Unsupervised Learning and Dimensionality Reduction

Xujian Liang GTID: 903421366

Abstract:

This paper explores clustering and dimensionality reduction techniques to pre-process the data and uses such techniques to train artificial neural networks. K-means and expectation maximization are two clustering algorithms used. Four dimensionality reduction techniques are: principal component analysis; independent component analysis, random projection and information gain. The paper is organized in three parts: part one explore two clustering algorithm; part two applies four dimensionality reduction techniques and cluster the dimension reduced data; part three applies both dimensionality reduction methods and clustering algorithms, and use the new data to train neural networks.

Datasets

Bread cancer Wisconsin diagnostic dataset and letter recognition dataset are used in this assignment. The letter recognition dataset was used in assignment 1 as well.

Breast cancer dataset

Despite the recent research advancement, breast cancer continues to be one of the most common cancers and second largest cancer deaths among women. Over 1 in 8 women in the United States will be diagnosed with breast cancer in her life time. The breast cancer victim's survival chance is improved by early detection and increased awareness.

The breast cancer dataset contains two classes as diagnosis: malignant and benign. It has 569 instances and 30 real-valued features. It is an interesting dataset with respect to machine learning because it has many features and thus a good candidate for dimensionality reduction.

Letter recognition

Computer vision and image recognition is an interesting field in machine learning. Many industrials use character recognition to help with process automation and improvement. The scanner is able to use letter recognition to convert image to text.

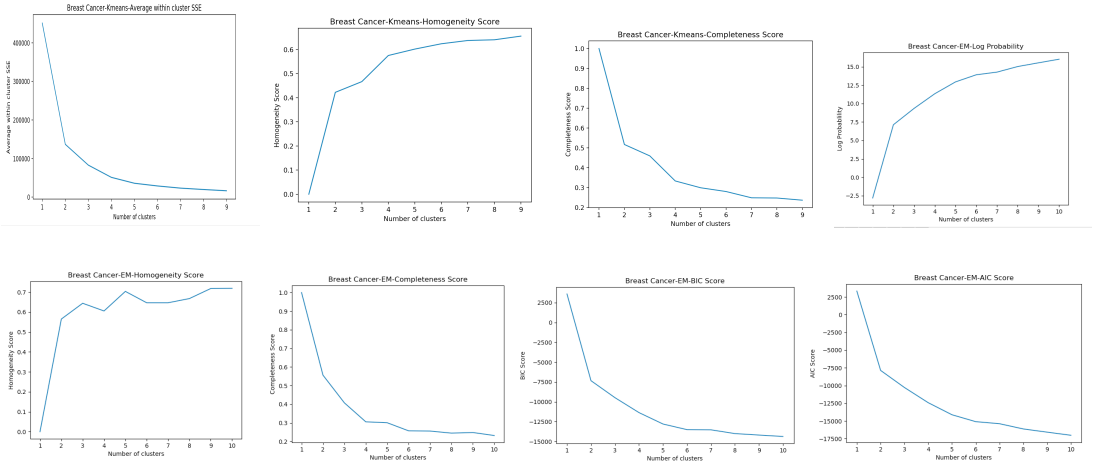
The dataset has 26 classes and each class is one letter in alphabet. It also has 16 features and 20000 instances of user-generated letters. It is interesting with respect to machine learning because it has many numeric features and thus a good candidate for dimensionality reduction and neural network.

Part 1: Clustering

Clustering is a method of grouping the instances together such that instances which belong to same cluster are more similar to each other than those in other clusters. In this section, K-means cluserting and EM algorithm are explored. In K-means, Euclidean distance is used because other distance function might not converge. Besides, K-Means is implicitly based on pairwise Euclidean distances between data points, because the sum of squared variance from centroid is equal to the sum of pairwise squared Euclidean distances divided by the number of points. Contrast to K-Means, EM is structured with probability distribution. It uses maximum likelihood parameters. EM alternates between estimating the log-likelihood of current estimates (E step) and maximizing the likelihood based on the E step (M step).

