Unsupervised Learning and Dimensionality Reduction

Georgia Institute of Technology CS7641 Machine Learning Assignment 3

Yan Cai GTID: ycai87

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 explores two clustering algorithms; 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

Breast 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 filed 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 text 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.

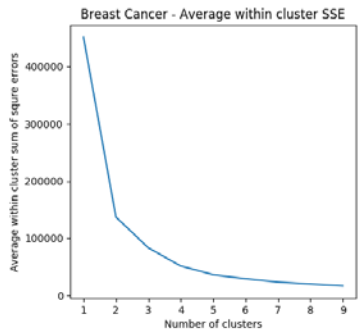
Part 1: Clustering

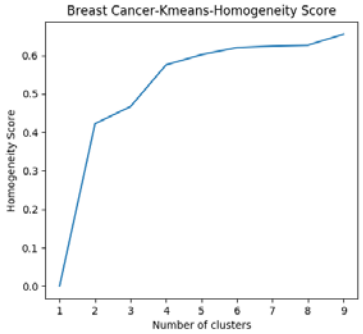
Clustering is a method of grouping the instances together such that instances which belong to same cluster are more similar to each other than those in other clusters. In this section, K-Means clustering and Expectation Maximization (EM) algorithms are explored. In K-Means, Euclidean distance is used because other distance functions 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 distributions. It uses maximum likelihood parameters. EM

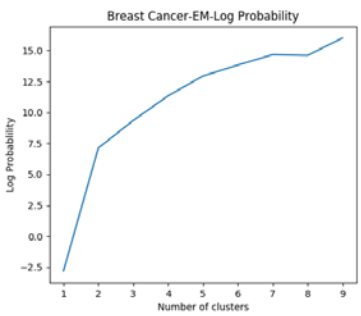
alternates between estimating the log-likelihood of current estimates (E step) and maximizing the likelihood based on the E step (M step).

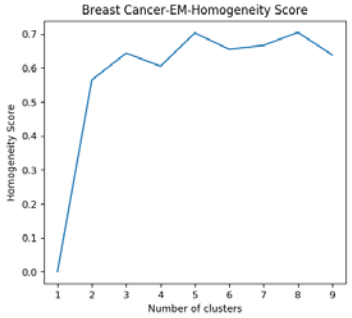
Both K-Means and EM are implemented using Scikit-learn. Clusters are evaluated using average within-cluster sum of square errors for K-means and log likelihood for EM. Homogeneity and completeness and adjusted RAND score are also used to evaluate the cluster. Homogeneity describes how each cluster contains 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) are provided to evaluate EM. In our implementation for GM, the covariance type is diagonal.

