Comparative Study of Unsupervised Learning and Dimensionality Reduction Algorithms

Xiangnan He

Nov 5, 2017

I. Abstract

In this study, unsupervised learning and dimensionality reduction algorithms are implemented and analyzed with two datasets, which are UCI repository Breast Cancer Wisconsin (Diagnostic) and Letter Recognition datasets. Breast Cancer dataset was studied in both my assignment 1 and 2 reports. Two unsupervised learning algorithms are implemented and tested on the datasets, K-means Clustering and Expectation Maximization. Four algorithms, including Principal Component Analysis (PCA), Independent Component Analysis (ICA), Randomized Projections, and Information Gain, are implemented to reduce dimensionality. The project uses python sklearn for clustering, J48 tree, and ANN analysis, meanwhile, Weka is used for generating dimension reduced datasets with the four different algorithms. The performance before and after dimension reduction are explored and discussed.

II. Introduction

Clustering can group the instances with certain similarity together in a collection of unlabeled data. The points that are dissimilar will belong to a different cluster. Python sklearn [1] is used for clustering. Clustering algorithms used in this analysis report are described as follows.

1. K-means

In this analysis, unsupervised K-means algorithm will be used, both Euclidean and Manhattan distances were explored. It has the advantage of easy to learn and quick analysis of the data, instead of training algorithm with the data first. It aims to partition the data/instances into k clusters in which each instance belongs to the cluster with the nearest mean. These centers should be placed in smart way since different location causes different results. The better choice is to place them as far as possible from each other. Next step, we need to take each point in a given data set and associate it with the nearest center. After associating all the data points, the first round is complete. At this point, we need to recalculate k new centroids as barycenter of the clusters from previous step. Then a new association will be done with the same set of data points and the nearest centers. Because of this loop. The k centers gradually change their position until no more changes are done. The algorithm targets at minimizing an objective function - sum of squared error (SSE) function.

2. Expectation Maximization

Expectation maximization (EM) is an iterative tool to look for maximum likelihood or maximum a posterior (MAP) of parameters in a statistical model with hidden variables (usually missing data or latent variables). Two steps will be done. Firstly, an expectation (E) step will generate a function, which takes in consideration the log-likelihood (LL) from the current estimate of the parameters. Secondly, a maximization step will calculate parameters maximizing the LL from E step. The algorithm find k distributions of the given data such that the LL of data given distribution is maximized. The LL can be negative or positive based on the probability density of the function, in that LL represents products of probability densities instead of probabilities. Densities can be large and far exceeding 1, which can make LL positive. Common applications are fitting mixture models, learning Bayes net parameter with latent data, and learning hidden Markov models, and etc.

Dimension reduction is reducing the number of random variables using various mathematical methods from statistics and machine learning. It is often employed for problems with large data sets. The Curse of dimensionality suggests we limit or reduce the dimension of the dataset during analysis, in that when the dimension increases, the volume of the space increases so fast that the available data become sparse, which is troublesome for the methods requiring statistical significance. Therefore, to make sure the result is statistically solid and reliable, the amount of data points has to increase exponentially. Weka GUI [2] with FastICA [3] as an added plugin is used for implementing the dimensionality reduction algorithms. Dimensionality reduction algorithms used in this analysis report are discussed in the following paragraphs.

1. Principal Components Analysis (PCA)

PCA uses orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables – principal components. The transformation is constructed in a way that the first principal component possesses the largest possible variance, and each following component has the highest variance under the condition that it should be orthogonal to the preceding components.

2. Independent Components Analysis (ICA)

ICA is a tool for revealing hidden factors that are behind sets of random variables. The data variables are assumed to be linear mixtures of some unknown latent variables, and the mixing systems is also unknown. The latent variables are assumed non-Gaussian and mutually independent, and they are called the independent components of the observed data. These independent components can be found by ICA. ICA approaches the data such that the variables are the output of many unobserved sources. The applications of ICA include digital images, document databases, economic indicators and psychometric measurements. Typical examples are mixtures of simultaneous speech signals been picked by several microphones, brain waves recorded by multiple sensors, interfering radio signals arriving at a mobile phone, and etc.

3. Random Projection (RP)

RP is a technique for dimensionality reduction from a set of instances in Euclidean space. The advantages include its simplicity and less erroneous output. The original high dimensional data is projected onto a lower-dimensional subspace using a random matrix whose columns have unit lengths. RP is found to be computationally efficient, yet sufficiently accurate for dimensionality reduction.

4. Information Gain (IG)

IG is an entropy-based feature evaluation technique, widely used in machine learning. It is used in feature selection for evaluating the amount of information provided by the features or attributes. A rank of features will normally be generated. IG uses decision trees in which the nodes are organized from top to bottom with information gain, which is measured by how much information a feature render with regard to the classification.

III. Data analysis and preprocessing

1. Breast Cancer

Breast Cancer dataset from assignment #1 and #2 was chosen to perform the analysis.

Breast cancer is one of the diseases that causes a high number of deaths per year. It is the most common type of cancers and the major cause of women's death worldwide. The early diagnosis and prognosis of a cancer type have become a necessity in cancer research, as it can facilitate the subsequent clinical management of patients. Machine learning is frequently used in breast cancer diagnosis and detection by classifying cancer patients into high or low risk groups, helping to model the progression and treatment of cancerous conditions. In addition, the ability of ML tools to detect key features from complex datasets reveals their importance. The dataset contains 569 instances, 30 attributes, and the classification column contains two categories: M = malignant, B = benign [4].