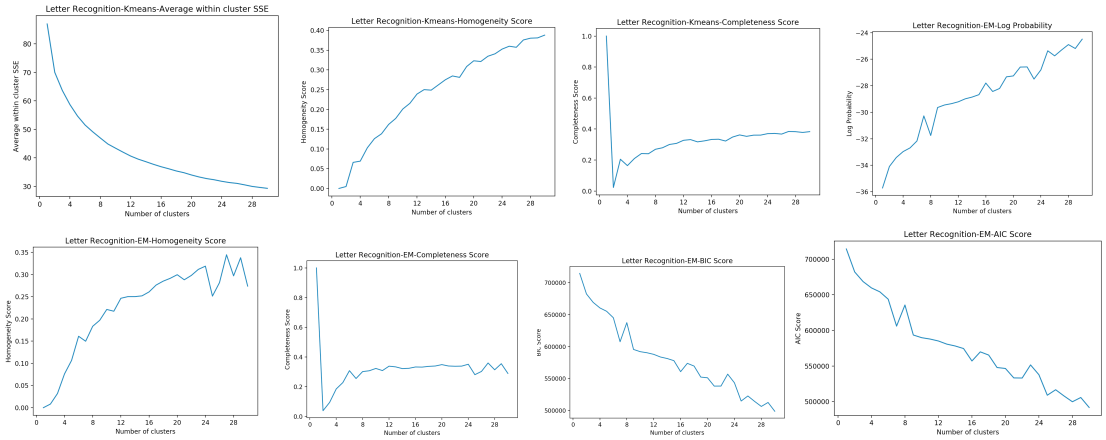
Both K-Means and EM are implemented using Scikit-learn. Clusters are evaluated using average within-cluster sum of square errors for K-means and log likelihood for EM. Homogeneity and completeness and adjusted RAND score are also used to evaluate the cluster. Homogeneity describes how each cluster contains only members of a single class, completeness describes the degree in which all members of a given class are assigned to the same cluster. Akaike Information Criterion (AIC) and Bayesian information criterion (BIC) as provided to evaluate EM.

Breast Cancer



From the above plots, we can use the elbow idea to evaluate the cluster. For almost all the plots, the elbow methods indicate that cluster = 2 seems to be the best choice, that is because when cluster number = 2, we can see the angle in SSE and log probability curves and after that the curve starts to flatten. This actually makes sense because there are only two classes in the breast cancer datasets.

Letter Recognition



For completeness score, we actually see the score improves as the number of cluster increases. This is because the clustering algorithms recognizes more than 26 different letters and some letters may have more than one appearance, thus adding the cluster numbers actually considers different appearances and styles of single letters and styles of single letters and differentiate it in more detail. In log probability and AIC and BIC scores, it is reasonable for us to assume the good cluster numbers are around 26.

Part 2: Dimensionality Reduction and Clustering

Dimension reduction algorithms transform the input data to fewer dimensions. Four algorithms are chosen: principal component analysis (PCA), independent component analysis (ICA) and random projections (RP) and information gain.

Methodology

All dimension reduction algorithms use Weka and are applied on both dataset. The procedure is shown as follow:

1. Apply the dimension reduction algorithm to the original dataset to get transformed dataset.

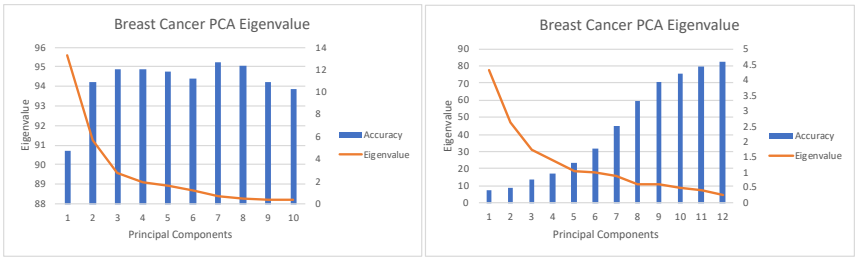
2. Apply J48 classifier to get the optimum choice of number of components for each algorithm. The newly transformed feature is removed one by one until the classification accuracy drops. 10 fold cross validation is used.

3. Apply K-means and EM clustering analysis on the newly transformed data based on the previous search on the optimum number of principle components.

