CS 7641 Unsupervised Learning

Hyesung Ji (hji61@gatech.edu)

Abstract— The unsupervised learning is a kind of Machine Learning technique. Compared to the supervised learning, it does not have the labeled data. It is to figure out the structure of data and clusters unlabeled dataset. Thus, it usually used for understanding given dataset. In this project, the K-means and Gaussian Mixture algorithms are studied before and after dimensionality reduction. For dimensionality reduction, PCA, ICA, Random Projection, and Locally Linear Embedding algorithms are used. Also, Neural Network is applied the dataset to investigate the effects of dimensionality algorithms and clustering algorithms.

I. INTRODUCTION

The unsupervised learning is used to cluster and analysis given dataset. To compare before and after dimension reduction, four dimension reduction techniques are used in this project. As it is widely known, dimension reduction can be more effective for the K-means method to cluster dataset [1]. Therefore, how effective the dimensionality reduction algorithm is analyzed in this project in terms of quantitative and qualitative factors. For this analysis, two datasets(Breast Cancer data and Wifi localization data) are used [2, 3]. The Breast Cancer dataset (B-cancer) has 30 features and the Wifi datset (Wifi) has 7 features. The B-cancer dataset has a various range of values and the Wifi dataset has a similar range of values.

A. *Principle Components Analysis(PCA)*

The PCA is widely used for dimension reduction. It can reduce the dimension of a given dataset by calculating each eigenvalue for each feature in data [4]. It also can make the data less complex and reduce noise of the data [4].

B. *Independence Components Analysis (ICA)*

The ICA algorithm computes the 'non-gaussianity' of a given data [5]. The points of the data will be projected to a new axes on higher dimension and the ICA will find the maximize non-gaussianity (kurtosis) of the data [5]. The maximum non-gaussianity can be the number of components.

C. *Random Projection (RP)*

The RP is also widely used for dimension reduction. This algorithm is to map a higher dimension vector to a lower dimension vector by using 'Johnson-Linderstrauss' lemma [6]. This algorithm is known that it can preserve the information of the original data [6]. In this project, the Gaussian Random Projection is used.

D. *Locally Linear Embedding (LLE)*

The LLE is a kind of unsupervised learning that can preserve neighborhood. It calculates and corresponds from nonlinear structured data (normally high dimensional data) into linear grid (lower dimensional data) [7].

II. PART-1: K-MEANS AND EXPECTATION MAXIMIZATION(EM) WITHOUT DIMENSIONALITY REDUCTION

As we discussed in the introduction, two unsupervised learning algorithms (K-Means and EM) are applied to two datasets.

A. *K-Means*

For K-Means algorithm, the number of clusters is the most important parameter. Normally, to determine of the number of clusters, it is necessary to minimize the sum squared error (SSE) of distance between each point and center point(centroid) in a cluster [8]. The python library function provides 'inertia' to calculate the SSE. After calculating the SSE plot according to the number of clusters, the 'Elbow' method can be applied [8]. There can be a convex point that the SSE changes from drastically decreasing to slowly decreasing. In this case, the number of clusters corresponding that point can be an optimal number of clusters [8]. Additionally, this quantitative may not be a perfect way to compute the optimal cluster number. So, the qualitative method is added to this project. The shape of silhouette is added to find optimal number of clusters.

1) *Breast Cancer*

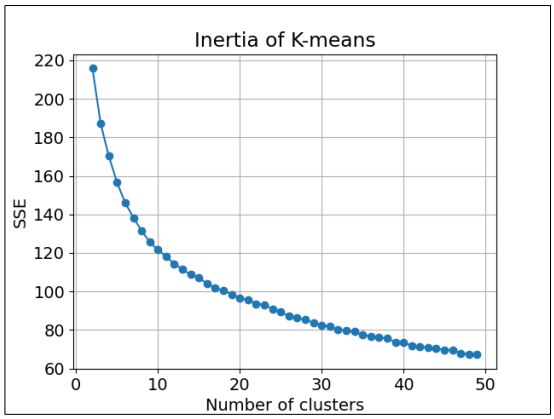

Figure 1. Inertial of K-Means

In Figure 1, the 'elbow' point seems to be near by 7 to 10. However, it is not clearly seen in this figure. So, the shape of silhouette and centroids should be investigated.

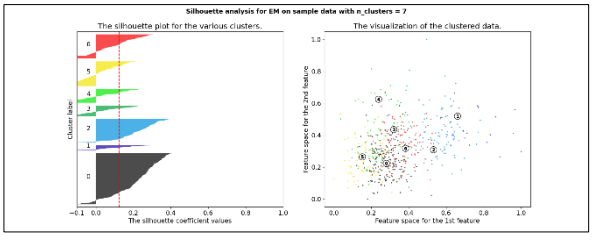

Figure 2-1. Silhouette and centroids when K=7

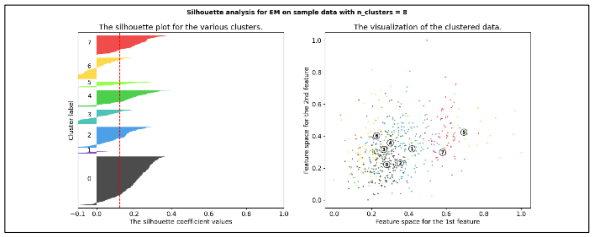

Figure 2-2. Silhouette and centroids when K=8

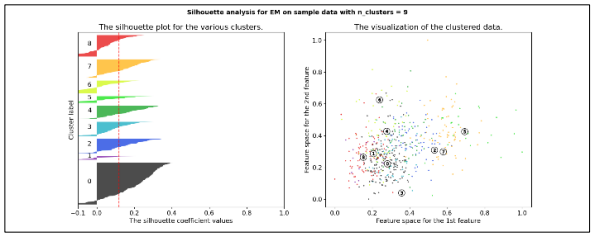

Figure 2-3. Silhouette and centroids when K=9

From Figure 2-1 to 2-3, we can see the silhouette has the best figure when K is 7. All clusters are higher than the average score (Red dashed line) and the distances between centroids are proper, not overlapped.

