Unsupervised Learning and Dimensionality Reduction

Exploration of unsupervised learning algorithms

Roshan Gajurel
Georgia Institute of Technology
Rgajurel3@gatech.edu

Abstract—This paper explores two clustering and four dimensionality reduction algorithms and makes comparison among them.

I. INTRODUCTION

Unsupervised learning is the is the use of various algorithms to draw inferences from unlabeled data. This paper focuses on two of the commonly used Clustering techniques:

- K-means Clustering
- Expectation Maximization (EM)

We will also explore four dimensionality reduction algorithms

- Principal Component Analysis (PCA)
- Independent Component Analysis (ICA)
- Randomized Projections (RP)
- Factor Analysis (FA)

We will explore these algorithms using two different datasets and analyze how it behaves under a variety of circumstances. Scikit Learn libraries were used exclusively to run all the analysis and graphs.

II. Data Sets

I am using two previously used datasets for this analysis, Breast Cancer Wisconsin Dataset and Credit Card Fraud Detection Dataset. The models derived from breast cancer dataset was used in the University of Wisconsin Hospitals to diagnose malignant vs benign cancerous cells [1]. This dataset consists of 569 datapoints with 32 features of which 357 are benign and 212 malignant. Credit card fraud detection dataset is a real-world dataset collected over 2 days in September 2013 from European credit card companies. This dataset contains 284, 807 data points of which only 492 are fraudulent cases. Since the dataset is highly unbalanced with fraud accounting for only .172% of the dataset. First thing we will do is make the dataset more balanced by having the ratio of fraudulent to non-fraudulent 1:10 or 5500 datapoints on different algorithms. This dataset has 31 features.

III. Clustering Algorithms

Clustering is a method of creating groupings or clusters of a dataset set so that the data within a grouping are more similar than other clusters.

A. K-Means Clustering

K-mean is an unsupervised learning algorithm that groups data into clusters. It does this by randomly selecting points in the vector space. The distances to other points around the center are calculated to find the closest points around the center. Then the center for the cluster is recalculated by averaging the clustered points. This process is repeated until it converges. Clustering algorithm tries to group the data into n clusters of equal variation by minimizing

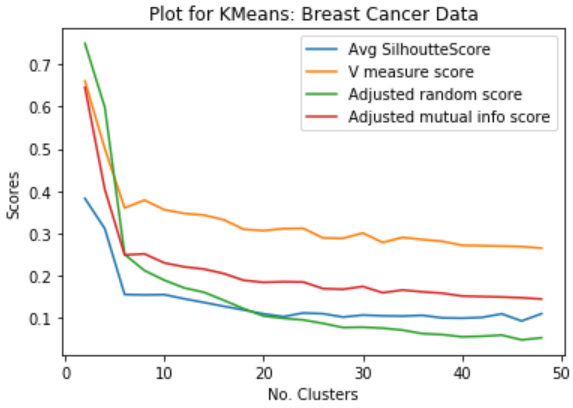
inertia or within cluster sum of squares. Some of the scoring methods that I used are:

Silhouette score: It uses the mean intra-cluster distance and mean nearest-cluster distance for each sample to calculate a value between 1 and -1. Where 1 is the best value.

V-measure: It is the harmonic mean of completeness and homogeneity. Switching the true labels from the predicted labels return the same score value. Here 1 is the perfect complete labelling.

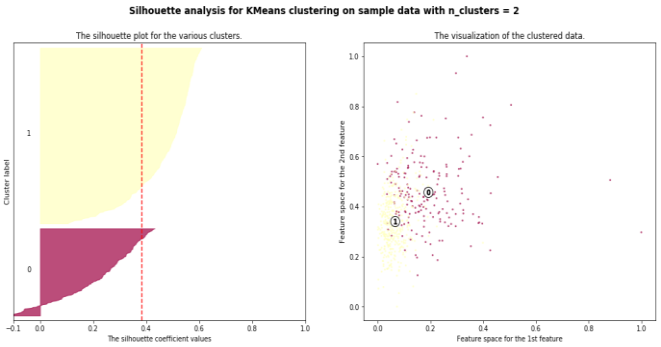
Adjusted random score: It is the similarity measure between two clustering. It is measured by taking into account the pairs of sample and counting pairs are assigned to same or different clusters in the predicted and true clustering. It has a value of 0 when the labels are random independent of the number of clusters and samples and is 1 when the clustering are identical.