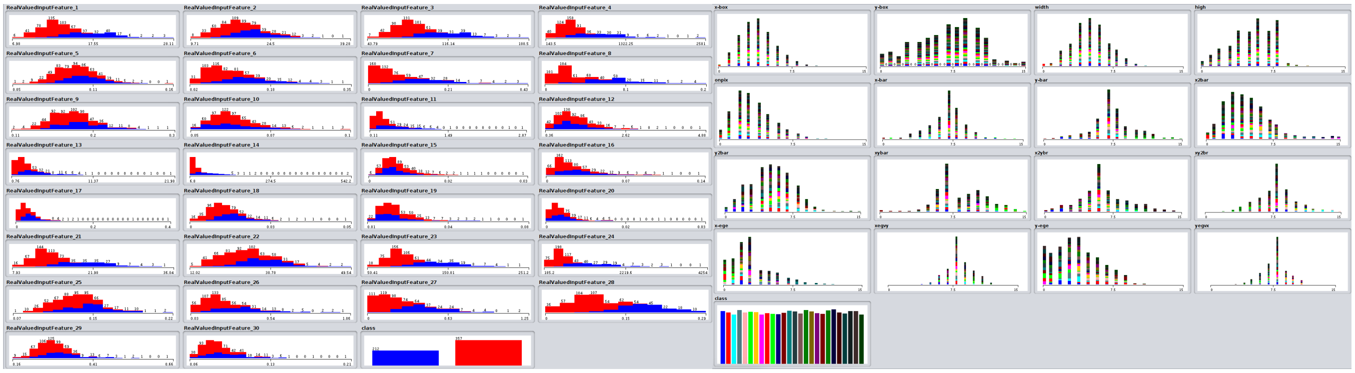


Figure. 1 Visualization of Breast Cancer dataset on the left, and letter recognition on the right with Weka.

2. Letter Recognition

In the field of machine learning, computer vision is growing very fast, with applications, such as, virtual reality, augmented reality, autonomous vehicles, and medical imaging. Optical character recognition was developed to be more and more important for supporting different industries, such as finance, law and construction, and healthcare to help reducing paperwork, improving processes, and automating tasks. Optical character recognition paves the way for computer vision industry and guides the development of numerous sub-domains.

The letter recognition dataset we used is from UCI repository [5]. The objective is to identify each of a large number of black-and-white rectangular pixel displays as one of the 26 capital letters in the English alphabet. The character images were based on 20 different fonts and each letter within these 20 fonts was randomly distorted to produce a file of 20,000 unique stimuli. Each stimulus was converted into 16 primitive numerical attributes (statistical moments and edge counts) which were then scaled to fit into a range of integer values from 0 through 15.

In assignment 1, the datasets were split into train and test sets to compare the performance of different classifiers by looking at both training score and test score to address the overfitting issue. However, in this assignment, the datasets are not split due to the consideration of computational burden over selection bias.

IV. Results

Part I. Clustering before Dimensionality Reduction

Both K-means and EM uses sklearn and the evaluation criteria include average within cluster SSE for K-means and LL for EM, homogeneity and completeness scores. Homogeneity shows how the clusters contain only members of a single class, completeness shows the degree that all data from a given class are associated to the same cluster. Akaike information criterion (AIC) and Bayesian information criterion (BIC) are also discussed for EM. AIC and BIC are useful to select the value of the regularization parameter by making a trade-off between the goodness of fit and the complexity of the model. A good model should explain the data well and stay simple. AIC does not provide a test of a model in the sense of testing a null hypothesis, thus it tells nothing about the absolute quality of the model, but the quality relative to other models. The covariance type for EM is diagonal.

1) Breast Cancer & Letter Recognition K-means

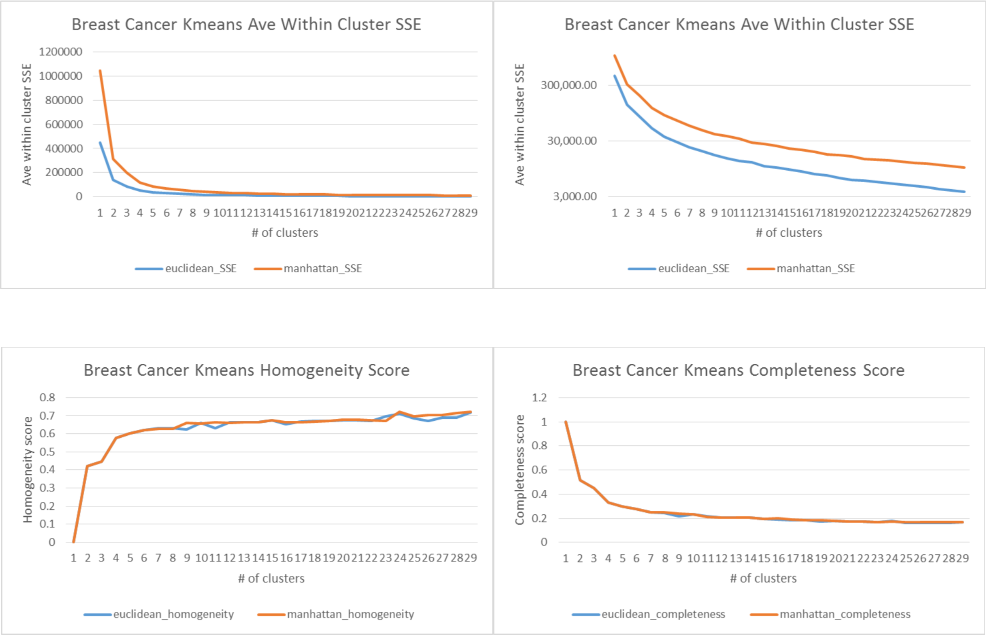


Figure 2. Top row left is the Breast Cancer K-means average SSE plot, and its log form on the right. Bottom row is the homogeneity score (left) and completeness score (right)

From Fig.2, with elbow method, we can see that cluster number = 2 is probably what we want to choose. At cluster = 2, the curves has the highest slope whereas when cluster number increases, the curves flatten out. On the other hand, we know there are only two classes in the dataset, M and B. The top row plots (the right side is with log Y axis) shows the difference between Euclidean and Manhattan, although the trend of decreasing within cluster SSE is the same. Euclidean SSE is significantly lower than Manhattan. This is expected, since Euclidean distance is the shortest between two points, whereas Manhattan take the sum of the step distances in each dimension, like the city blocks. Homogeneity means all of the observations with the same class label are in the same cluster. The higher the score, more likely each cluster contains only members of a single class, which make sense since the method converges when each cluster contains the same class. We can see the homogeneity score starts at 0 due to the fact that 1 cluster contains all the classes. Homogeneity score is growing with more and more clusters, because more clusters tends to distinguish the classes better. Meanwhile, completeness score means all members of the same class are in the same cluster. With cluster # = 1, the cluster include all members of all classes, therefore, the score is the highest = 1. We can see the completeness score is decreasing, indicating the members of the same class are more and more split into different clusters due to more and more clusters are added.

