classification/clustering problems have been identified for the purpose of this assignment. The details of the problems and their respective data sets have been discussed in detail in the following sections. One of the main factors that makes the datasets interesting is that the data sets carry a non-trivial amount of data as well as features, and thus provide numerous opportunities to use different algorithms to gain deeper insights into the data. This attribute also helps in identifying the limitations and strengths of various unsupervised and dimensionality reduction algorithms as a by-product.

1.1 Dataset 1

Problem1 attempts to study the effects of breast cancer on the proteomic(protein expression) landscape of those affected by breast cancer. Somatic(genetic) mutations have been extensively characterized in breast cancer and such mutations change the protein expression landscape of affected persons. The dataset, obtained from kaggle presents 77 breast cancer samples with roughly 12000 protein values for each of the samples. Analysing the proteomic landscape helps identify certain subtypes of breast cancer as opposed to the subtypes identified by the conventional RNA-expression based clustering. The RNA based clustering identifies 4 subtypes of cancer and has been used as a benchmark for comparing the clustering achieved based on proteomic data. The need to identify subtypes different from the conventional ones arises in order to distinguish therapeutic targets from non treatable ones.

1.2 Dataset 2(From Assignment1)

Problem 2 is a multilabel classification problem involving the Avila data set from UCI(the same dataset used in Assignment 1). The dataset has been extracted from 800 images of the 'Avila Bible', an XII century giant Latin copy of the Bible. The prediction task is to associate each pattern of the script to a copyist. The features include numerical data such as intercolumnar distance, margins, interlinear spacing, etc. There are 12 classes to which the data belong to and the distribution is imbalanced with the majority class representing 4286 instances and the minority class representing 5 instances.

2 Problem 1- Analysis

2.1 Clustering

In order to find an efficient way to cluster the data, different values of k were passed to the algorithms to study the silhouette score and the homogeneity scores. Homogeneity scores were considered since a clustering similar to that of the conventional RNA based clustering is desired.

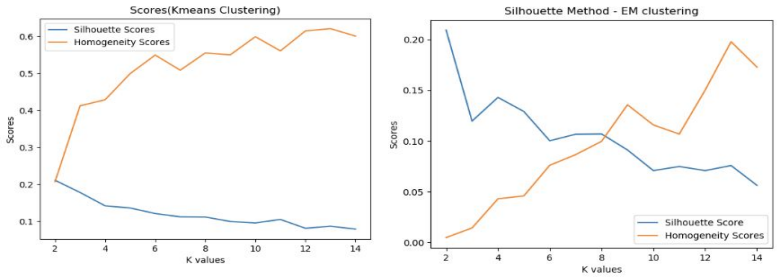

Figure 1— Silhouette, Homogeneity scores as a function of k

As can be observed from the figure, the silhouette score is still high and the homogeneity score jumps ~2-fold at k=3 for kmeans. For EM, the homogeneity score is very low but the silhouette score is the same as kmeans. This shows that the gaussian mixture model clusters the data in a way very different from the conventional clustering. K=3 was chosen to study the clustering of the data.

2.1.1 Reason behind the obtained clusters

From Figure 2, it can be observed that both the algorithms are able to find clusters based on proteome data that are different from the conventional clustering based on pam50RNA data(the second column in both figures).

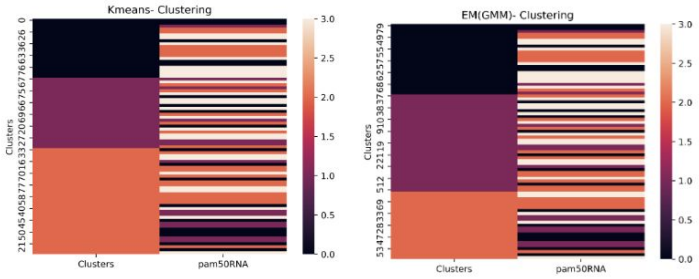

Figure 2— Heatmap of clustering achieved by Kmeans and EM along with PAM50 labels

However since the clustering is based on proteomes, the clusters are broadly based on the difference in the underlying protein signature and samples having similar signatures are grouped together.