Adjusted mutual info score: It adjusts the mutual information to so that it includes chance in the score. Since two clustering with large number of clusters have higher MI regardless of if more information is shared, this score takes that into account.

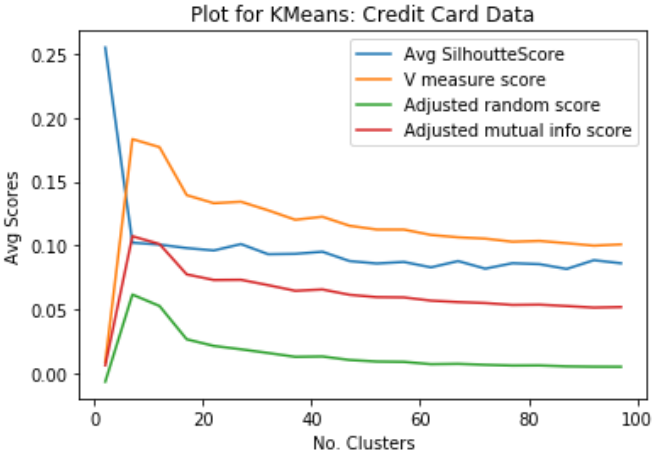


For all the scores it seems like the k-mean algorithm performs the best when the number of clusters is k=2. This makes sense because it is a binary classification problem. Increasing the number of clusters only decreases the scores.

Using 2 as the number of clusters he k-mean clustering algorithm was able to create the clusters with an accuracy of **93%** for the breast cancer data set. It took **0.12** seconds and converged in 3 iterations. The below graph shows the visualization of clusters when the number of clusters is 2. The visualization is interesting and shows how the datapoints are clustered.



Below is the graph for credit card data.



These score measures if a given cluster contains only the datapoints which are members of the target class. We can see that that the scores are better when the number of clusters is k=4 which is interesting because this is a binary classification problem as well and it is expected to produce the best clusters when the number of clusters is 2. This can be because the distribution of credit card data is not correctly linked to the output. There might be

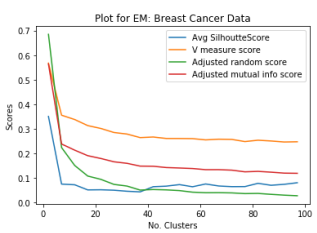
features that are causing this to happen. It will be interesting to see if we can improve this by running dimensionality reduction to remove unimportant features. Also, credit card dataset is unbalanced. We can also assume there is noise in the dataset that is causing it.

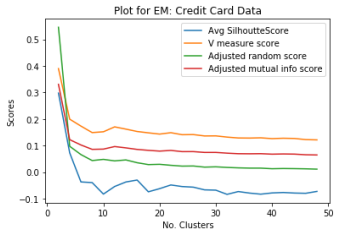
Running the credit card data with the 4 clusters I can see the accuracy of 95% and silhouette score of 0.29. It took 0.35 seconds to run. Which is more than it took for breast cancer data which is obvious because this is a larger data set. The other interesting thing is that it took 17 iteration to converge to the center which says that there is noise in the data.

B. Expectation Maximization

Expectation Maximization is an unsupervised learning algorithm that uses probability distributions can be used for clustering. It works in two phases that are expectation and maximization. Expectation probabilistically assigns z variables from centers and the z variables are sent to maximization. The maximization function does a weighted average of the data points and sends it back to the expectation function which estimates the z again. This assumes that data is going to be more and more likely to be in the cluster over iterations. During iterations it will not converge but will not diverge either. Maximum Likely hood mean of a gaussian is the mean of the data.

We will see that k-mean will produce clearer boundaries of clusters than EM because it tries to fit the data. I am using sklearns Gaussian mixture which uses expectation maximization algorithm for fitting.