2) Wifi-Localization

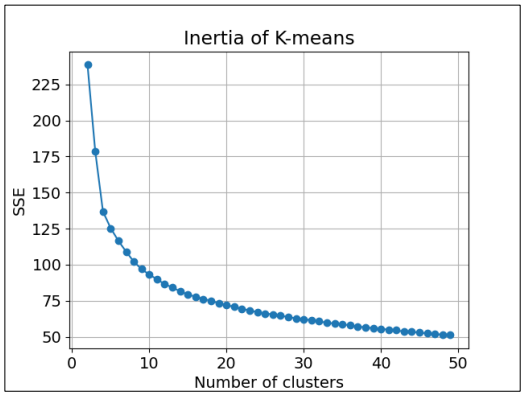

Figure 3. Inertial of K-Means

In this case, the elbow point can be from 4 to 6. The silhouette shape and centroids are calculated.

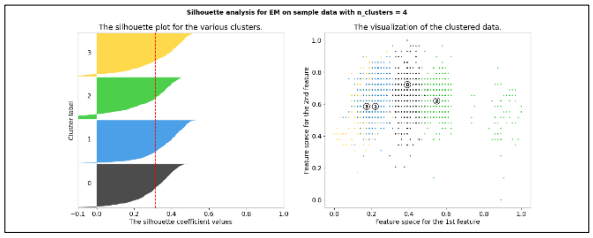

Figure 4-1. Silhouette and centroids when K=4

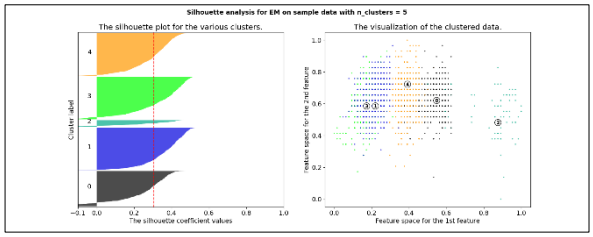

Figure 4-2. Silhouette and centroids when K=5

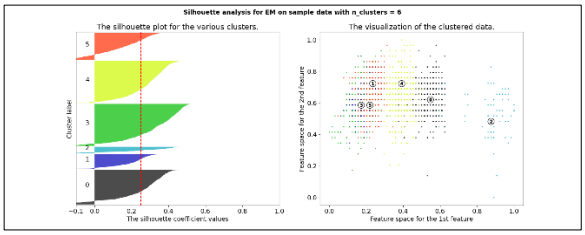

Figure 4-3. Silhouette and centroids when K=6

From Figure 4-1 to 4-3, when the K is 4, the shape of silhouette seems to be proper. The shapes of other cases are irregular.

B. Expectation Maximiztion (EM)

For EM algorithm, the number of components is also the most important parameter. There are some ways to determine of the number of components. In this project, BIC score was selected to determine the number of components. Normally, the smallest value is selected as an optimal number of components. Also, as the K-Means method, the qualitative method is analyzed.

1) Breast Cancer

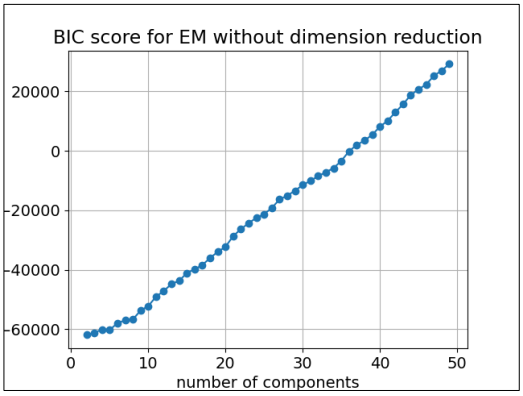

Figure 5. BIC score of EM

In this Figure, the number of components is considered from 2 to 4.

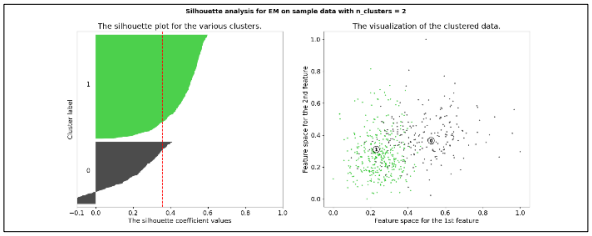

Figure 6-1. Silhouette and centroids when K=2

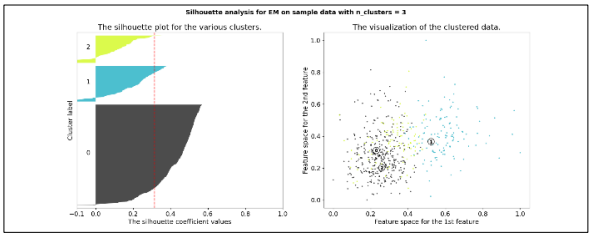

Figure 6-2. Silhouette and centroids when K=3

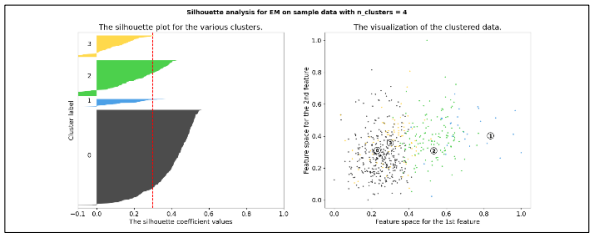

Figure 6-3. Silhouette and centroids when K=4

Among these figure, the number of components can be 2 considering the shape of silhouette and centroids.

2) Wifi-Localization

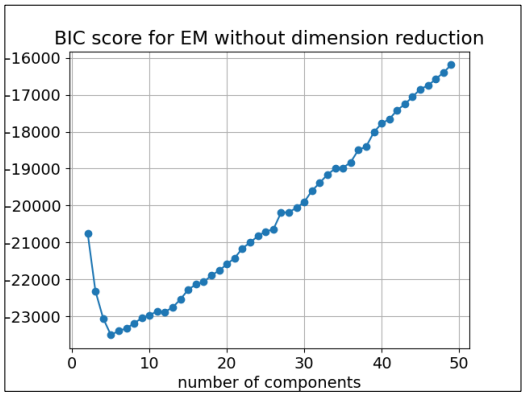

Figure 7. BIC score of EM

The smallest number is 5 and the number of components is considered from 4 to 6.