2.1.2 Do the clusters make sense?

As shown in Figure 2, the clusters do not line up with the conventional clusters. However they still make sense and the difference is rather expected and even desired in order to gain more insights into the field of cancer studies. As per one of the studies conducted on the dataset, these categories thus identified by clustering based on proteomes seem to represent three novel subtypes of breast cancer viz., basal-enriched, luminal-enriched and stromal-enriched as opposed to the conventional subtypes viz., basal-like, luminalA, luminalB and HER2-enriched.

2.2 Dimensionality Reduction

In order to find the number of components that can best express the data after applying reduction, a number of graphs showing the distribution of variance among the components and

reconstruction errors were plotted for all the dimensionality reduction algorithms. They are presented in the following sections.

2.2.1 Principal Components Analysis

As shown in Figure 3, the first component shows just ~28% variance and this explains the high reconstruction error. However, the first 5 components contain the majority of information and the corresponding RMSE is comparatively low. Hence 5 components were chosen to do the clustering.

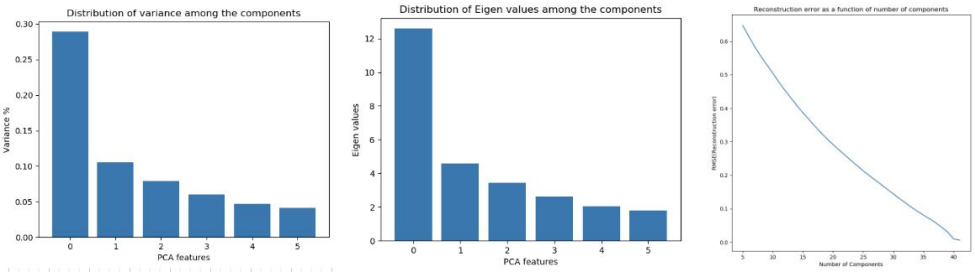

Figure 3— (Left)Variance, (middle)Eigenvalues, (right)Reconstruction error

2.2.2 Independent Components Analysis

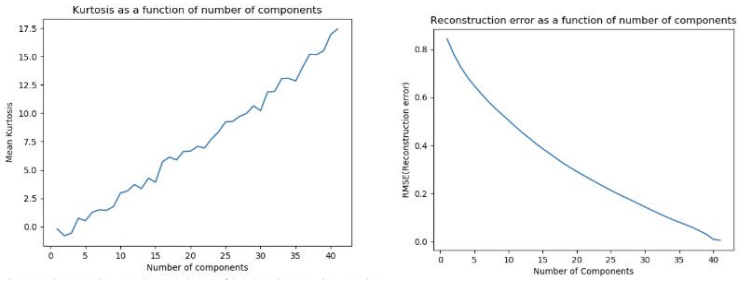

Figure 4— (Left)Kurtosis, (right)Reconstruction error

From the above plots, the optimum number of components chosen for ICA was 25 as the mean kurtosis is still high representing a non-gaussian distribution and the reconstruction error is also low. When compared to PCA, which was efficient with just 5 features, ICA needs a lot of features. This can be explained by the nature of data itself with features inherently interdependent on each other.

2.2.3 Random Projections and SVD

From Figure 5, it can be observed that the lowest reconstruction error for RP(error averaged over 10 runs) is at n=40 and the 4th algorithm, SVD is at n=30(with some acceptable reconstruction error). Even with 40 components and parameter tuning, the error of RP is higher than the other algorithms. This happens to be one of those problems where RP is not a good fit since all the other algorithms outperform RP by a significant margin.

2.3 Clustering after Dimensionality Reduction

As shown in figure 6, all the algorithms are able to cluster the data as separate blobs. The components that best explained the spread of data were chosen for plotting the figures. With RP, the data is not well separated and

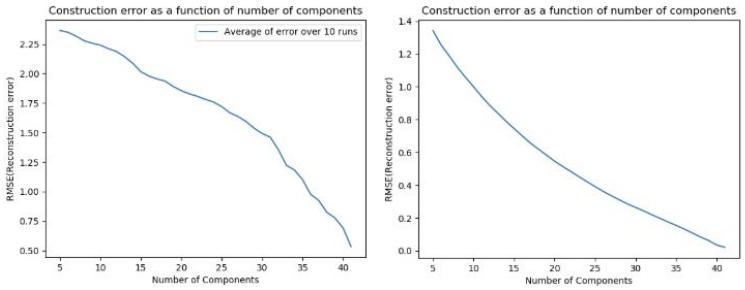

Figure 5— (Left) RMSE- RP, (right) RMSE- SVD

as evident from the plots, the corresponding silhouette score is also lesser comparatively(which is not shown here due to space constraints).