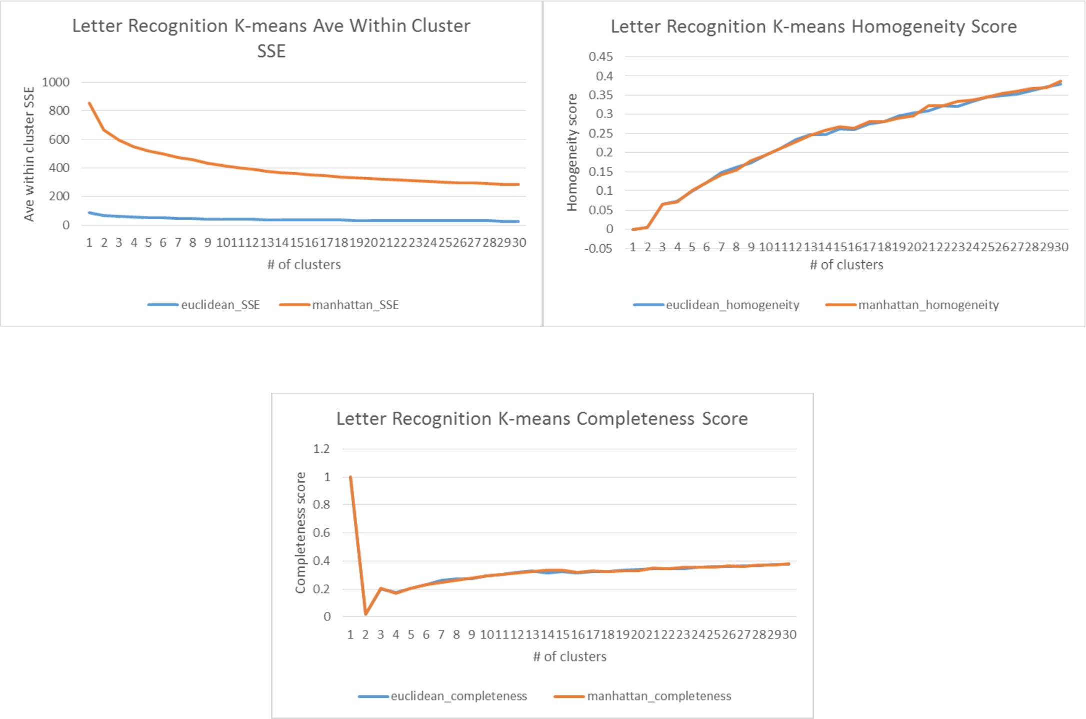


Figure 3. Letter recognition K-means plots.

From Fig.3, the elbow methods cannot be easily applied in this scenario as the curves are smooth. This algorithm captures even different styles of the same character, therefore, the SSE keeps decreasing with increasing # of clusters. It makes sense to choose the # of cluster = 26 as the best cluster number for the dataset for this study. Again, the homogeneity score is gradually increasing due to that more clusters can distinguish better of the classes, meanwhile, completeness score decreases at the beginning from cluster # = 1 to 2, however, the completeness score keeps increasing with more clusters due to the fact that the clustering algorithm can find more than 26 letters or some letters show more than one style.

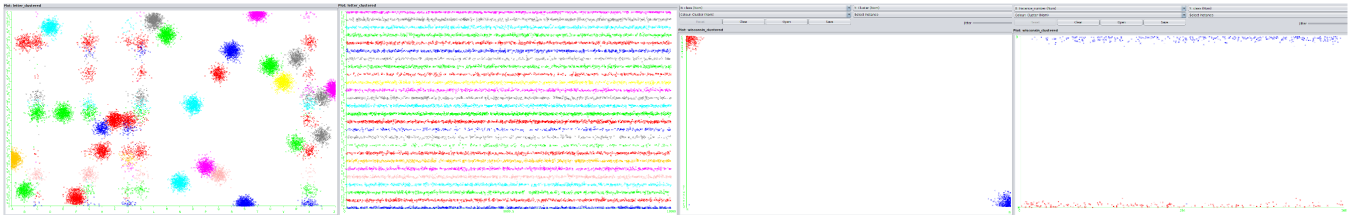


Figure 4. Letter recognition and breast cancer K-means visualization using Weka.

In Fig.4, the clusters of K-means are visualized. Cluster number was set at 26 for letter recognition and 2 for breast cancer. The first image from left compares the 26 classes on the X and 26 clusters on the Y axis with 50% jitter. We can see there are more than 26 clusters. We can see multiple clusters with regards to the same class, which suggests the same letter is clustered to different centroids. Meanwhile, there are also points in the cluster belong to different class, suggesting the algorithm is not able to distinguish between classes on certain letters. The second image from the left compares 20000 instances on the X-axis against 26 classes on the Y axis. The two images on the right are cluster vs class and class vs instances. They show clear clustering of the breast cancer data, there is not much incorrectly clustered points given the 50% jitter.

2) Breast Cancer and Letter Recognition EM

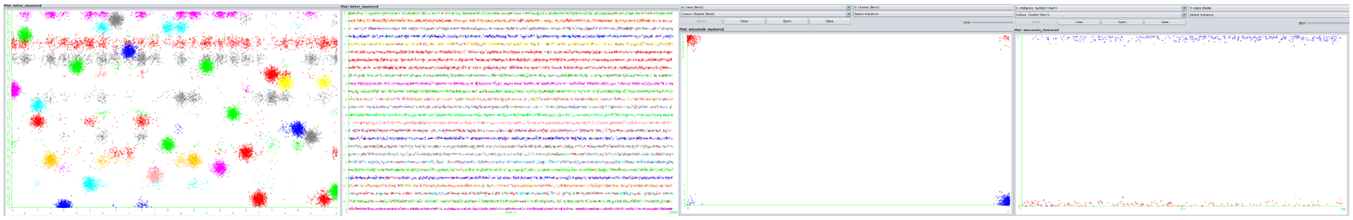


Figure 5. Breast cancer and letter recognition EM visualization with Weka.