Principal Component Analysis

Principal component analysis finds the orthogonal eigenvectors that best explain the maximum amount of variance. I used Weka to apply PCA. The maximum number of attribute in names is 5.

Dimension Reduction Analysis

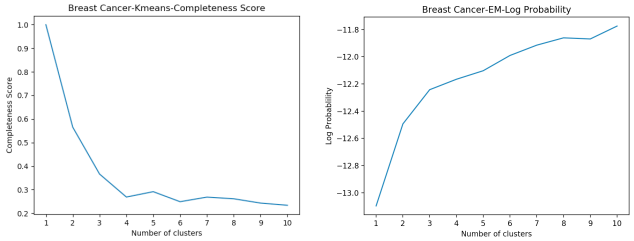


For both datasets, the eigenvalues for the last few components are relatively small, giving the possibility of removing to apply classification.

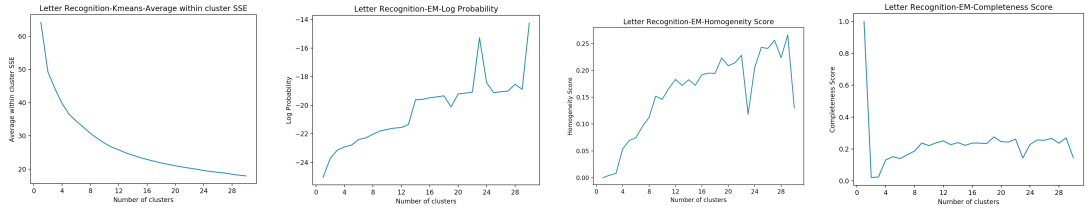
From the classification results, we can see for breast cancer dataset, after 7, the accuracy starts to drop and this suggests that the remaining principal components actually contain some noise that impacts the classification. For letter recognition dataset, the accuracy becomes flat when the number of principal components is equal to 9. It suggests the remaining components do not contain worthy information that helps classification.

Cluster Analysis

Cluster algorithms are applied on the transformed data after PCA with number of components = 7 for breast cancer datasets and 11 for letter recognition datasets.



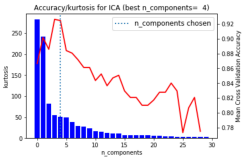
For breast cancer dataset, as we can see from SSE and log probability, the curve has its angle when cluster number = 2. PCA transformed data has similar performance curves as the original dataset. However, with PCA, SSE is lowered and log probability is increased. This indicates that PCA makes it easier to cluster the data.

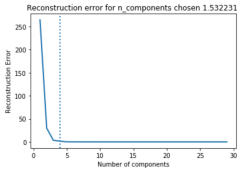


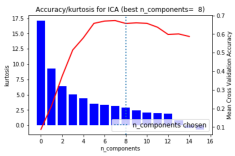
For letter recognition, again the SSE decreases and log probability increases with PCA transformed data. Homogeneity and completeness score curves become smoother as well. It is not easy to identify the best cluster number by elbow method, but as we look at spikes in the curve, we see the big spikes when cluster = 26 which is similar to original dataset.

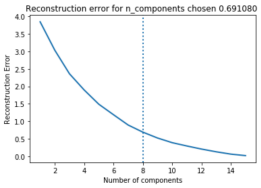
In dependent Component Analysis

Independent component analysis tries to reconstruct the data by maximizing the difference between components and find independent components of the original data. I use fastICA in Weka. The independent components are sorted by kurtosis values from highest to lowest.