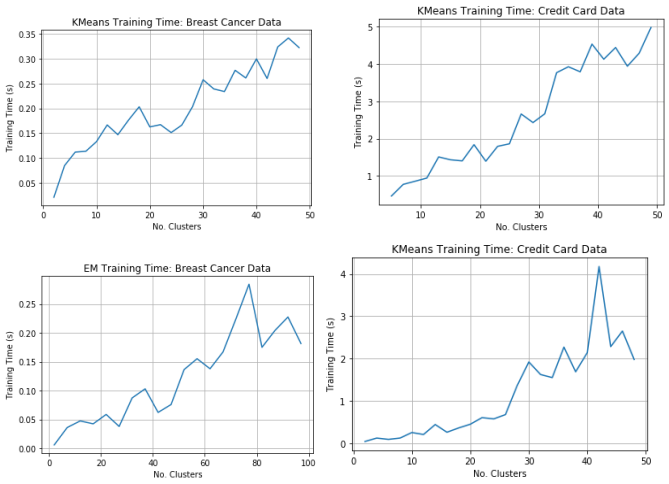


As we can see that the Sihouellte score for both breast cancer and credit card data are very similar. As the number of clusters increases, the socre decreases. We can see here that the best score is when the number of cluster is k=2. It is true for Credit card data as well.

This is interesting because k-mean did not form the best cluster for credit card data at k=2 clusters. We can see that EM does better for credit card data where the data is more unbalanced and might be noisy. I if made the dataset more balenced, we would see increase in performance.

If we look at the model training time, EM is significantly faster than k-mean. It takes around 0.01 seconds for breast cancer data and .02 for Credit card data. The accuracy for breast cancer data was .95 and .92 for credit card data which is not good as k-means. This could be because there could be datapoints on the boundaries which are not clustered appropriately. Since k-means completely fits the data into clusters and EM does not. If we compare the silhouette scores to k-means, we see that EM did better for credit card dataset and worst for breast cancer data.

Below graphs compare the training times of the datasets for both algorithms.

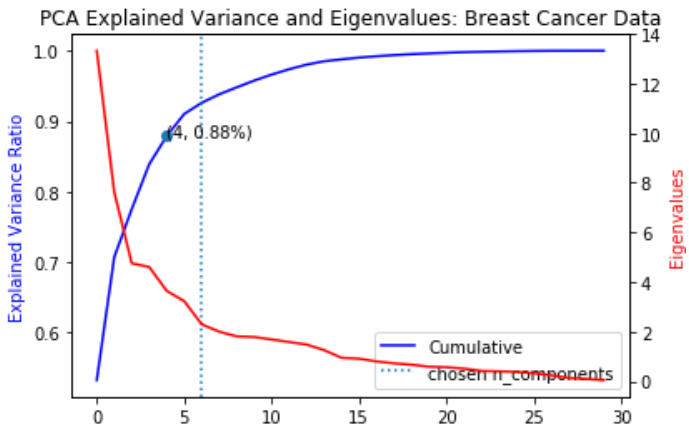


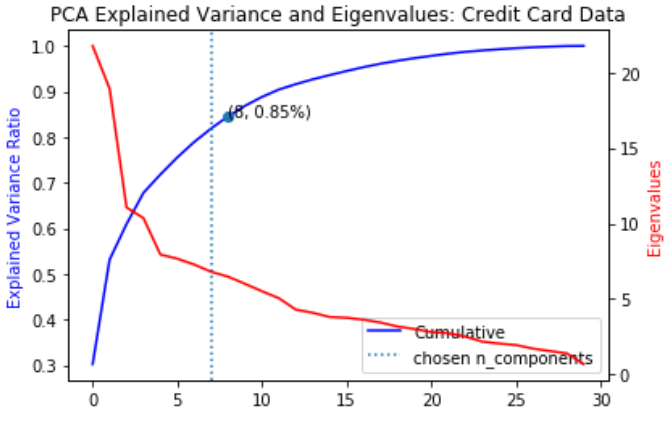
We can see that EM is faster than k-means for larger datasets.