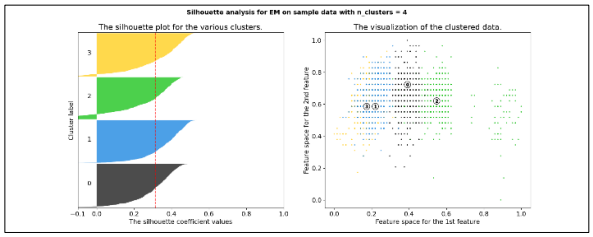

Figure 8-1. Silhouette and centroids when K=4

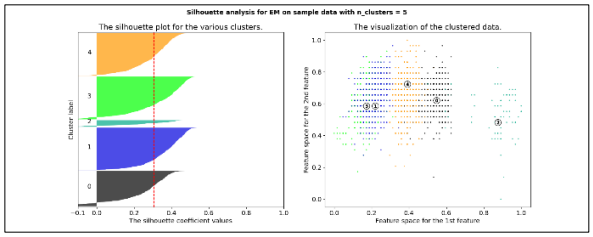

Figure 8-2. Silhouette and centroids when K=5

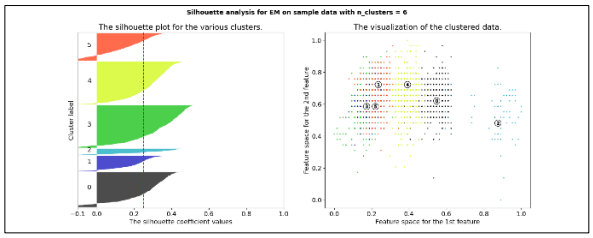

Figure 8-3. Silhouette and centroids when K=6

Considering the shape of silhouette and centroids, the number of cluster can be 4.

III. PART-2: FOUR DIMENSIONALITY REDUCTION ALGORITHMS

In this section, the four dimensionality reduction algorithms (PCA, ICA, RP, and LLE) are applied to the original data prior to the clustering algorithms to be applied. And then, the effect of each dimensionality reduction algorithm is investigated.

A. PCA

For PCA, the most important parameter is the number of components. Also, to determine the number of components, the individual and cumulative variance are normally considered. Because, a eigenvalues can represent a variance value [9]. A variances projected on the first some principal components can be expressed as the sum of the eigenvalues [9]. It means that the higher variance a component has, the higher eigenvalue it has. In this project, 90% of the cumulative variance and the individual variance which is greater than 5% were considered simultaneously.

1) Breast Cancer

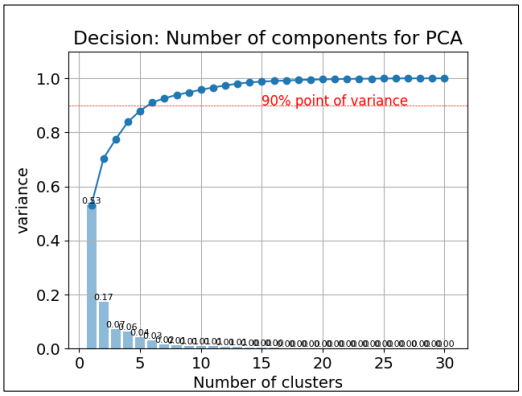

Figure 9. PCA variance of B-cancer dataset

In Figure 9, when the number of components is greater than 6, the cumulative variance is higher than 90%. However, the variances of numbers 1 to 4 are greater than 5%, so the number of components is selected as 4.

2) Wifi-Localization

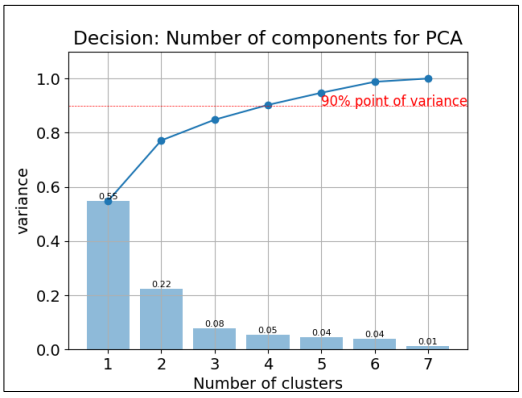

Figure 10. PCA variance of Wifi dataset

Similarly to the previous figure, the variances of numbers 1 to 3 is higher than 5%. So, the number of components is selected as 3.

B. ICA

1) Breast Cancer

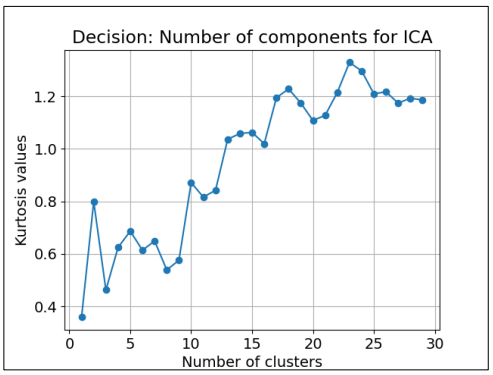

Figure 11. Fitness function for different restarts in RHC

In Figure 11, the highest value is 23 and it is selected as the number of components in ICA.

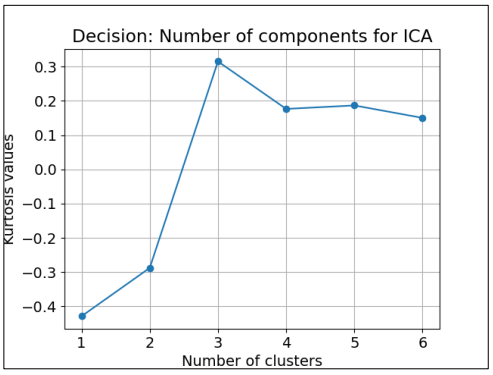

Figure 12. Kurtosis of ICA components

In Figure 12, the number of components is 3.

C. Random Projection(RP)

In this project, the Gaussian Random Projection was used. Also to avoid biased results, the reconstructed error was calculated 100 times and averaged. The standard deviation is plotted as a blue color.

1) Breast Cancer

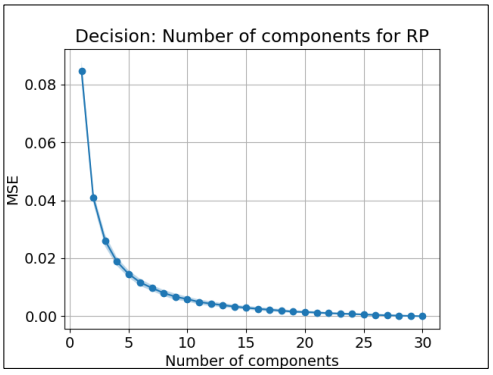

Figure 13. Reconstructed error of RP for B-cancer dataset

Considering the Elbow method, 5 can be considered as the number of components for Gaussian random projection in Figure 13.