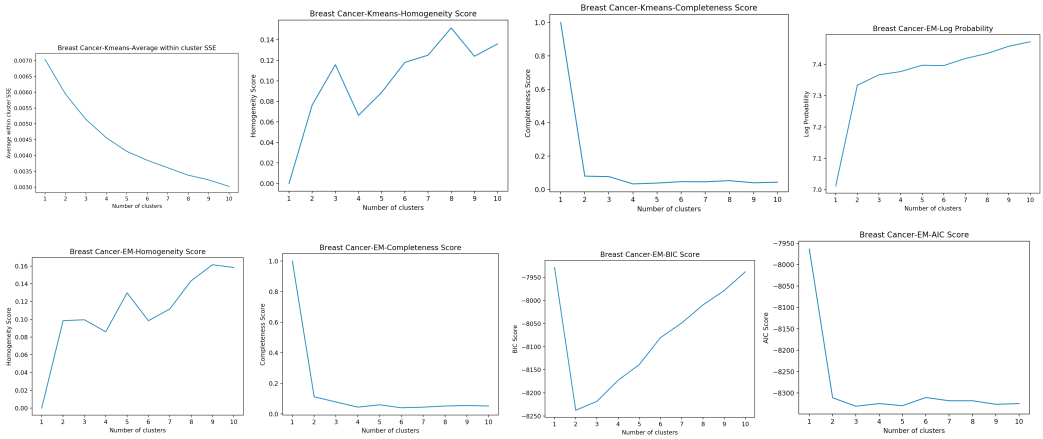


Breast Cancer Letter

For breast cancer dataset, we can see the distribution of kurtosis values, which measures the degree of the non-Gaussianity. The accuracy curve and kurtosis shows that the best components number is 4 for breast cancer dataset. For letter recognition dataset, at the 8th component, the accuracy starts to be flat and the remaining kurtosis start to drop to close to zero, and 8 is chosen as the reserved number of independent components.

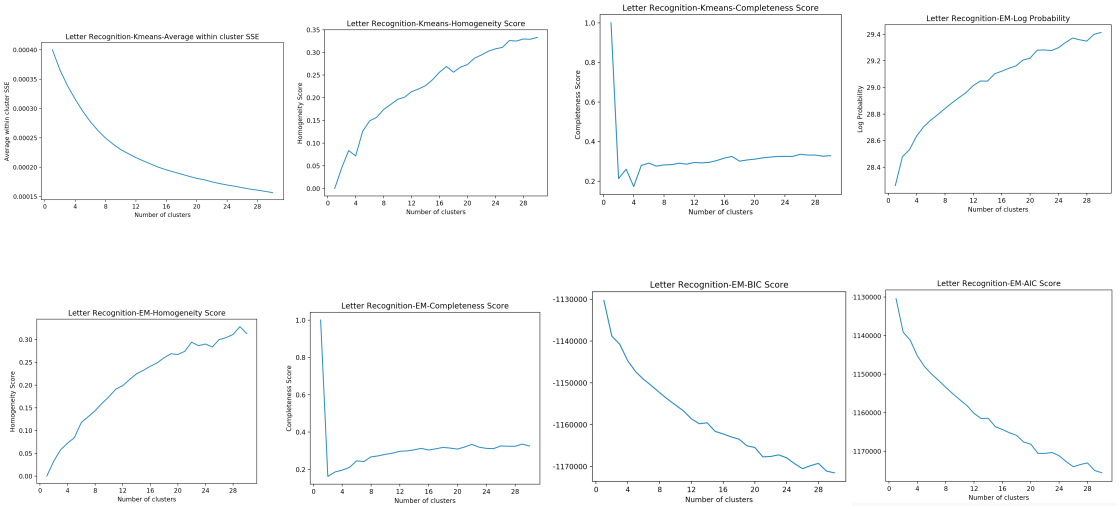
From the reconstruction error diagrams, breast cancer dataset curve has already become flatten at 4 and letter curve is close to flatten. Thus the number of independent components are quit reasonable.

Clustering Analysis



For breast cancer dataset, from the SSE and EM log probability plots, we use elbow method and cluster

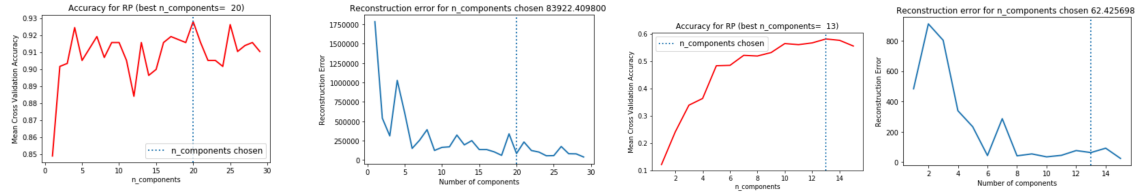
= 2 has the obvious angle. SSE further decreases and log probability increases for ICA in general.