In Fig.5, the clusters of EM are visualized. Cluster number was set at 26 for letter recognition and 2 for breast cancer. The first from left plots the 26 classes on the X and 26 clusters on the Y axis with 50% jitter. We can see there are more than 26 clusters. The horizontal line in first image from left suggests there are many letters in the same cluster, which suggests the algorithm has difficulty distinguish them. The second image from the left compares 20000 instances on the X-axis again 26 classes on the Y axis. The two images on the right shows clear clustering of the breast cancer data. For the image second from right, we can see more data in the cluster from the incorrect class compared with K-means method in Fig.4, suggesting the EM method is likely less efficient on distinguishing between the two classes.

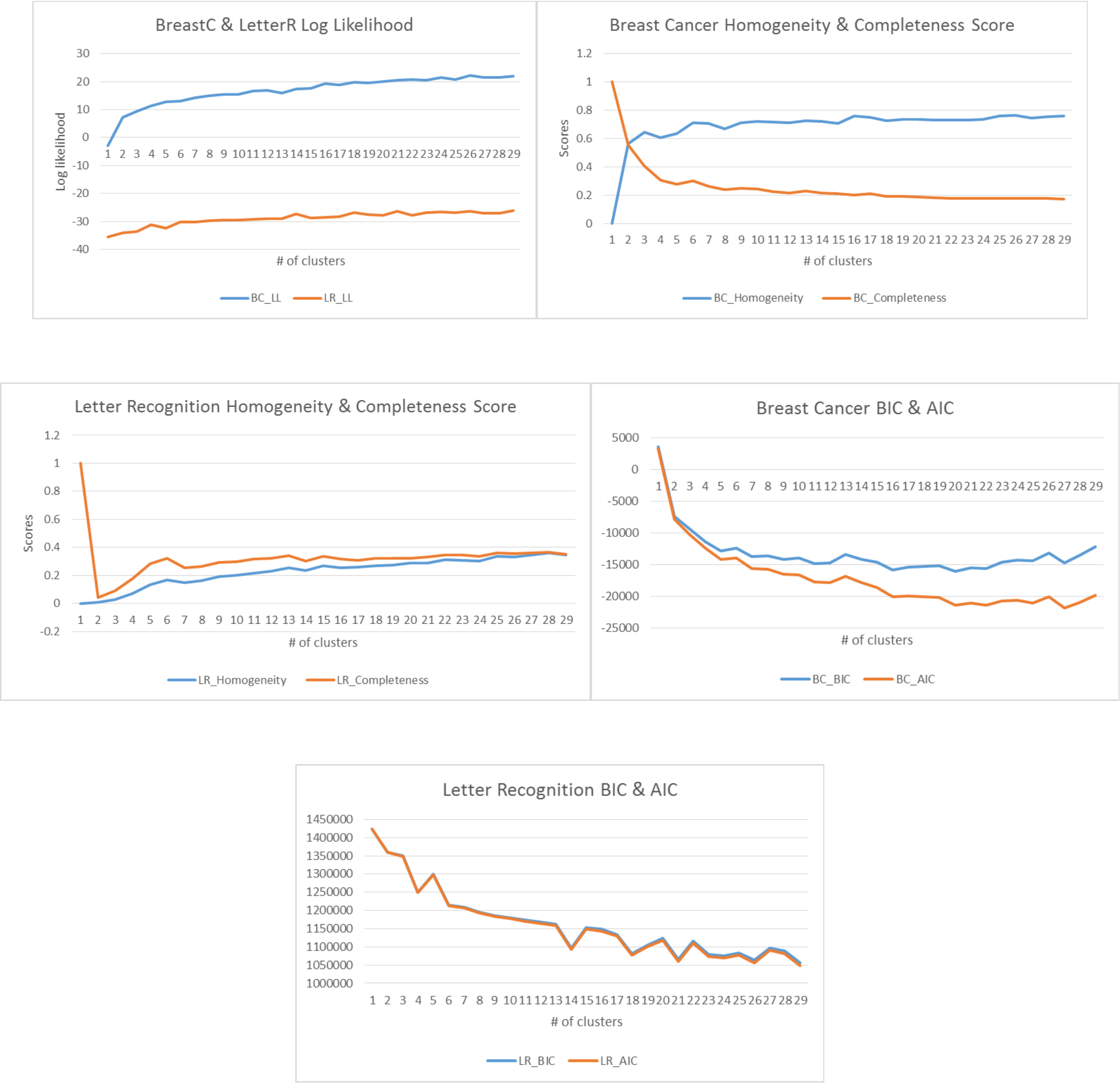


Figure 6. Breast cancer and letter recognition EM plots.

In Fig.6, we can also find out the best choice of clusters in the LL plot for breast cancer with elbow method. The LL of breast cancer is positive after number of clusters >= 2, suggesting the probability densities are greater than 1. The homogeneity score increase with increasing cluster # due to better distinguishing capability with more clusters. The completeness score decreases with more clusters due to more clusters separate the classes into different clusters. For breast cancer dataset, the BIC and AIC plot shows a difference in the decreasing trend, where AIC decreases faster. The formula for BIC is similar to the formula for AIC, but with a different penalty for the number of parameters. With AIC, the penalty is 2k, whereas with BIC the penalty is ln(n)k, where n is the sample size.

For letter recognition, LL grows with more cluster# because more capability of distinguish with more clusters. The homogeneity score increases with cluster # and completeness score decrease but increase after cluster # = 2. Both scores reach a peak at cluster # = 6. Interestingly, the two score becomes identical at high cluster # around 29. The AIC and BIC show similar the trend, and there is a peak around 27 and 28, which is close to 26.

Part II. Clustering with Dimensionality Reduction

The DR algorithms with Weka and the procedure includes:

1) Use DR algorithm to generate the transformed dataset;

2) Train J48 classifier with the transformed dataset for the optimal number of attributes for each algorithm. In this step, the transformed attributes are removed and added in one at a step for plotting the curve of the accuracy of J48.

3) Use K-means and EM on the transformed datasets based on the optimal number of attributes or components.

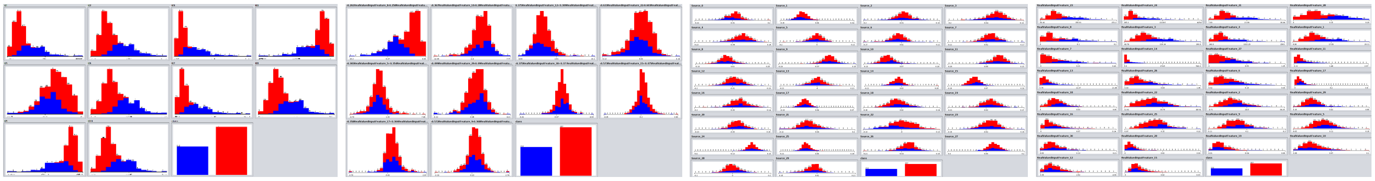


Figure 7. Breast cancer dimensionality reduction methods visualization with Weka. From left to right. PCA, ICA, RP, and IG.