IV. Dimensionality Reduction

Dimensionality Reduction is performing pre-processing on the data to remove the random variables and keep principal variables. Some dimensionality reduction methods that I have used are listed below.

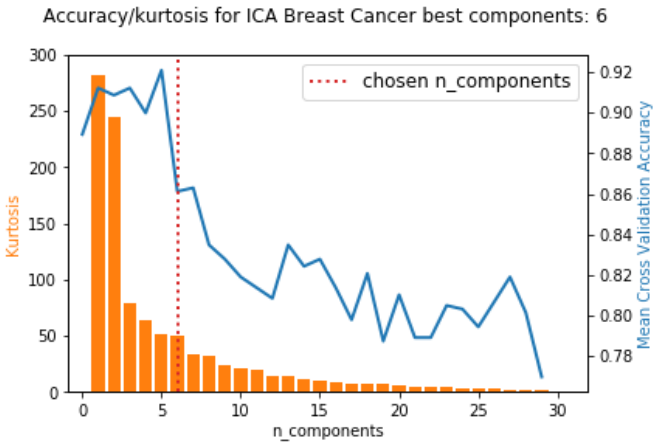
Principal Component Analysis: PCA is a method of linear dimensionality reduction which uses eigenvalue value decomposition of the data to a lower space. It does this by projecting the data to axes which maximizes variance. The components of the PCA are in order of highest variance to lower variance.

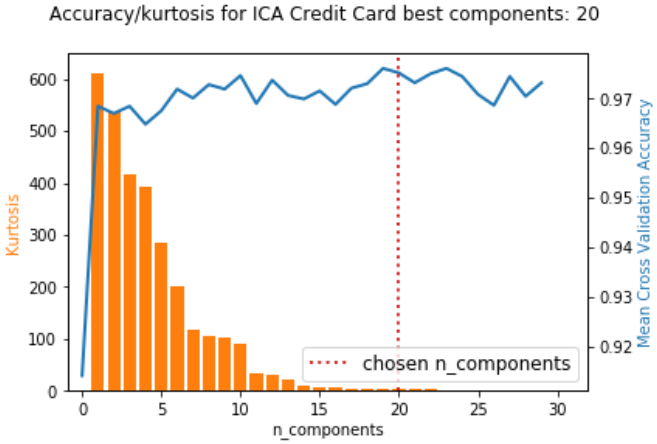




Above graphs show the results of running PCA for both breast cancer data and credit card data. We can see that the variance of the cumulative data increases for higher number of components. We can see that 4 is the optimal number of components that gets 88% or more of the variance for credit card data and 8 components has a variance of 85% for credit card data. We can also see the dotted blue line which has the best cross-validation accuracy of a decision tree for the number of components returned after dimensionality reduction. It is interesting to note that the line is very close to the optimal number of clusters that maximizes the variance with the least number of clusters. Eigenvalues for both datasets have similar curves. These determine the variance of the data on the new feature axes. The eigenvectors with lower eigen values carry the least information and thus can be dropped. In the case of breast cancer data, the features after 10 components and around 15 components can be dropped because they carry very little information. It is interesting to note that only 4 out of 32 components contain the most data for breast cancer data and 8 out of 31 for Credit card data.

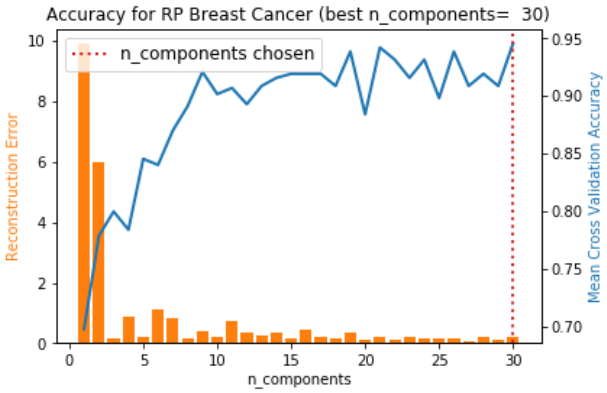
Independent Component Analysis: ICA is another dimension reduction method which separates multivariate signal into subcomponents. It is different from PCA in that features are projected onto axes which reconstruct the features, and these are maximally independent.