For letter recognition dataset, it is not very obvious to tell the good cluster number using elbow method for SSE and EM log probability plots. From homogeneity and completeness curves, when cluster number is 26, the curve starts to stay flat. This is consistent with the 26 alphabetical letters. It also indicates that ICA helps cluster the dataset closer to the number of classes.

Random Projection

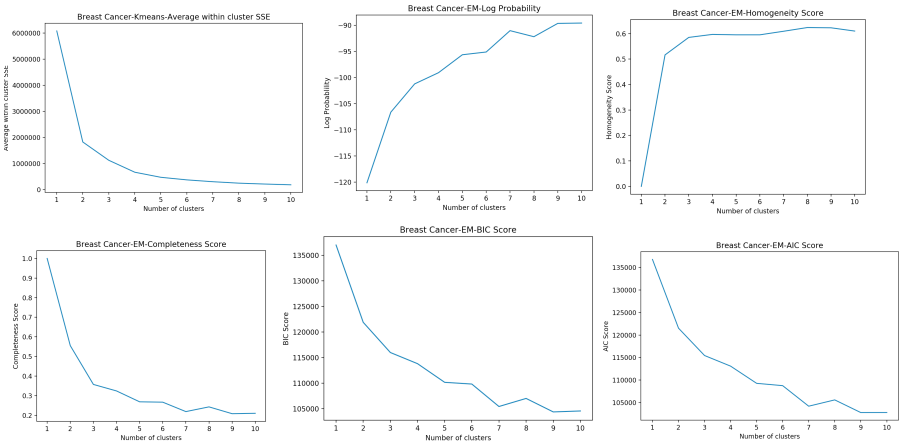
Random projection is a dimensionality reduction method that projects the total number attributes to a lower dimensional space. As opposed to PCA, random projection projects the original input space on a randomly generated Gaussian matrix.



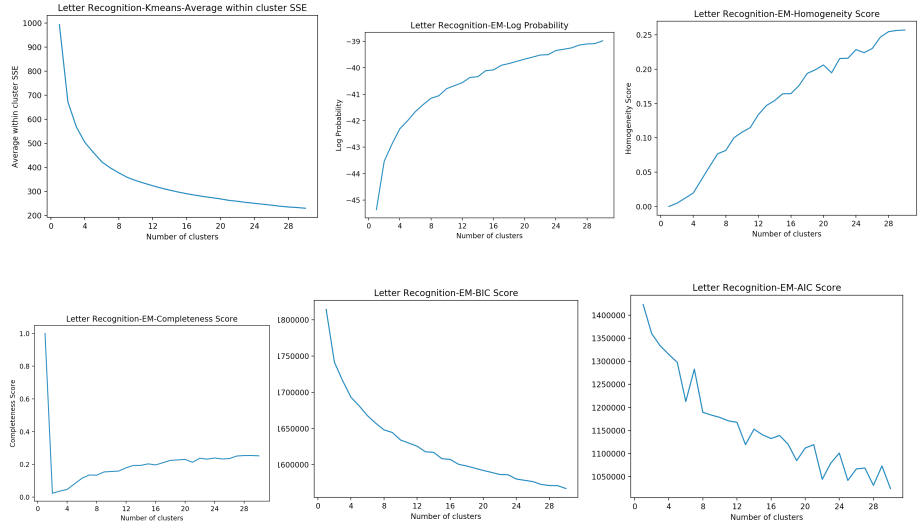
Breast Cancer Letter

From above accuracy plots, for breast cancer, dimension = 20 got highest accuracy and it also got low reconstruction error. Thus, I will pick dimension = 20 for this dataset. For letter reconginition, when dimension = 13, the accuracy begins to decrease and thus we will pick 13 as the number for dimension.