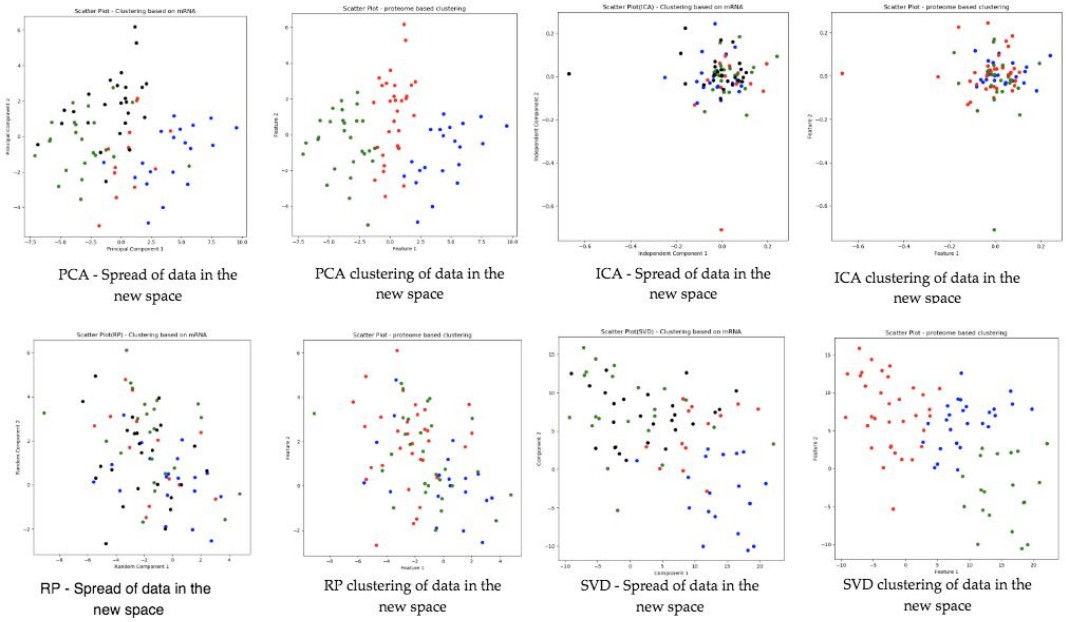

Figure 6— Spread of data in spaces reduced by various algorithms and the same after clustering.

With the reduced data, kmeans and EM both produced similar clusters with all types of reduction algorithms. The silhouette and homogeneity scores were analysed for different values of k as before and the optimum value of the k remains the same at 3 with the same scores as obtained before. This shows the efficiency of the dimensionality reduction algorithms as the same clusters have been obtained with same scores but with reduced number of features.

Particularly, data reduction by PCA has been the most efficient for this particular dataset as the number of components was reduced to just 5 from 42.

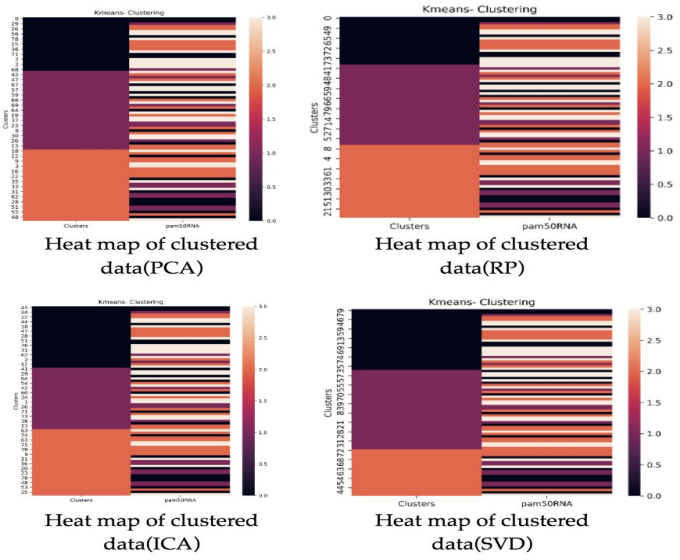

Figure 7— Heatmaps of kmeans clustered data after dimensionality reduction.