2) Wifi-Localization

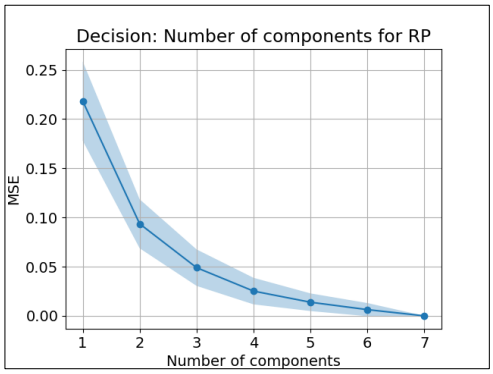

Figure 14. Reconstructed error of RP for Wifi dataset

As we can see Figure 14, 3 can be considered as the number of components for Gaussian random projection.

D. Locally Linear Embedding (LLE)

For LLE, the number of neighbors and the number of components are the most important parameters to be considered.

1) Breast Cancer

To determine the number of neighbors and components, two plots were computed. First, considering the number of components in the previous PCA experiments, the number of components is set to 7 temporarily.

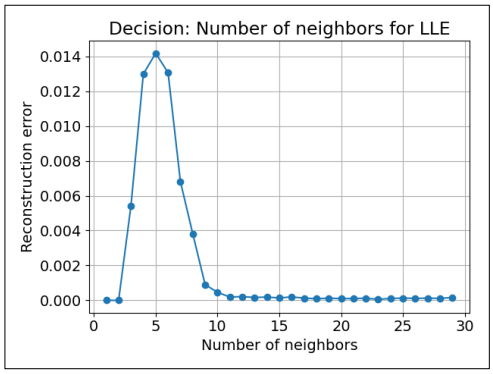

Figure 15-1. Reconstructed error of LLE for B-cancer dataset

In Figure 15-1, there is a very interesting plot. Except for 3 to 9, other error values are very closed to zero. After 9, the error plot is stabilized, so 11 is selected as a number of neighbors. After the number of neighbors is set to 11, the number of components is calculated.

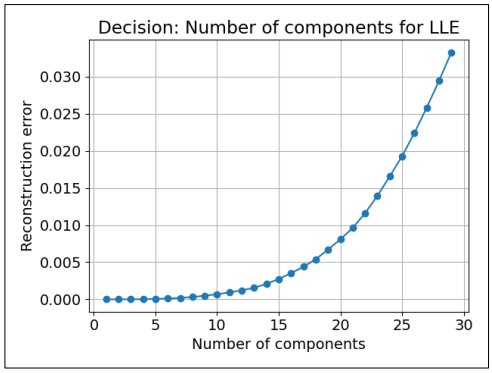

Figure 15-2. Reconstructed error of LLE for B-cancer dataset

In Figure 15-2, considering the Elbow method, the number of components is determined as 17.

2) Wifi-Localization

Like the B-cancer dataset, the number of components is initially set to 3.

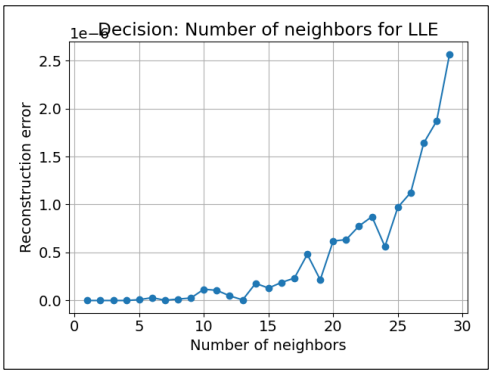

Figure 16-1. Reconstructed error of LLE for wifi dataset

In Figure 19, the Elbow point was selected as 19.

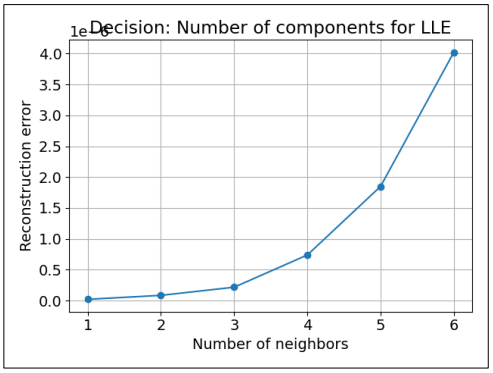

Figure 16-2. Reconstructed error of LLE for wifi dataset

As the number of components, 4 can be the number of components.

IV. PART-3: K-MEANS AND EM ALGORITHM AFTER DIMENSIONALITY REDUCTION

In this section, the unsupervised learning method (K-Means and EM) is applied the processed data by dimensionality reduction algorithms. After that the results will be compared to the results from Section II.(Part-1) method. Each algorithm is applied to two datasets simultaneously. In this section, the qualitative analysis using the silhouette shape and centroids is not implemented. The best performance of these methods is compared to the performance of Part-1.

A. K-Means (KM)

1) PCA

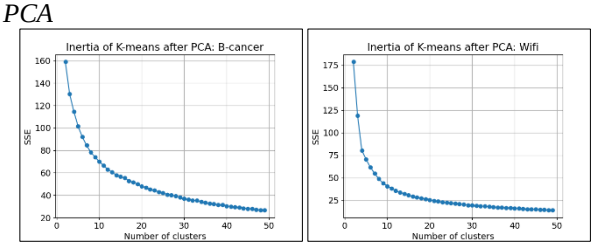

Figure 17. Inertial of K-Means after PCA

Compared to Figure 1 and Figure 3, the overall shapes are exactly the same as the Figure 1 and Figure 3. The only discrepancy is that the values of inertia are decreased. So, the number of clusters are the same as the previous case (no dimension-reduction cases).

2) ICA

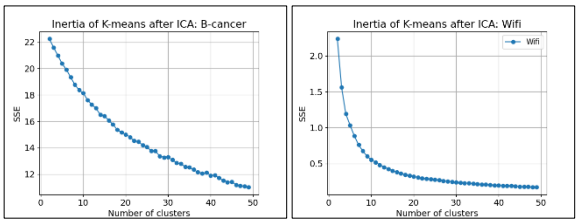

Figure 18. Inertial of K-Means after ICA

Figure 18 shows that the left shape(B-cancer) is changed a lot and the right shape (Wifi) is very similar to the previous case. It is difficult to find the number of clusters for B-cancer case due to its linear shape, not elbow shape. Despite of the shape, the number of clusters is set to 15.