Cluster analysis



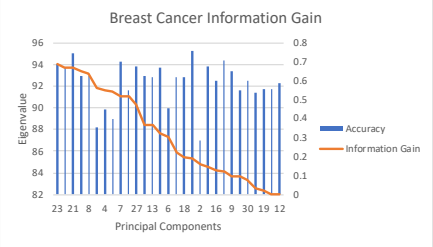
For the breast cancer dataset, we can see from the SSE and EM log probability curves that cluster = 2 is obvious angle by using elbow method. The SSE increases a lot than PCA and ICA and log probability decreases a lot too. This is because in random projections, the projected vectors are chosen randomly, thus variance between instances increases.

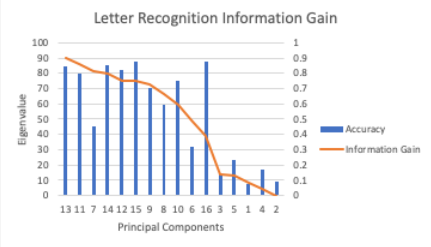


Likewise, the SSE increases for letter datast and log probability decreases. We saw at cluster = 26 in log probability it got high value which matched the class number.

Information Gain

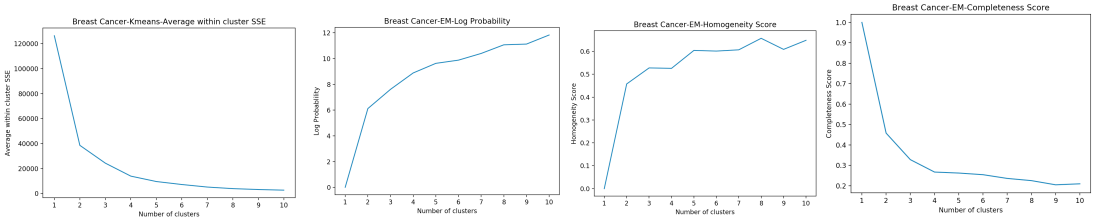
Information gain attribute selector evaluates the attributes by measuring the information gain respecting the class. This algorithm ranks the attributes based on the calculated information gain. We use the same method to drop the attribute on my own and evaluate the performance.



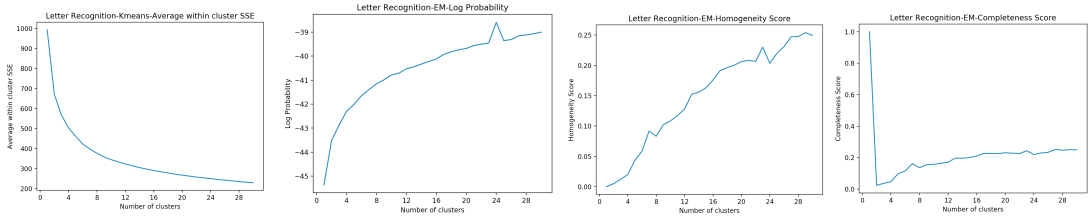


For breast cancer dataset, the last three attributes have information gain = 0, and we can see the peek classification accuracy at 18 is 96.13 and the cutoff information gain is 0.1881. Thus we will select 18 attributes out of 30. For letter recognition dataset, only attribute 2 has 0 information gain. We could find out that at attribute number = 15, the classification accuracy reaches its peak at 88.33, the reaming classification performance stays flat. Since the remaining attribute have relatively low information gain, they do not give much information with respect to classification.

Clustering Analysis



As we can see for the breast cancer datasets, the elbow is obvious at cluster 2. And SSE decrease but log probability also decrease compared with original dataset. That is because we do not transform the data but only select datasets.



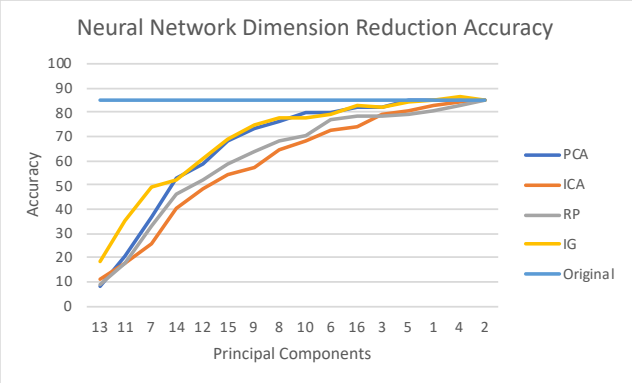
For the letter recognition datatset, we can seet the SSE and log probability values are close to the orginial dataset, since we only delete one attribute and attribute 16 is also useless. And based on the position of spike, we can pick cluster number = 26.