Clustering obtained from EM is not shown as they are the same for all algorithms. It can be observed from the above figure that data from all the algorithms are clustered the same way except for RP. The borders obtained from RP data differs from the ones obtained from the rest of the data resulting in very low homogeneity scores due to the reasons cited earlier.

3 Problem 2- Analysis

3.1 Clustering

As with problem 1, silhouette and homogeneity scores were plotted for different values of k in order to determine the optimum value of k. Since all features are numeric and none are categorical, the metric used for measuring distances was euclidean as it directly translates to the physical distance between two points and it also makes sense for this problem as the features used are inter-columnar distances, margins, etc.

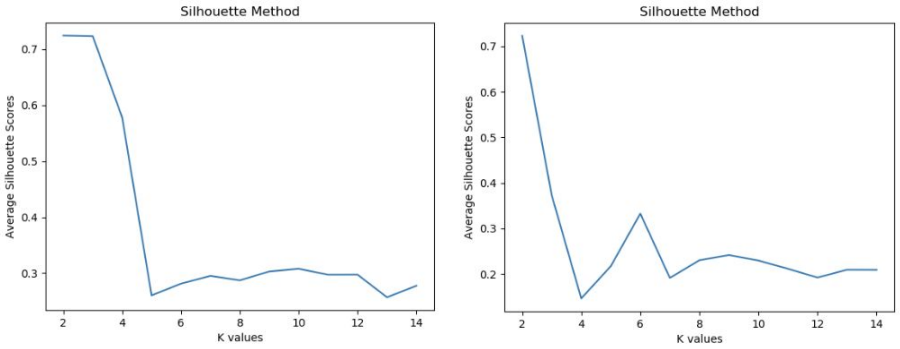

Figure 8— (Left)Silhouette score kmeans, (Right)Silhouette score EM

It can be observed from Figure 8 that the silhouette scores are at their peaks for k=2 for kmeans and EM. Since the original dataset has 12 labels and the clustering shows only 2 sets, it does not make sense to take homogeneity scores into account. Hence the clusters were plotted for k=2 for kmeans and EM to analyse the same.

3.1.1 Reason behind the obtained clusters

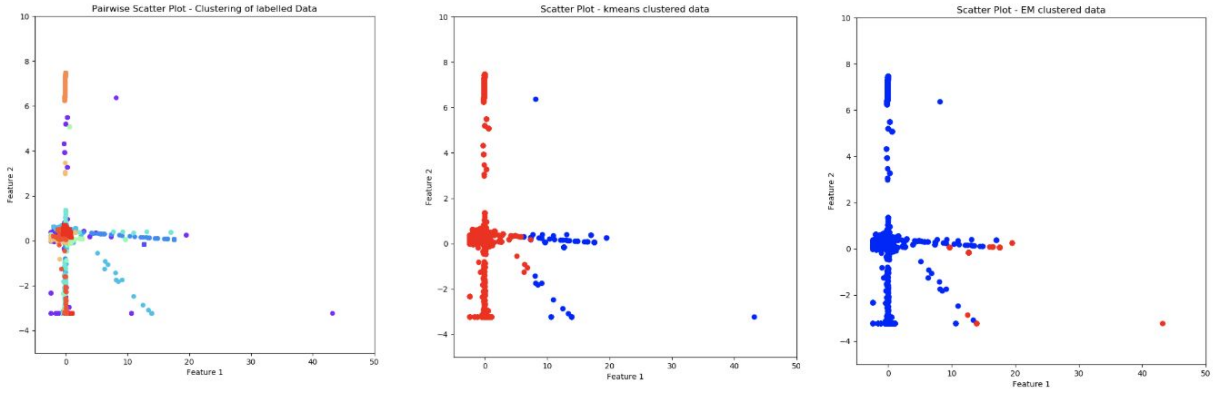

Figure 9— (Left)Spread of actual data, (Centre)Kmeans-Clustered data, (Right)EM-Clustered data

The features that best explains the spread of data were chosen from the pair plots for illustration purposes. The above figures were plotted using the chosen features. As shown, the clusters are made based on the range of values the features are in. The original labelled data is of 12 clusters with minute differences between each of them, which can be learnt using supervised learning algorithms. However, the clusters still make sense as they identify supergroups and help gain an understanding of the data, which is the primary goal of unsupervised learning.