3) Random Projection(RP)

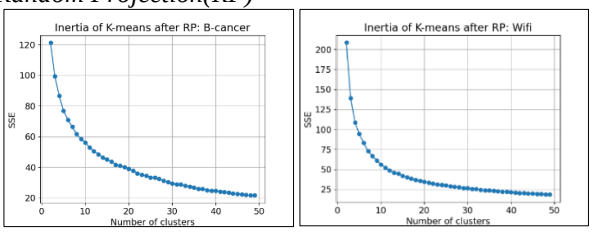

Figure 19. Inertial of K-Means after RP

As we can see in Figure 19, the number of clusters for K-Means is the exactly same as the previous case.

4) Locally Linear Embedding(LLE)

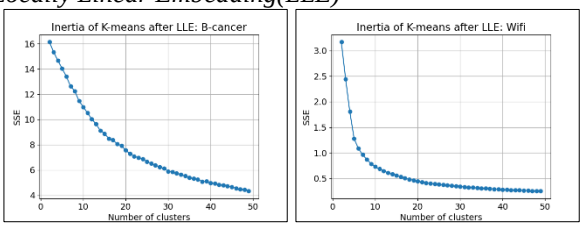

Figure 20. Inertial of K-Means after LLE

The shapes of inertia is similar to the ICA case. However, the overall inertia values are lower than the ICA cases. So, the number of clusters for B-cancer is 15, but the number of cancers for Wifi cannot be 4 like the previous cases. Because, the discrepancy between 4 and 5 is large, so 5 can be the number of clusters in this case.

B. Expectation Maximiztion (EM)

1) PCA

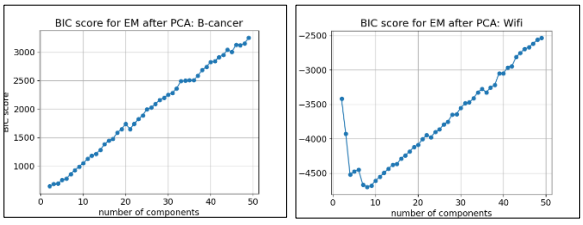

Figure 21. BIC score of EM after PCA

In Figure 21, the overall shapes of BIC score is similar to the Figure 5 and 7. The overall absolute value of BIC scores are significantly lower than the previous cases. In the case of Wifi dataset, the number of component can be 8.

2) ICA

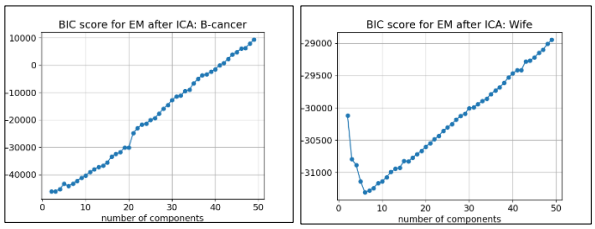

Figure 22. BIC score of EM after ICA

Figure 22 shows that the overall shape of BIC score is similar to the PCA case. The number of components for B-cancer can be 2 and the number of components for Wifi can be 6.

3) Random Projection(RP)

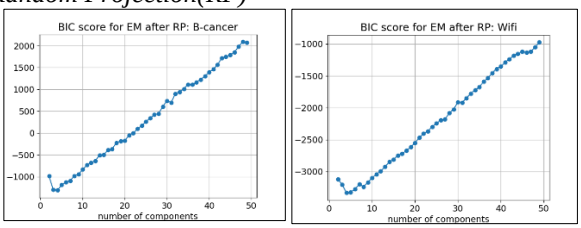

Figure 23. BIC score of EM after RP

As we can see in Figure 23, overall shapes of BIC score is changed and the overall absolute value of BIC scores are significantly lower than the previous cases. The number of components for both cases can be 4.

4) Locally Linear Embedding(LLE)

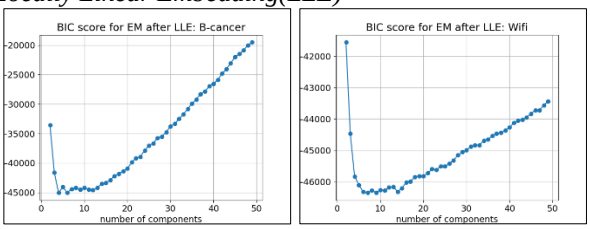

Figure 24. BIC score of EM after LLE

In this Figure, the shapes of BIC score are quite different from the previous cases and the lowest values of BIC scores are 4 and 7 respectively.

C. Comparison and analysis

The results from part 3 and part 1 are compared to analyze the difference the clustering no dimensionality reduction and the clustering after dimensionality reduction.

1) Comparison (D.R. means Dimensionality Reduction.)

Method	Data	K-means			EM		
		Silhouette score	Complete score	# of cluster	Silhouette score	Complete score	# of component
No D.R.	B-Cancer	0.1574	0.2734	7	0.3561	0.6676	2
	Wifi	0.3290	0.8467	4	0.3068	0.8714	4
PCA	B-Cancer	0.2157	0.2559	7	0.3852	0.5055	2
	Wifi	0.4335	0.8244	4	0.3393	0.6306	8
ICA	B-Cancer	0.0214	0.1312	15	0.1417	0.0290	2
	Wifi	0.3561	0.5897	4	0.2786	0.6174	6
RP	B-Cancer	0.2053	0.1953	7	0.2879	0.3017	4
	Wifi	0.3533	0.5439	4	0.3174	0.5808	4
LLE	B-Cancer	0.1191	0.1776	15	0.0895	0.4106	4
	Wifi	0.4021	0.5759	5	0.1991	0.6270	7

Table 1. The results from no D.R. and each D.R. method

According to the table 1, even though the dimensionality reduction was applied to the original data, the silhouette and complete score are not improved significantly. And the numbers of clusters/components tend to be increased than the no. D.R. method. A possible reason is that the data 'filtered' by D.R. has relatively significant information (such as higher variance and kurtosis, etc). So, the clustering algorithm is likely to recognize more clusters/components are needed. Relevant to the number of clusters/components, the scores of the ICA and LLE were worse than the no D.R clustering. In these cases, the number of clusters/components are increased. On the other hand, the interesting result is the clustering after PCA. In this case, the number of clusters/components is increased only for just EM algorithm and Wifi data. The silhouette scores are improved than the no D.R. clustering.

2) Analysis

Why the result of clustering after PCA is better than other results? As discussed above, a possible reason is the number of clusters/components was not changed much and it can improve the silhouette score. In the case of other dimensionality algorithms such as ICA and LLE, the numbers of clusters/components are more changed than other cases, and it may result in a decrease in silhouette score. The more clusters/components increase, the worse the accuracy of labels of clustering can be. Because, the B-cancer data is binary classification and Wifi data is 4-class classification. So, these datasets have two and four labels respectively. If the number of clusters/components increases, the labels would not match to the original labels. This can cause the worse completeness score and silhouette score. So, the quantitative scores may not be perfect metrics to estimate the performance of clustering algorithm. The qualitative analysis should also be considered. The silhouette shape of clustering after PCA is the following.