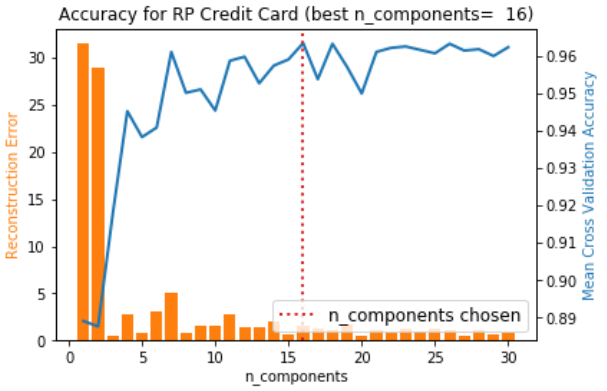




The above graphs show the results of running ICA on both data sets. Kurtosis, 4th normalized central moment, for the number of components is plotted in orange. Kurtosis can be used to measure the mutual independence between components. More mutually independent components carry more information. As we can see that the first 2 components of the breast cancer data are completely mutually independent and this decreases as the number of components increases. Similar trend can be seen for credit card data. However, the interesting thing about this graph is that the cross-validation accuracy of breast cancer data is the best when for the first two independent features and it goes down. It can be observed that it is good to just have these components for breast cancer dataset. However, for credit card data even though the first two components have the highest kurtosis, we can see that the cross-validation accuracy goes up with the number of components. Kurtosis may not be the best measure for dimensionality reduction for credit card data set when using decision tree as the classifier. However, it seems to work well for breast cancer which is a smaller dataset with less noise and is more balanced.

Random Projection: RP is a method that reduces the dimensionality of data by projecting to axes but instead of using maximum variance as in PCA it uses random axes.

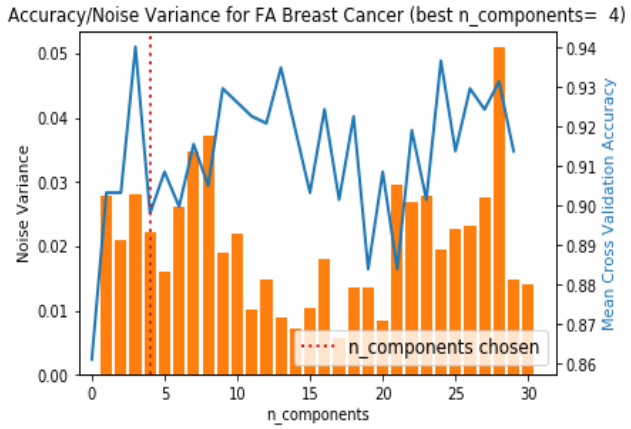


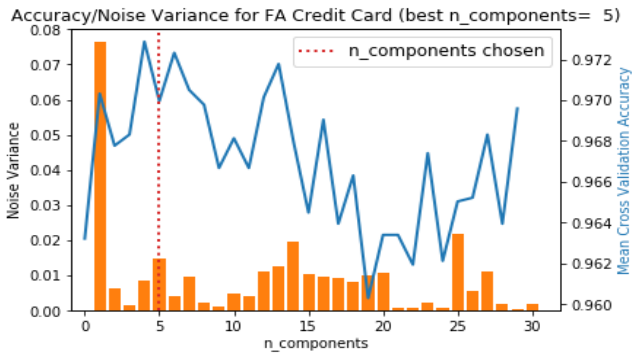


In the above graph, the orange bar denotes the reconstruction error of the components. It was selected by averaging at the reconstruction error with iterative restarts equal to the number of components. I did not see much variance when I reran RP multiple times. We can see that the error for the first 2 components is very high but after that the error decreases. The minimum number of components that can minimize the reconstruction

error can be selected as the reduced dimension. In the case of both datasets the error decreases substantially after the first two components. However, it is interesting to note that the best number of components is 30 for breast cancer data which is just 2 less than all the components. The cross-validation score verifies this as well since that is when the score is highest. It is interesting to see 16 as the best n_components for credit card data even though it is more complex and imbalanced compared to breast cancer data.