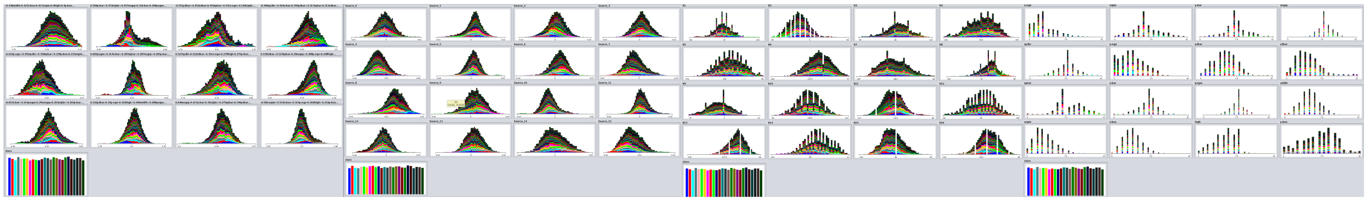


Figure 8. Letter recognition dimensionality reduction methods visualization from left to right, PCA, ICA, RP, and IG.

The breast cancer data in Fig.7 are transformed by PCA, ICA, RP and IG. The transformation is not so obvious in this dataset compared with the letter recognition. The PCA and ICA reduces the features to 10 from 30. The letter recognition data in Fig.8 are transformed by PCA, ICA, RP and IG. The original data we not Gaussian like, they are sharp peaks. The PCA transformed data shows smooth peak with tails on both ends. PCA also reduces the features to 12 from 16. ICA show much sharper peaks, we cannot see it in the images because the images has tiny scales on X axis. RP is not so smooth. We can see that the IG simply reordered the features without altering the features themselves.

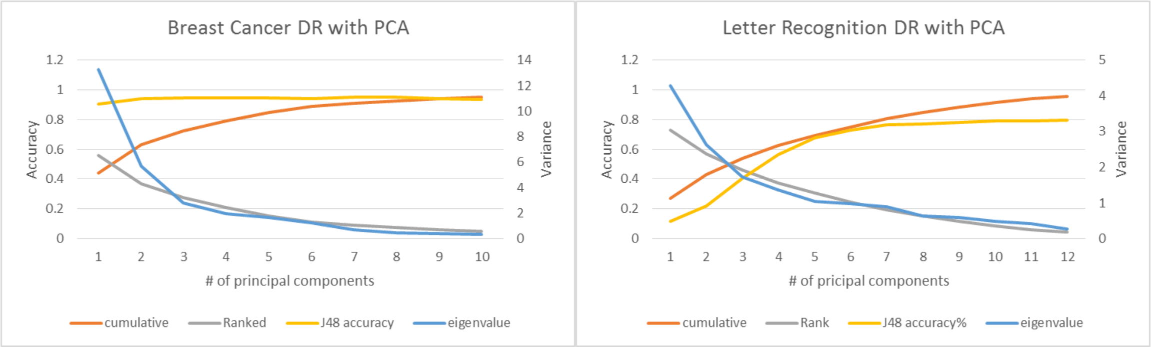
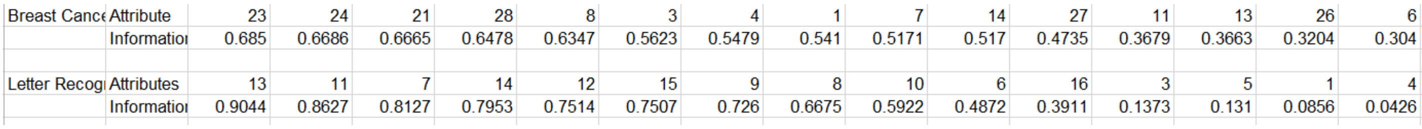


Figure 9. Breast cancer and letter recognition DR with PCA.

In Fig.9, the principal components are sorted by variance/eigenvalues. The slope of cumulative variance decreases with increasing number of PCs. J48 decision tree accuracy can efficiently point out when the reducing of PCs starts to make less impact. For breast cancer dataset, the J48 accuracy increased slightly from 1 PC to 2PCs and then flatten out. The J48 accuracy is above 0.9 even with just 1 PC, suggesting the PC method can distinguish the data really well. For letter recognition dataset, the J48 accuracy drop significantly when # of PCs drop below 5.



Breast Cance	Attribute	23	24	21	28	8	3	4	1	7	14	27	11	13	26	6
	Informatio	0.685	0.6686	0.6665	0.6478	0.6347	0.5623	0.5479	0.541	0.5171	0.517	0.4735	0.3679	0.3663	0.3204	0.304
Letter Recog	Attributes	13	11	7	14	12	15	9	8	10	6	16	3	5	1	4
	Informatio	0.9044	0.8627	0.8127	0.7953	0.7514	0.7507	0.726	0.6675	0.5922	0.4872	0.3911	0.1373	0.131	0.0856	0.0426

Figure 10. Breast cancer and letter recognition DR with IG

In Fig(or table).10, due to plotting issue, I just put the data in this table. We can see the ranking of the attributes with the information gain sorted from high to low using the IG method from Weka GUI.