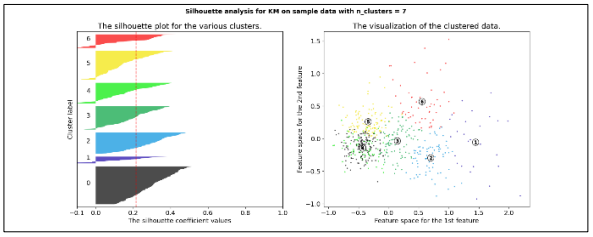

Figure 25-1. KM Silhouette and centroids when K=7 for B-cancer

Compared to Figure 2-1, the data groups are relatively divided well. The distance between centroids is more

visible. And the shapes of silhouettes are more similar to each other.

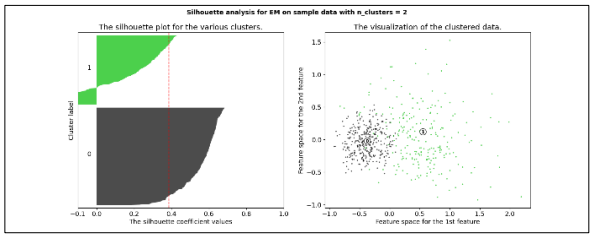

Figure 25-2. EM Silhouette and centroids when K=7 for B-cancer

Compared to Figure 6-1, the shape of silhouette is pretty similar and the silhouette score is slightly increased. It means that the distance between label 0 and label is increased.

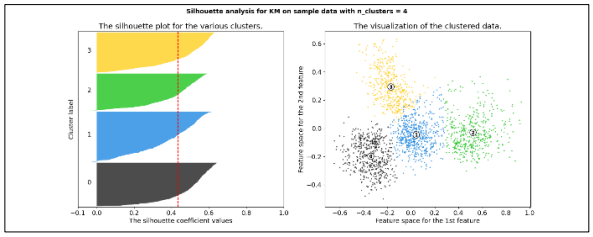

Figure 25-3. KM Silhouette and centroids when K=4 for Wifi

Compared to Figure 4-1, the shape of the silhouette is more even. Also, the silhouette score is improved. Especially, the distribution of dataset is significantly changed. In Figure 6-1, the data is just aligned as a column, but the data distribution of Figure 25-3 is much better divided to be grouped.

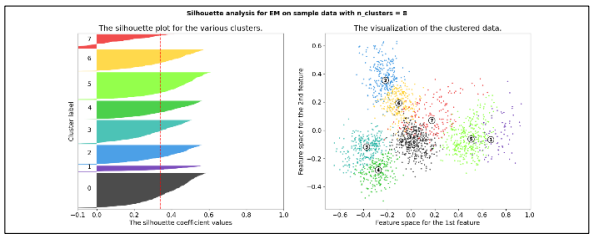

Figure 25-4. EM Silhouette and centroids when K=8 for Wifi

Compared to Figure 8-1 to 3, the number of centroids is greater, however, the distance between centroid is more even than the previous cases. Also, the data is well-distributed.

In summary, after the PCA process, some low-variance features are eliminated. So, the meaningful data is remained and is distributed relatively better than the original data. As a result, the clustering with PCA method becomes a better algorithm than the clustering without dimensionality reduction.

V. PART4: NEURAL NETWORK AFTER DIMENSIONALITY REDUCTION

In this section, the Neural Network is applied to the data which was processed by each dimensionality reduction algorithm. And then the effect of each dimensionality reduction algorithm is investigated in terms of the learning curve and loss curve. Also, each accuracy and time to train is compared to each other. For this work, the dataset Wifi-localization is selected because its structure is simpler than the B-cancer dataset, so it can represent the effect of characteristic from each algorithm.

A. *PCA*

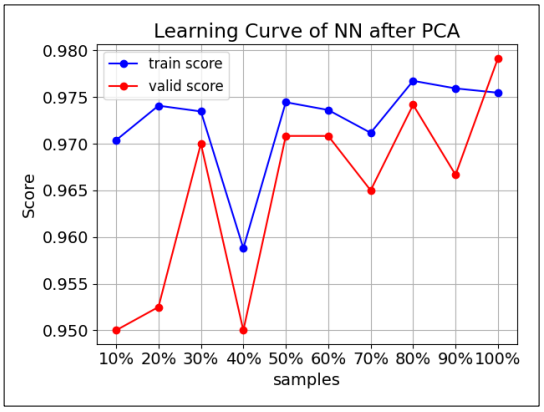

Figure 26. The learning curve of NN after PCA

The gap between train score and validation score is small and it means that there are not too much variance. Compared to the Project 1, the variance is slightly reduced. In terms of bias, it still has a little bias. However, the valid score is higher than the train score, so it has a little underfitting.

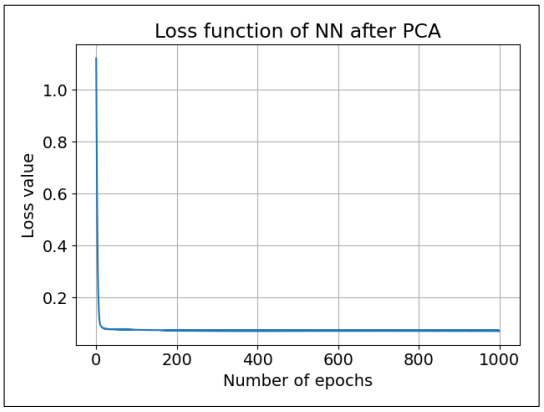

Figure 27 The loss function of NN after PCA

The loss function seems to be stable and converged well.

B. *ICA*

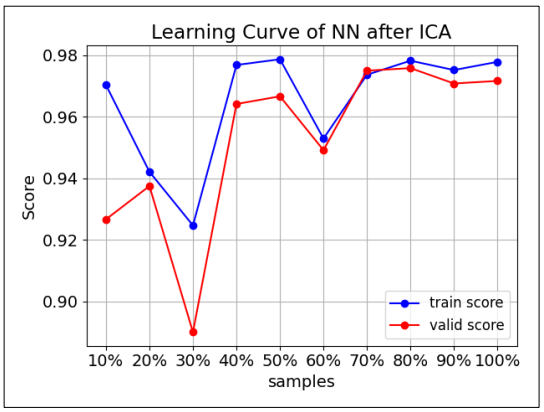

Figure 28. The learning curve of NN after ICA

The Figure 28 has similar shape with the Figure 26. The variance is small and it has a little bias. It shows that the ICA algorithm can also be an effective to reduce the variance between training data and validation data.