Factor Analysis: The general theory for FA is that by determining the interdependence of different components, we can later reduce the set of variables in the dataset. FA is a method of linear transformation of lower dimensional factors which are distributed according to a gaussian with zero mean and zero covariance. It is useful to determine the smaller number of latent variables that are important.





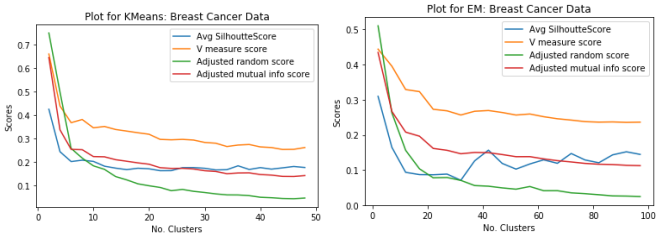
Graphs for both datasets can be seen above. We can see that the number of components is 4 and 5 for breast cancer and credit card data respectively. Which is a huge improvement compared to RP. 5 is the least predicted components for credit card data yet. We can see that the cross-validation score when running a decision tree for the selected features is maximum around the best n_component. We can also see that the change in noise variance is less in the breast cancer dataset compared to credit card. The first component of the credit card dataset has the highest noise and others have fewer, compared to breast cancer data where the noise is almost uniform. It could be because of the noisy and unbalanced nature of credit card dataset.

V. Clustering after Dimensionality Reduction

In this section we will run both the clustering algorithms using the datasets that we have already run dimension reduction on already.

A. K-means and EM clustering after PCA
   a. Breast Cancer Data

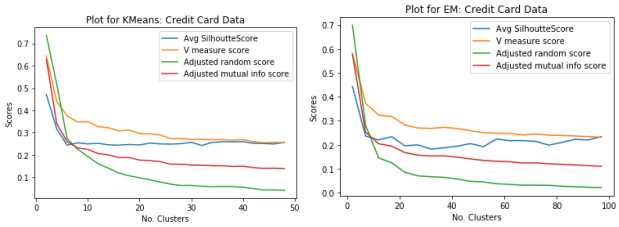
After applying PCA dimensionality reduction to the breast cancer dataset, there were 5 features that were selected. Below is the graph of running K-mean and EM on the data.



We see very similar graphs that we saw as before but if we notice the Silhouette score, and others, for k-mean it is better than the before. It's is true for EM as well. For breast cancer dataset it can be concluded that by reducing the number of features using PCA, the resulting clusters are categorized more accurately.

b. Credit Card Data

For credit card 8 features was selected as a result of PCA.



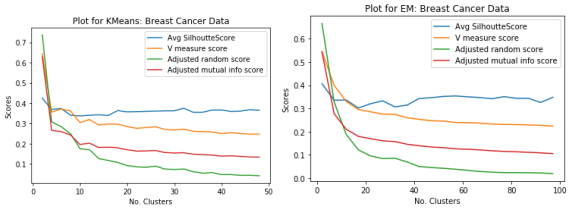
We can see significant improvement in modeling time as well. Since the number of features were reduced the time taken to model the credit card data for k-means reduced from 0.07 to 0.01 and for EM which is already rather fast it reduced from 0.02 to less than 0.01.

Also, if we investigate the scores, they are significantly better as well. Removing the features with less information has significantly improved the clusters.

B. K-means and EM clustering after ICA

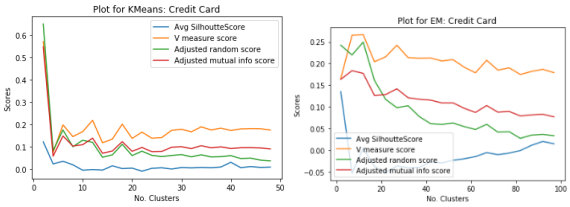
a. Breast Cancer Data

6 features were selected as result of ICA for K-means and EM. The outcome of the graphs can be see below. It is interesting to note that the scores of the for 2 clusters seem almost the same as before. However, the accuracy has dropped to .93 for k-means and to .72 for EM. Like before there is significant improvement in the speed of the algorithm because of the decrease in the number of dimensions.