Part 3: Neural Network Performance

Dimensionality Reduction and Neural Network

In this part, we pick letter recognition dataset to train a neural network and with four dimensionality reduction algorithms. I apply four different algorithms on the dataset, then do a forward search from

only one component or attribute to all 16 components or attributes with neural network classifier. The classifier uses default (attribute + class) / 2 nodes in the hidden layer, with learning rate 0.3 and momentum 0.2. The PCA only has 12 components though. We use full dataset as training and did not use cross validation because the dataset is large and cross validation takes much longer time and the neural network model is not the main focus here. All algorithms are implemented in Weka.



	Training Accuracy	Training Time (s)
Original	84.95	19.69
PCA	82.2	16.37
ICA	84.45	17.75
RP	84.85	18.07
IG	86.25	17.63

The plot and table indicates that random projection and information gain algorithms have the best classification performance, but shorter training time. PCA has lower accuracy but also shorter training time. ICA has comparable performance against original dataset and also shorter training time. We also not that IG grows very fast in the beginning and it suggests that by ranking the attributes through information gain, we are able to get the performance faster. PCA can be used as a trade-off for shorter training time though it has a 2% performance decrease.

Clustering and Neural Network

In this part, we explore how performance changes as clustering is introduced as an attribute. We apply two different case here: first, we use clusters as an additional attribute in addition to the original 16 attributes; second, we use clusters as only attributes for the whole dataset. The two methods are implemented in Weka as AddCluster and ClusterMembership filter. For the ClusterMembership filter, the available clustering algorithm is EM.

Cluster As addition Attribute	Training Accuracy	Training Time (s)
Original	84.95	19.69
Cluster = 2 KMeans	87.3	18.61
Cluster = 2 EM	87.3	16.84
Cluster = 15 KMeans	88.85	26.66
Cluster = 15 EM	87.95	26.89

Cluster = 26 KMeans	90.55	35.95
Cluster = 26 EM	90.3	36.53

We can see that adding the cluster helps us increase the accuracy. We pick 2, 15 and 27 as the cluster number. Meanwhile, the training time also siginificanly increases as the cluster number increases. K-Means and EM has similar performance and training time.

Cluster as Only Attribute	Training Accuracy	Training time in seconds
Original Dataset Only	84.95	19.69
Cluster Number = 2, EM	74.9	45.42
Cluster Number = 26, EM	4	8845.31

As we can see, when we add cluster as the only attribute, the number of attribute depends on the number of class * cluster number. Since we have 26 classes, if we choose cluster number is equal to 2, we will have 52 attributes and if we got 26 clusters, there will be 676 attributes. This indeed will dramatically increase the computation complexity due to the curse of dimensionality. Thus adding cluster as the only attribute is not a good method for this particular dataset because it has too many classes. When cluster number is equal to 2, we have lower accuracy and much longer training time. This is because the cluster itself is not giving enough information. When cluster number = 27, the accuracy is only 4%, suggesting using clustering as the only attribute is not a useful technique.

Conclusion

Information gain is shown to have the best accuracy performance among all four dimensionality reduction algorithms. PCA also shows relatively good performance and it has shorter training time. For PCA and ICA, we also found out that the low raking components will have not worthy information and can be discarded for further dimension reduction. Information gain helps identify the more important attributes also achieving very good performance. Random Projection performance varies but it actually shows some good accuracy results. I also found out PCA and ICA transform the data such that it will have lower K-Means SSE and higher EM log probability. The clustering added as an additional attribute generally helps achieve better performance as it provide extra information at cost of extra computation, but using cluster as only attributes depends on the number of classes. If the number of class is too large, it would exponentially increase the computation complexity.

Reference

1. Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for

"Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.

2. Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011

3.Breast Cancer Wisconsin (Diagnostic) Data Set. https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)

4. Letter Recognition Data Set. https://archive.ics.uci.edu/ml/datasets/Letter+Recognition