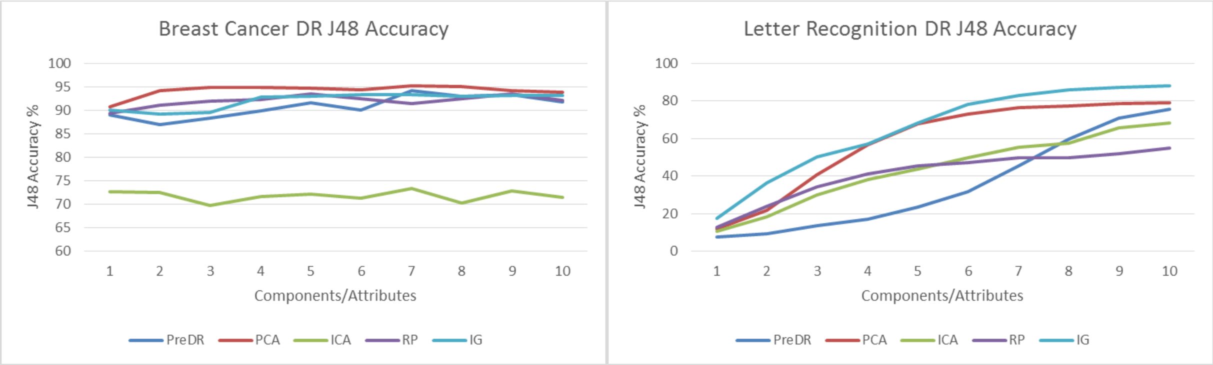


Figure 11. Breast Cancer and Letter recognition dimensionality reduction J48 performance.

As shown in Fig.11, the J48 scores are shown on the datasets with the components/attributes added one by one. One of the curve is named PreDR, which represents the J48 accuracy by adding the features one by one according to the feature order on the original dataset. The best for breast cancer is PCA showing improvement from PreDR data. RP and IG is similar compared with PreDR data, suggesting the original features are already very efficient in distinguishing the classes. The ICA performs the worst due to its independent components do not represents meaningful information, which causes low performance on J48. Overall, we can choose the # of components or attributes = 2, since it will render the similar J48 accuracy compared with # of components or attributes = 10.

For letter recognition on the right, the best is IG and PCA, showing improvement from PreDR data. Since J48 decision trees use information gain in the algorithm, it is expected that IG will perform best of the four methods. RP performs the worst. RP is used to project the data into a new space of different directions. Looks like the RP method projected the data into directions not efficiently classifying the letter classes. Interestingly, the accuracy slope reduces after component # >5 for the DR data, whereas the PreDR data shows accuracy curve slop reducing at component # > 9. Thus, we want to choose # of components/attributes = 5.

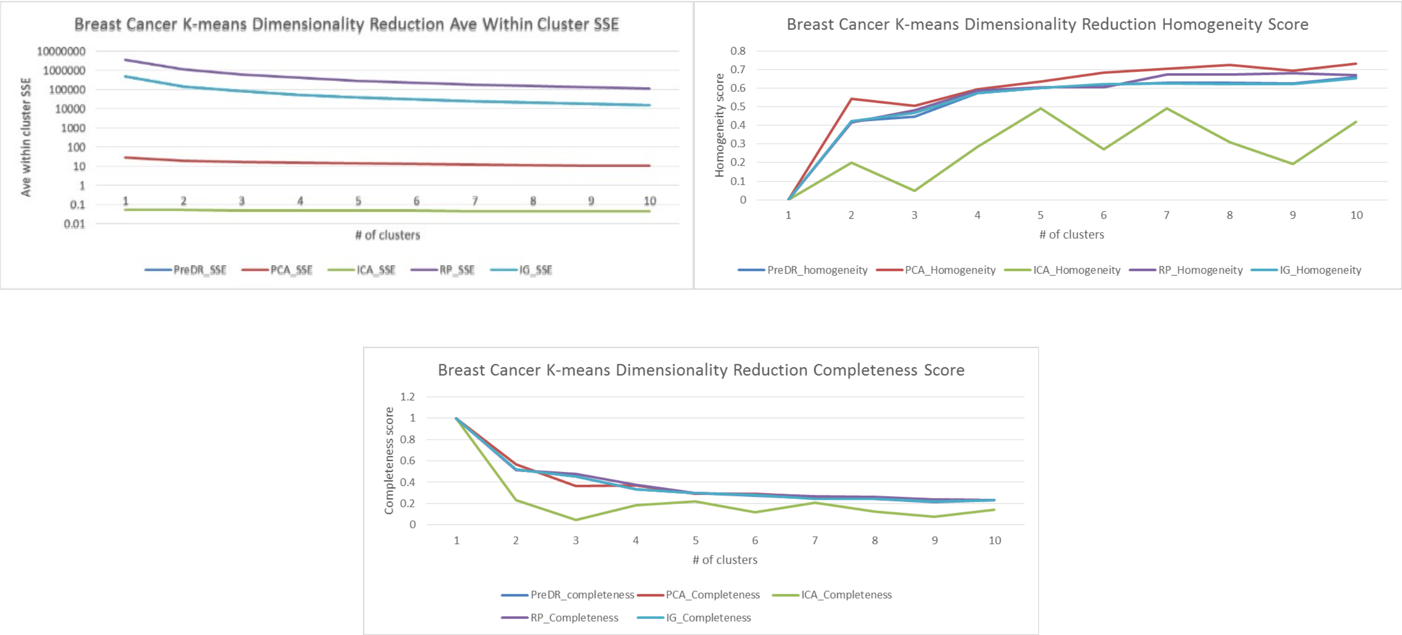


Figure 12. Breast cancer K-means DR analysis plots with comparison between PCA, ICA, RP and IG.

For the following analysis, we only use Euclidean distance for calculating the SSE to save some space in the report, since it is the mostly used method. After choosing the cluster # = 2, the clustering methods are used again to the dimensionality reduced dataset. In Fig.12, interestingly, we can see the PreDR_SSE is identical to IG_SSE, due to the fact that IG only reorder the features instead of altering them, so the average within cluster SSE is very similar. The ICA and PCA have significantly lower SSE. This indicating the principal component method removes unnecessary features with more focus on critical features. On the other hand, ICA method also reduces the noise by even more significantly reducing the SSE. However, ICA method generate worst homogeneity score and completeness score, actually from Fig. 11, it is also the worst. ICA cannot distinguish well the different instances with different classes, causing worse scores. PCA is the best given that the overall performance on the SSE and score are the best.