3.2 Dimensionality Reduction

3.2.1 Principal Components Analysis

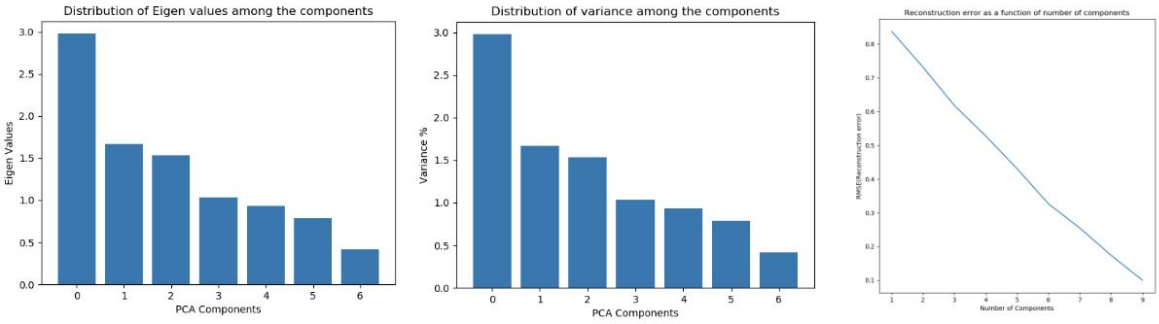

Figure 10— (Left)Eigenvalues, (middle)Variance, (right)Reconstruction error

As shown in Figure 10, the first component shows just ~30% variance and this explains the high reconstruction error. However, the first 6 components contain the majority of information and the corresponding RMSE is comparatively low. Hence 6 components were chosen to do the clustering.

3.2.2 Independent Components Analysis

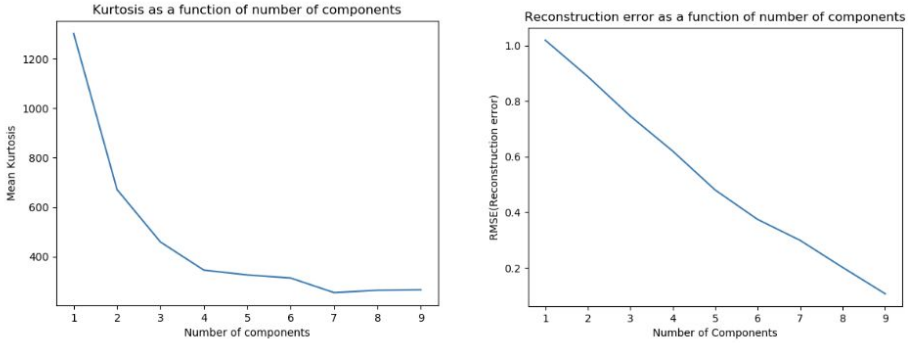

Figure 11— (Left)Kurtosis, (right)Reconstruction error

From the above plots, the optimum number of components chosen for ICA was 8 as the mean kurtosis is still high representing a non-gaussian distribution and the reconstruction error is also low.