b. Credit Card Data

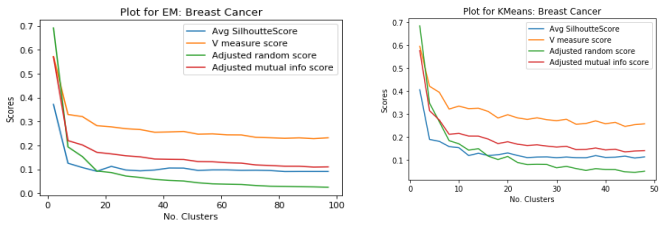
20 Features were was selected for credit card data as a result of ICA.



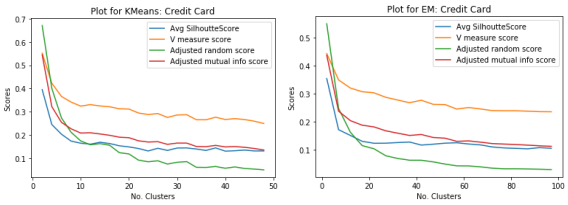
We can see that the average Silhouette score has dropped significantly after ICA for k-means. However, the adjusted random scores and v measure score remain almost unchanged. It gets more interesting for EM. The v measure score has the best accuracy around 5 clusters when the Silhouette score is the lowest. We also see that the accuracy as decreased significantly to 0.84 and 0.72 k-mean and EM respectively. However, it is faster as always. We can conclude that for an unbalanced dataset like Credit Card ICA does not perform as good as PCA.

C. K-means and EM clustering after RP

All components were selected from RP for K-mean and EM this is not very interesting because it kept almost all the features. Thus, we cannot see any change in this graph compared to the results of just running clustering.



For Credit card Data sets, after RP, credit card data was reduced to 16 features.

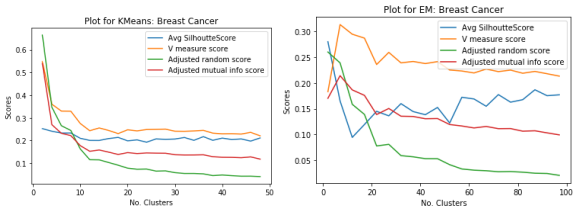


Comparing the outcome to previous graphs we can see that it does better than ICA but worst than PCA. We see that the K-means score are similar to after running PCA but the EM scores are worst. Since it had higher features it is slower than PCA and ICA as well.

D. K-means and EM clustering after FA

a. Breast Cancer Data

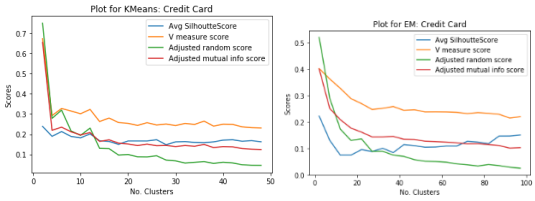
4 features were selected after FA for Breast cancer data.



We can see that the scores for the clusters h ave reduced significantly compared to PCA a nd the accuracy as also dropped to .72 and . 74 respectively for k-mean and EM. We see that v-measure has the highest score for ar ound 5 clusters.

b. Credit Card Data

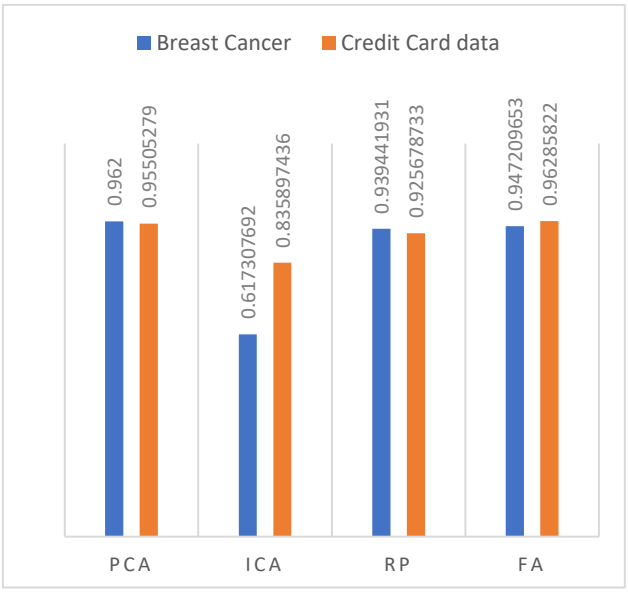
5 Features were selected after FA for Credit card data.