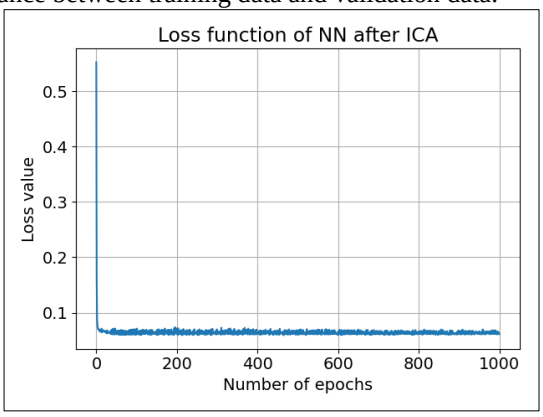

Figure 29. The loss function of NN after ICA

The loss function has a little fluctuation, and there may be a little local minima. However, the overall loss curve is relatively stable and the loss value is very small.

C. *RP*

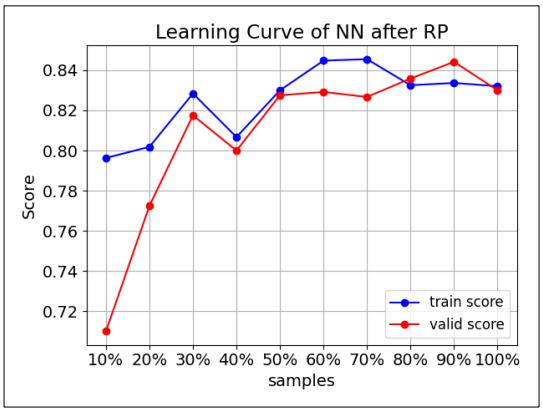

Figure 30. The learning curve of NN after RP

Figure 30 shows an interesting result. Even though the variance is very small, the bias is significantly higher. This is an underfitting and it means that the RP algorithm filtered the original dataset too much. In this project, the RP algorithm used 3 as the parameter (number of components). This may cause this underfitting result.

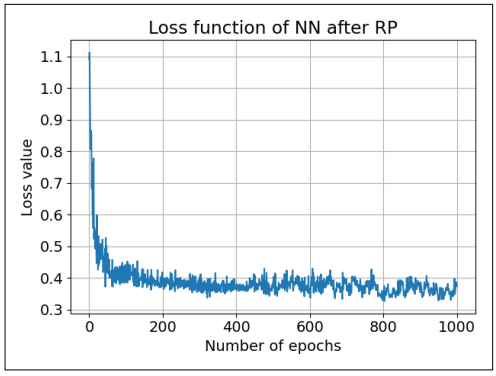

Figure 31. The loss function of NN after RP

As we can see, the loss function has significant large fluctuation and the value is relatively higher than the previous two cases. Therefore, this model is not a feasible model. To improve this drawback, more accurate tuning should be implemented or the parameter of data reduction should be changed.

D. *LLE*

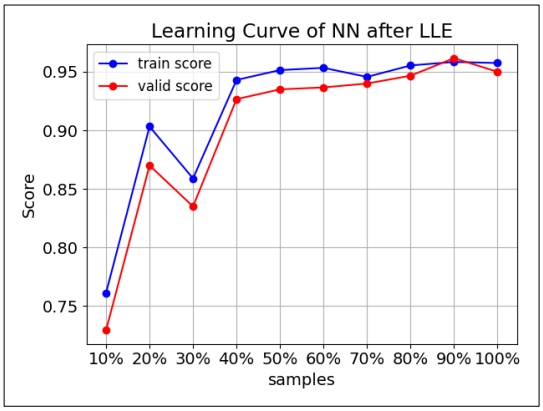

Figure 32. The learning curve of NN after LLE

In Figure 32, the result shows good variance and bias. However, the accuracy score is not good as the PCA and ICA results. It means that it is difficult to find optimal parameters such as the number of components and the number of neighbors.

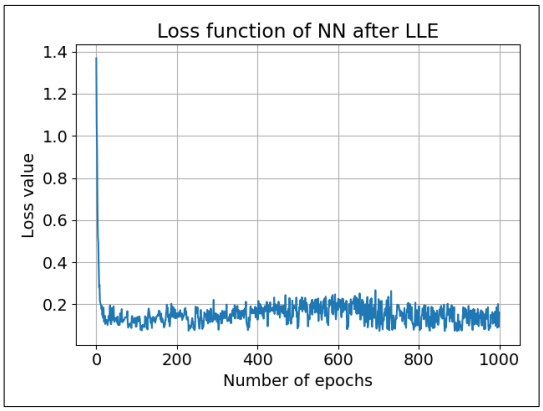

Figure 33. The loss function of NN after RP

In Figure 32, the loss function is similar to the Figure 31. The model is not a feasible model and more tuning is needed.

E. *Comparison: Accuracy and Time to train*

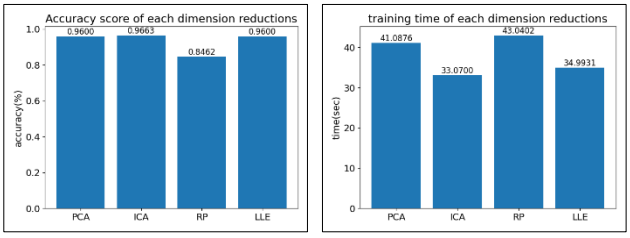

Figure 34. The accuracy and time of each dimensionality reduction

In terms of the accuracy score, the ICA is the best and PCA and LLE are the same. On the other hand, the RP does not plays a role as an optimal filter in this case. The time to train

PCA and RP takes relatively longer than the ICA and LLE. So, considering the accuracy and time to train, the ICA can be the best algorithm in this case.

VI. PART5: NEURAL NETWORK AFTER CLUTERING

In this section, the Neural Network is applied to the data which was included a output column came from each clustering algorithm. And then the effect of each clustering algorithm is investigated like the previous part. The learning curve, loss curve, accuracy, and time to train are investigated.