3.2.3 Random Projections and SVD

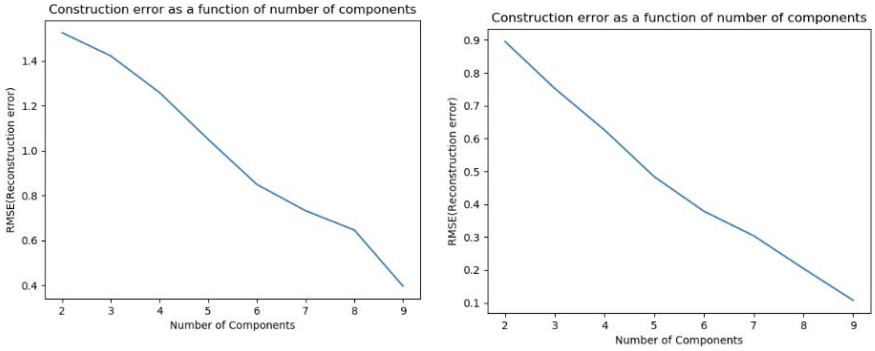

Figure 12— (Left) RMSE- RP, (right) RMSE- SVD

From Figure 12, it can be observed that the lowest reconstruction error for both RP(error averaged over 10 runs) and the 4th algorithm, SVD is at k=9. Even with 9 components and parameter tuning, the error of RP is higher than the other algorithms for this dataset as well.

3.3 Clustering after Dimensionality Reduction

In order to view the spread of data in the new spaces produced by various algorithms, pair plots of all the features were analysed and the components that best explained the spread of data were chosen for plotting the figures. The same has been shown in Figure13. Both kmeans and EM produced similar kind of clusters. EM clusters are not shown in the report due to space constraints.

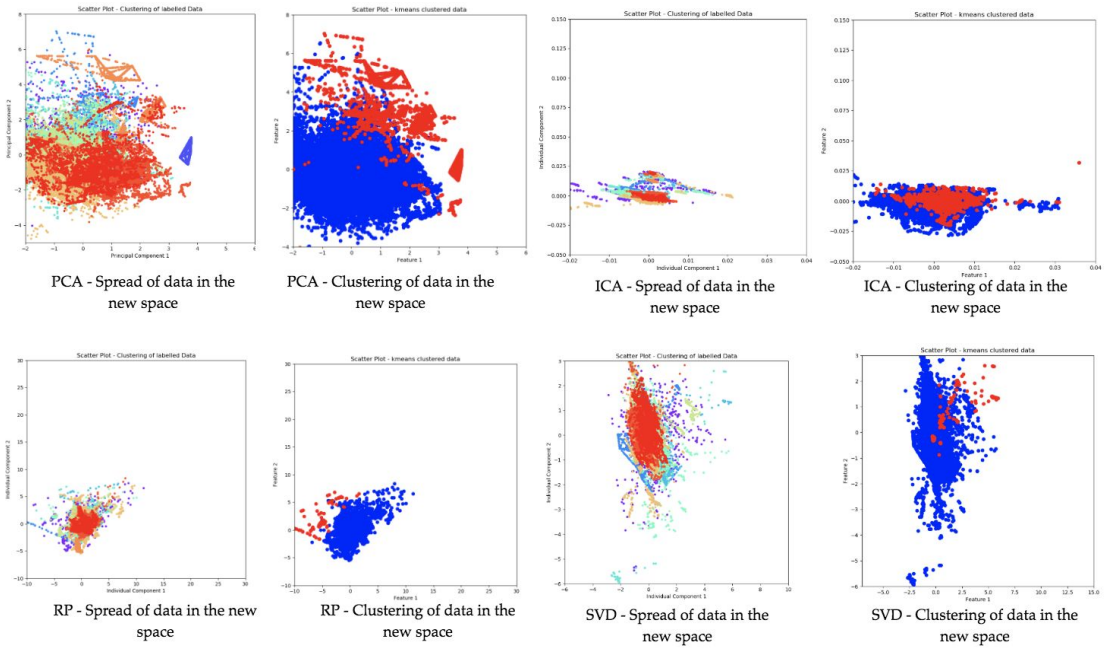

Figure 13— Spread of data in spaces reduced by various algorithms and the same after clustering.

Figure 14 shows the comparison of clustering obtained before applying reduction algorithms and after applying the same. As can be observed, the same kind of clusters are produced since components were chosen such that maximum information is retained while also reducing the number of features. Data from all reduction algorithms produced similar clusters and hence only PCA reduced clustering is shown in the figure.