The above graph shows that although the Silhouette scores have dropped, v-measure score and adjusted random scores have improved compared to clustering on the entire dataset.

VI. NN after Dimensionality Reduction

In in section I applied four dimensional reduction algorithm to both data types and used the transformed dataset to create a neural network model. I ran 10-fold cross-validation on the data for each algorithm using the best hyper-parameters that I had found from the first paper.

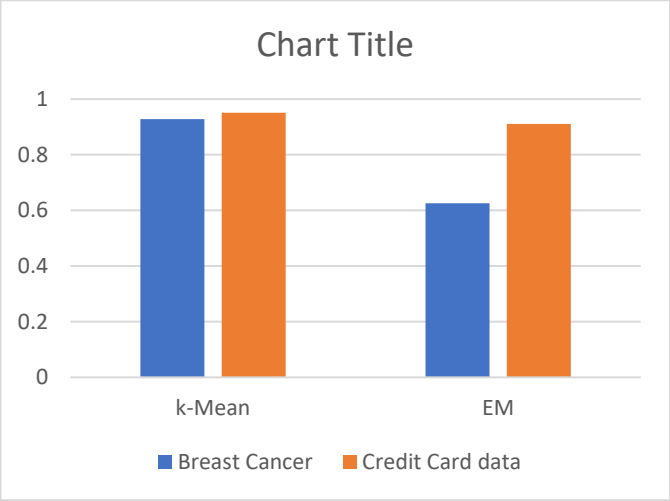


We see that PCA performs the best for breast cancer data and FA performs the best for Credit Card data. The 10-fold cross validation scores for Breast cancer data has improved significantly. Cross validation score for breast cancer data for my first paper was 92% which has increased to 96%. Also, since PCA only used 5 features to run this it was significantly faster. It is interesting to see that more data is not always better and by selecting only the important features we can increase the accuracy. During my first paper credit card data had a cross-validation score of 98%. FA comes close to that with 96% and does it with just 5 features. This significantly increases the speed of the algorithm. ICA does not perform as well, which was expected because it was not able select the features with maximum independence.

VII. NN after Clustering

In this section, I ran k-mean and Expectation Maximization clustering methods to both the datasets and used the best number of

clusters and transformed data to create a neural network modal. The 10-fold cross validation scores of the model can be seen below.



Breast cancer data does well after k-mean with a score of 92% which equals the baseline from my first paper for this dataset. Credit card data does well as well as it did after dimensionality reduction. The interesting thing is breast cancer model after EM performs really bad with the accuracy of just 62%. This could be because the

IV. Conclusion

We analyzed the breast cancer data set and credit card fraud detection dataset on 2 clustering and 4 dimensionality reduction algorithms. Both methods can be used to remove the data without information or least information. EM uses probabilistic methods to capture the maximum likelihood in contrast to k-means which finds the means. The purpose of dimensionality reduction is to reduce the number of features and prevent the curse of dimensionality. All the algorithms have pros and cons, some are faster than other however the main thing is they perform different on different datasets. So which method to use should be determined by observing the data.

For our breast cancer dataset, the best score for neural network was after PCA and FA for credit card dataset. Using unsupervised learning methods can to explore the data before learning from them we can significantly increase the accuracy as well as speed of the resulting models.

References

[1] Street, W. Nick, William H. Wolberg, and Olvi L. Mangasarian. "Nuclear feature extraction for breast tumor diagnosis." Biomedical Image Processing and Biomedical Visualization. Vol. 1905. International Society for Optics and Photonics, 1993.