A. *K-Means*

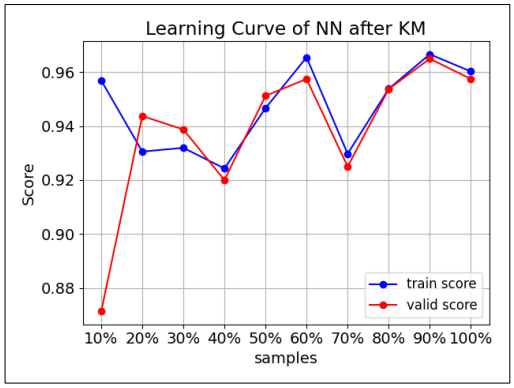

Figure 35. The learning curve of NN after K-Means

As we can see, the variance is very small. Also, a little bias is still remained in this data.

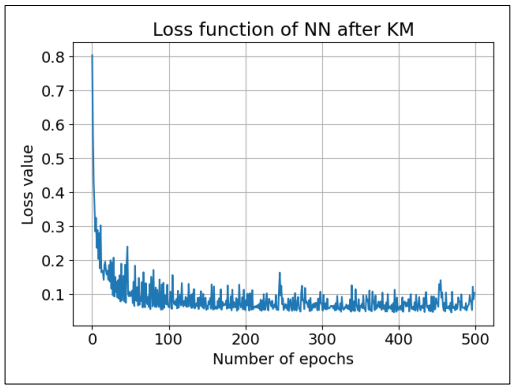

Figure 36. The loss function of NN after K-Means

B. *Expectation Maximization (EM)*

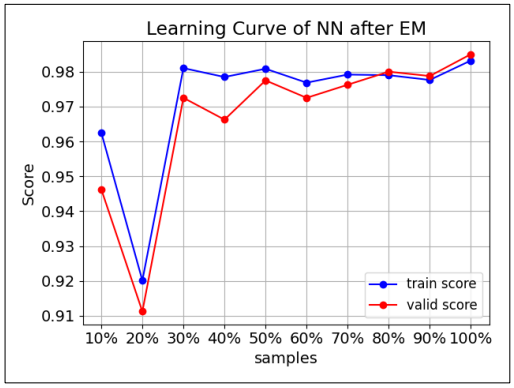

Figure 37. The learning curve of NN after EM

In Figure 37, the variance and the bias are very small, but the validation score is a little higher than the train score. This model has a little underfitting.

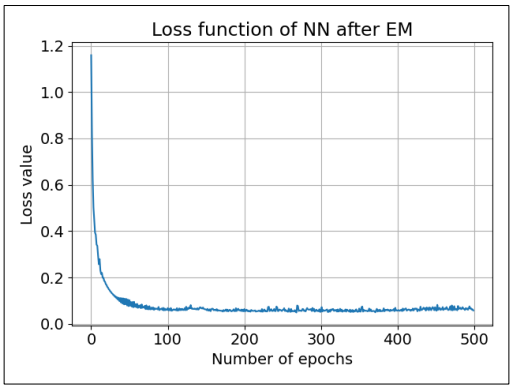

Figure 38. The loss function of NN after EM

In Figure 38, there is a small fluctuation, however the value is very small. So, it is relatively well trained than the K-Means case.

C. *Comparison: Accuracy and Time to train*

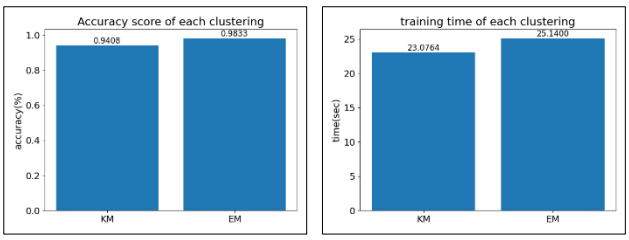

Figure 39. The accuracy and time of each clustering

In Figure 98, even though the EM algorithm takes a slightly longer time than the KM case, the performance of EM algorithm is better than the KM case. Therefore, the EM algorithm can be the best algorithm in this case.

VII. CONCLUSION

In this project, two clustering algorithms and four dimensionality reduction algorithms are investigated based on two datasets.

First, the clustering with dimensionality reduction algorithm can perform better than the clustering without dimensionality algorithm. The silhouette plots can show the improved result. Because, the dimensionality reduction algorithm can pick some ‘meaningful’ data from the original data. And it can reduce the higher dimension of the original data to a lower dimension. It also plays a role similar to noise reduction. In this case, picking the significant columns from the original data depends on the number of components of each dimensionality reduction algorithm. Therefore, determining the number of components is the most important.

Second, the dimension reduction and clustering algorithm can be beneficial way prior to the Neural Network implementation. Especially, all cases show lower variance than the variance of the Neural Network with the original dataset. However, some cases show underfitting, so accuracies for all

cases are not improved compared to Project 1. Tunned parameters for the dimensionality reduction algorithm and clustering algorithm in this project are not very effective, so it may cause this underfitting.

In summary, this project was very interesting to have an opportunity to analyze clustering and dimensionality reduction algorithm. Even though it was difficult to get improved results, the effects of clustering and dimensionality algorithms could be investigated. For better performance, determining parameters of each algorithm is the most important and it should be performed more detail than the method used in this project. For the next time, more detailed and accurate method should be introduced and these methods will be able to find the optimal parameters to lead the better performance.

REFERENCES

[1] Chris Ding and Xiaofeng He. 2004. K-means clustering via principal component analysis. In Proceedings of the twenty-first international conference on Machine learning (ICML '04). Association for Computing Machinery, New York, NY, USA, 29.

[2] UCI Machine Learning Repository: Wireless Indoor Localization Data Set

[3] Breast Cancer Wisconsin (Diagnostic) Data Set | Kaggle

[4] Mukherjee, Chandra Sekhar and Jiapeng Zhang. "Compressibility: Power of PCA in Clustering Problems Beyond Dimensionality Reduction." *ArXiv* abs/2204.10888 (2022): n. pag.

[5] Kairov, U., Cantini, L., Greco, A. et al. Determining the optimal number of independent components for reproducible transcriptomic data analysis. BMC Genomics 18, 712 (2017).

[6] Ky Vu, Pierre-Louis Poirion, Leo Liberti, Gaussian random projections for Euclidean membership problems, Discrete Applied Mathematics, Volume 253, 2019, Pages 93-102

[7] Saul, Lawrence K. and Sam T. Roweis. "An Introduction to Locally Linear Embedding." (2001).

[8] https://www.analyticsvidhya.com/blog/2021/01/in-depth-intuition-of-k-means-clustering-algorithm-in-machine-learning/

[9] https://www.stat.cmu.edu/~cshalizi/uADA/12/lectures/ch18.pdf