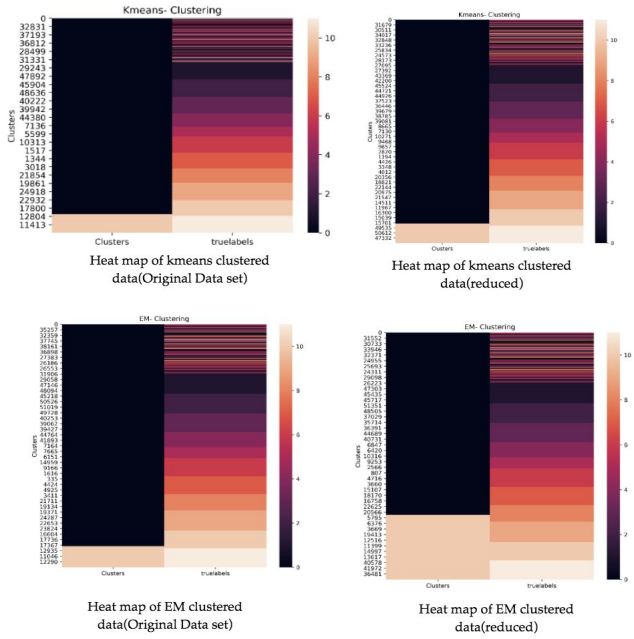

Figure 14— Heatmaps of clustered data after dimensionality reduction.

3.4 Neural Networks Training with Reduced data

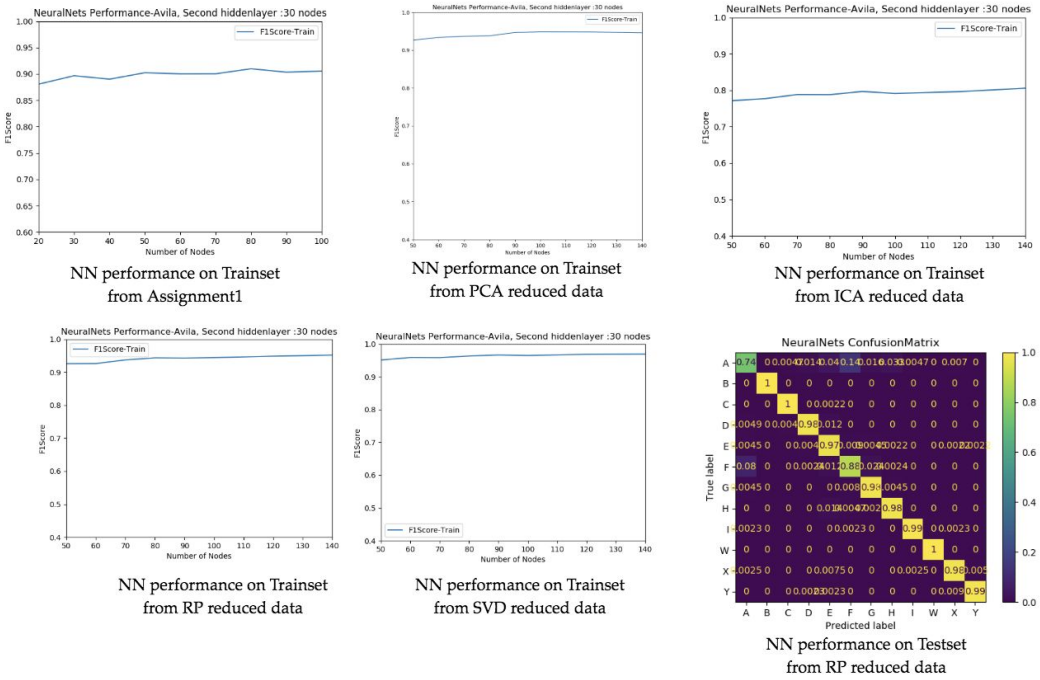

Figure 15—Training scores of Neural networks using data from various DR algorithms

In terms of speed, the time taken to train the networks using reduced data was almost the same as using full data. But performance wise, it can be observed that there is an increase in f1score in both train set and testset in all of the cases, except ICA reduced data(Test performance of other models are not shown due to space constraints). The increase in performance may be explained by the reduction of noise as a result of dimensionality reduction. The decrease in performance using ICA data is because of the loss in reconstruction observed earlier. With RP and SVD, there is a significant increase in performance. This is because both models required a large number of features to be retained in order to obtain minimum reconstruction error(9 features as compared to 10 in the original data). From the above observations, it can be inferred that even though the most efficient clustering was obtained from PCA reduced data, RP and SVD have proved efficient in reducing noise from data which inturn helps in training supervised learning models.