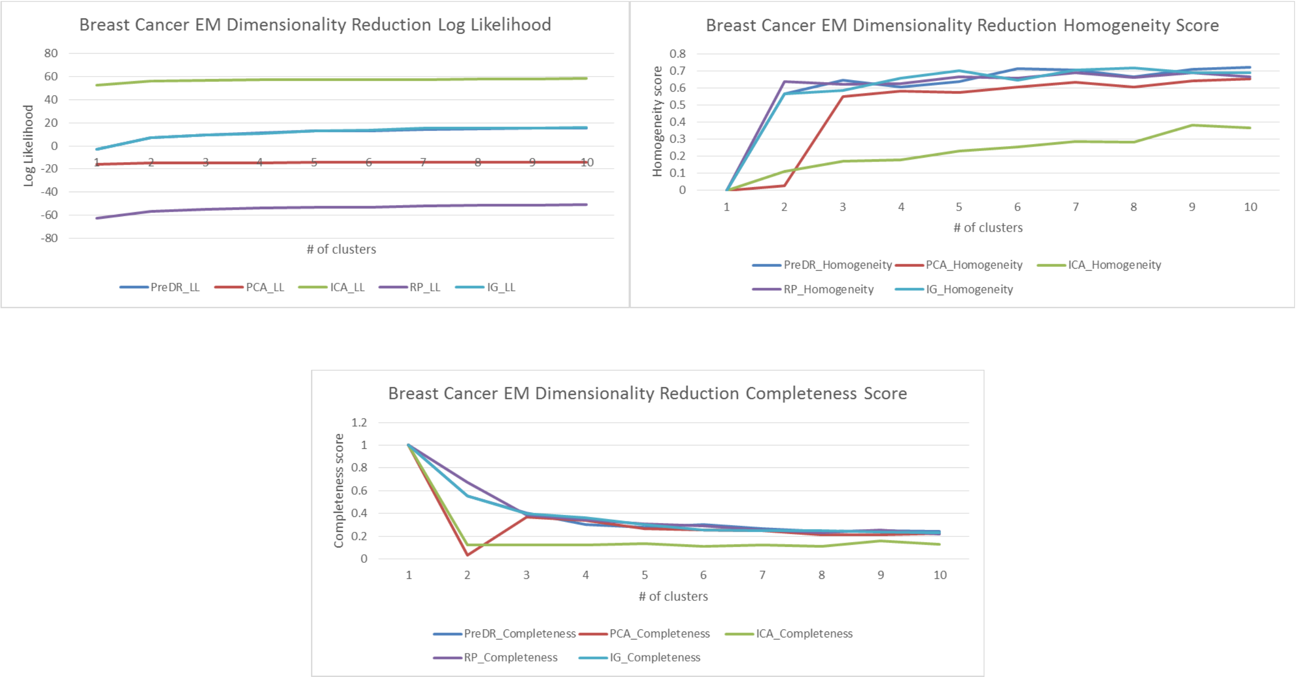


Figure 13. Breast cancer EM analysis with DR algorithm comparison

The Fig. 13 shows the best log likelihood is ICA, but again ICA has worst homogeneity and completeness scores. The LL of ICA is improved compared with original data, and it becomes positive, suggesting the transformation into independent components result in probability density greater than 1. This suggests the scores are a good metric to evaluate the DR algorithm by looking at more aspects of the data instead of just calculating the LL. On the other hand, RP has good scores, but the LL is quite low. From Fig. 11, we know that RP still performs well (much better than ICA) with LL much lower than ICA.

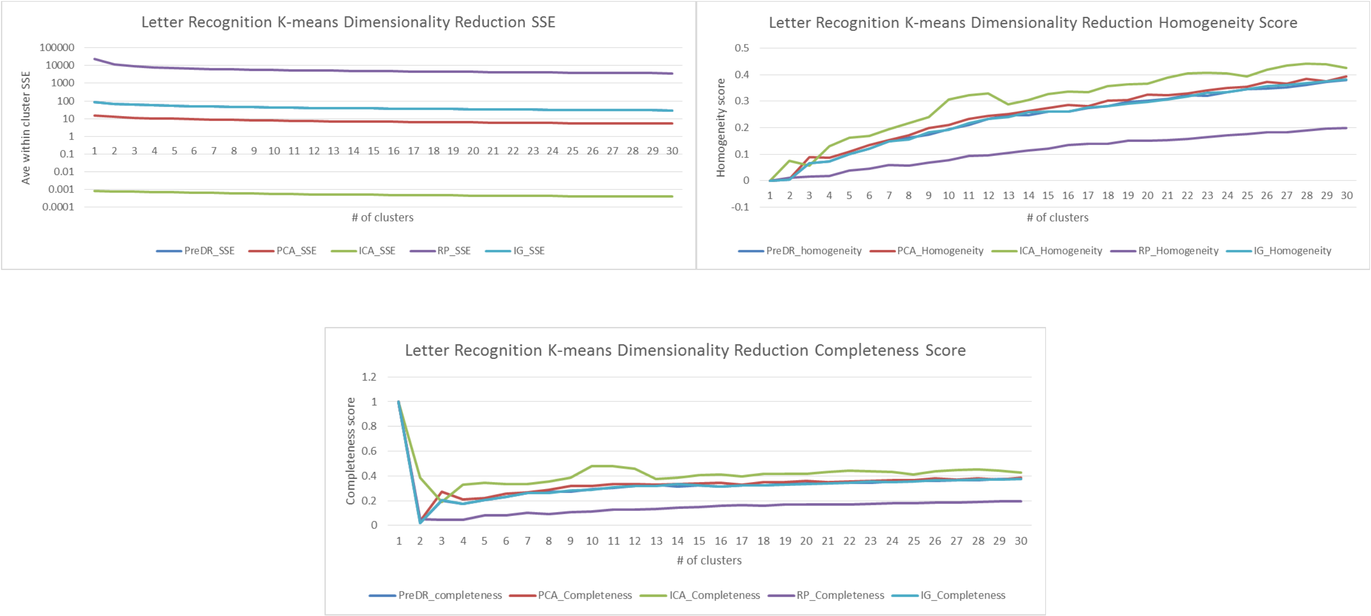


Figure 14. Letter recognition K-means DR plots with comparison between DR algorithms.

From Fig.14, SSE decrease with # of clusters, ICA has the lowest average within cluster SSE. ICA also has the highest homogeneity and completeness scores. It is possible that the independent components well represent the latent variables and more efficiently distinguish the data. There is a peak at cluster # = (10, 11, and 12). On the contrary, RP has high SSE and low scores, suggesting the randomly projected directions are not efficient in distinguish the classes.