3.5 Neural Networks Training using clusters as features

The labels obtained from the clustering algorithms were added to the original features and the neural networks were trained using the data. In terms of speed, there wasn't any significant difference from the models trained using the original data. But in terms of performance, a significant increase (scores almost close to 1.0) was observed as shown in Figure 16. This is because of the additional information obtained from the clustering algorithms. This can be seen as another way in which unsupervised learning algorithms can be used to augment supervised learning algorithms.

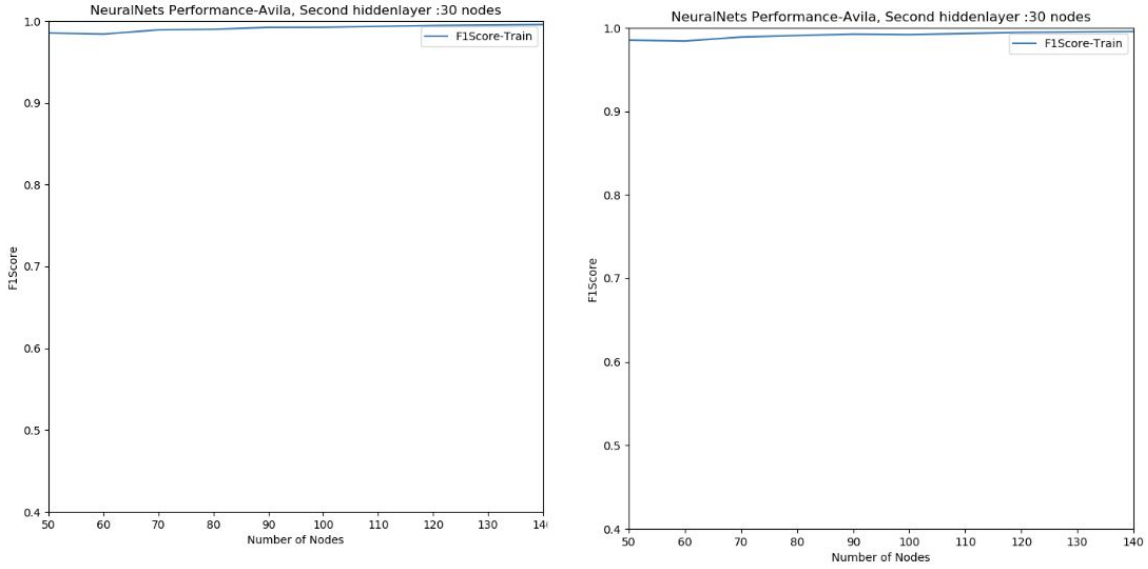

Figure 16—(Left) Performance of NN on Train Set with kmeans clustered data, (right) Performance of NN on Train Set with EM clustered data

4 References

1. Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.
2. Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake Vanderplas, Arnaud Joly, Brian Holt, Gaël Varoquaux, "API design for machine learning software: experiences from the scikit-learn project", *European Conference on Machine Learning and Principles and Practices of Knowledge Discovery in Databases*, 2013.
3. C. DeÂ Stefano, M. Maniaci, F. Fontanella, A. ScottoÂ diÂ Freca, Reliable writer identification in medieval manuscripts through page layout features: The 'Avila' Bible case, Engineering Applications of Artificial Intelligence, Volume 72, 2018, pp. 99-110.
4. Hyvärinen A., Oja E., Independent Component Analysis: Algorithms and Applications, Neural Networks Research Centre Helsinki University of Technology, 2000.
5. Isbell C.,Viola P., Restructuring Sparse High Dimensional Data for Effective Retrieval, AT&T Labs, Artificial Intelligence Laboratory, 1999.
6. Mertins, P., Mani, D., Ruggles, K. *et al.* Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* 534, 55–62 (2016). https://doi.org/10.1038/nature18003