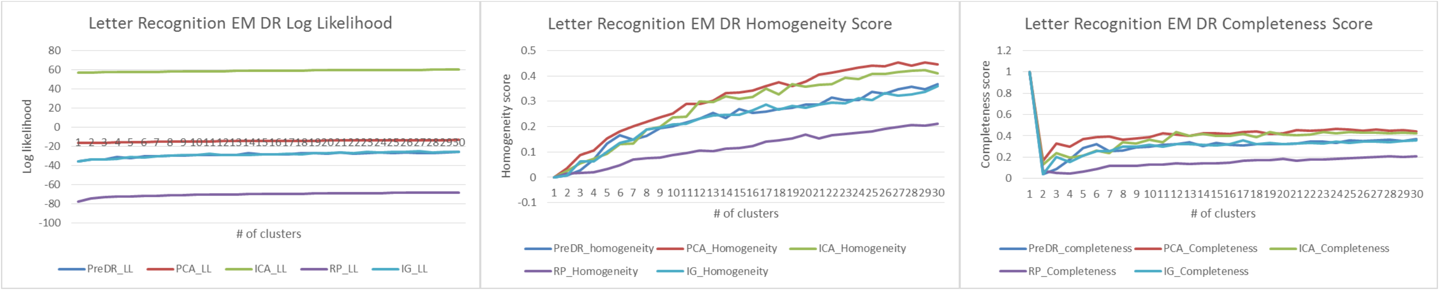


Figure 15. Letter recognition EM DR plots with comparison between DR algorithms.

In Fig.15, the best LL in letter recognition is ICA, again, it is positive, suggesting a probability density above 1. The worst is RP in this case from the overall performance on the LL and scores.

Part III. ANN with Dimensionality Reduction

In this part, we train the neural network with four DR algorithm reduced datasets. The four algorithms are compared by starting with only one components or attributes and then add back one by one of the attributes/components with the ANN classifier. The full datasets was used for training to reduce the time instead of cross validation. The accuracy of the breast cancer with ANN from the assignment 1 was 0.892 and the best score from assignment 2 was 0.95.

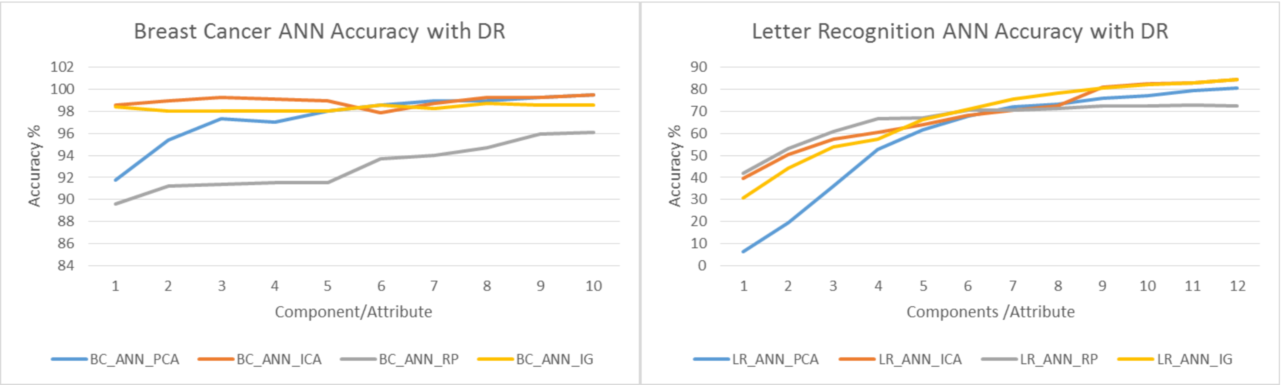


Figure 16. Breast cancer and letter recognition ANN plot with DR comparison

We can see the best DR algorithm for breast cancer is ICA and PCA (above 98%), whereas, the best for letter recognition is IG and ICA (around 80%). RP didn't perform well with a lower score on both datasets.

In the breast cancer dataset, interestingly, ICA and PCA performs better than the other two. Especially that ICA performs the worst in J48 classifier. It is possible that ANN is better at finding the global optima using the backpropagation with gradient descent learning, while J48 decision tree kind of stuck at a local optima.

Part IV. Further Discussion

K-means algorithm computes the distance between two points, it is difficult to use it on categorical (e.g. infant, youth, mid-age, senior) attributes/features. A simple way to work around is to calculate the percentage of times each variable matches the cluster centroid.

We didn't include the initialization/seed discussion in the plots, for EM, the performance can be further improved using k-means to initialize the best starting locations. Weka GUI actually has the option to choose k-means++ for initialization. Deciding cluster # in K-means is critical because additional cluster improves the quality of clustering but at a decreasing rate, and having too many clusters may be useless. Normally, initial seed number for K-means is the same as number of clusters, k. Seeds are randomly generated to determine the starting centroids of the clusters. The seed value in Weka GUI is different. In most cases, it is 10 in Weka GUI. The seed value is used in generating a random number which is, in turn, used for making the initial assignment of instances to clusters. In general, K-means is quite sensitive to how clusters are initially assigned, therefore, it is often necessary to try different values and evaluate the results.

Although not included in the study here, cross-validation can be used for further improve the results.

V. Conclusions

In this analysis, K-means and EM are analyzed before and after dimensionality reduction in this study. The performance of the transformed datasets are evaluated with J48 decision tree classifier by adding back lower ranked features one by one to determine the best number of components or attributes to use in the following evaluations. The J48 accuracy curve flatten out after 2 components or attributes for breast cancer, whereas the accuracy curve keeps growing with more components or attributes for letter recognition, but the curve slope reduces after component # of 5, which make sense in that more features are need to distinguish more instances and classes. ANN were explored with different datasets from different DR algorithms, and breast cancer training score is discussed and compared with previous assignments showing improvement with DR. PCS and ICA shows training score above 98% for breast cancer datasets. IG and ICA perform best in letter recognition dataset.

VI. References

[1] Sklearn: http://scikit-learn.org/stable/

[2] Weka: https://www.cs.waikato.ac.nz/ml/weka/

[3] https://github.com/cgearhart/students-filters

[4] This database is also available on UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

[5] UCI Machine Learning Repository for Letter Recognition Data Set:

https://archive.ics.uci.edu/ml/datasets/letter+recognition