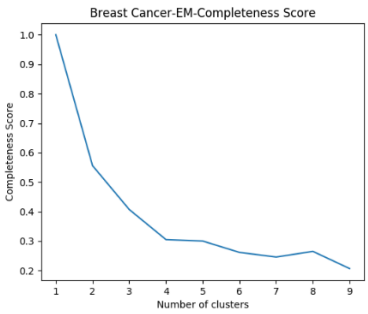
Breast Cancer

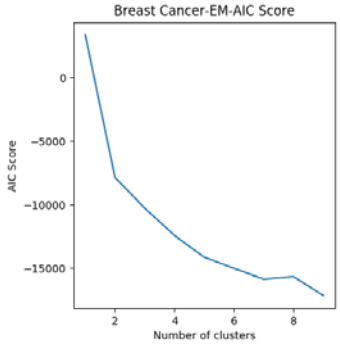


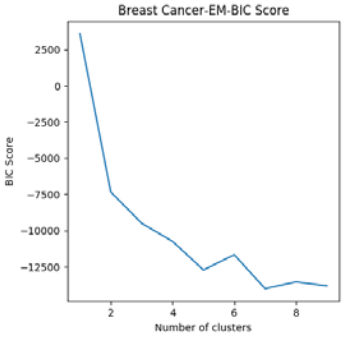






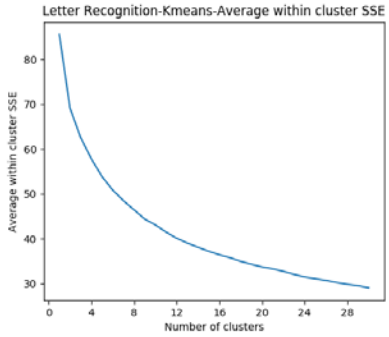


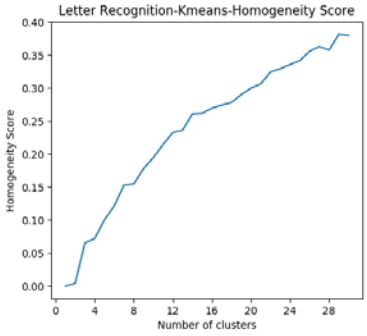


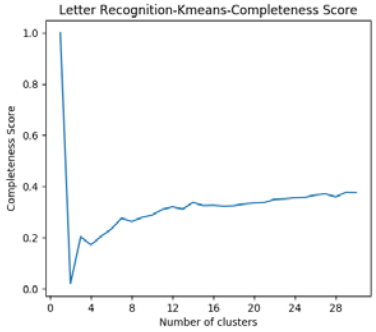


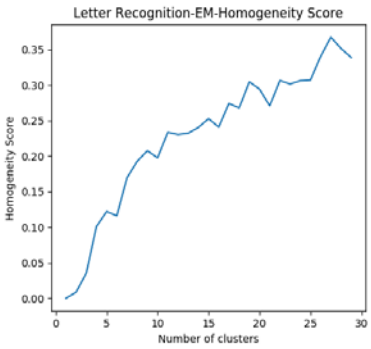
From the above plots, we can use the elbow method to evaluate the cluster. For almost all the plots, the elbow methods indicate that cluster = 2 seems to be the best choice, that is because when cluster number = 2, we can see the angle in the SSE and log probability curves and after that the curve starts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

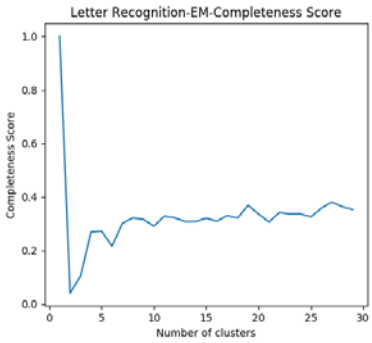
Letter Recognition

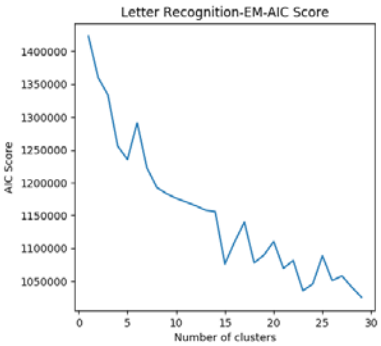












The k-means SSE curve is pretty smooth, using the elbow method is not easy to identify the angle. In terms of completeness score, we actually see the score improves as the number of cluster increases. This is because the clustering algorithms recognizes more than 26 different letters and some letters may have more than one appearance, thus adding the cluster numbers actually considers different appearances and styles of single letters and differentiate it in more detail. In log probability and AIC and BIC scores, we see spikes in the curve when cluster = 22, 25, 27. Given the class = 26, it is reasonable to assume the good cluster numbers are around 26. 25 is picked as the best cluster number for this dataset.

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 (IG).

Methodology

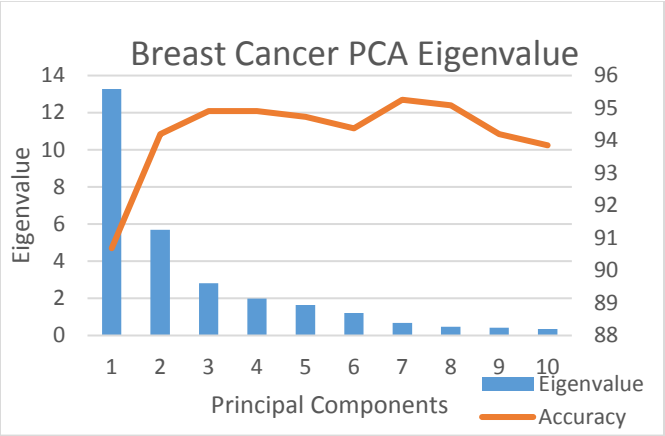
All dimension reduction algorithms use Weka and are applied on both datasets. The procedure is as follows:

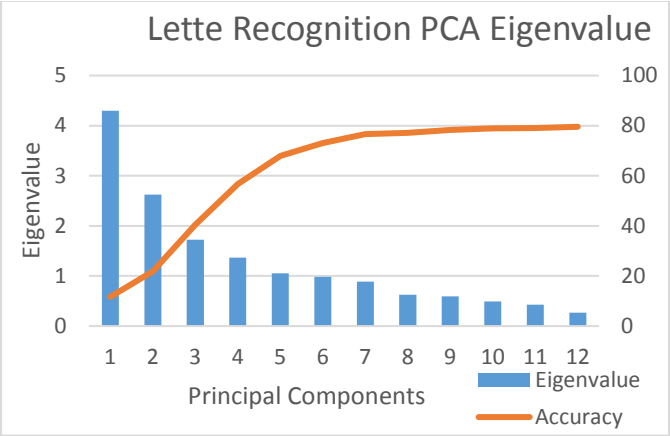
1. Apply the dimension reduction algorithm, get the newly transformed dataset.
2. Apply J48 classifier to get the optimum choice of number of components for each algorithm. The newly transformed feature is removed one by one until the classification accuracy drops. 10 fold cross validation is used.
3. 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. We use Weka to apply PCA. The maximum number of attributes in names is 5.

Dimension Reduction Analysis



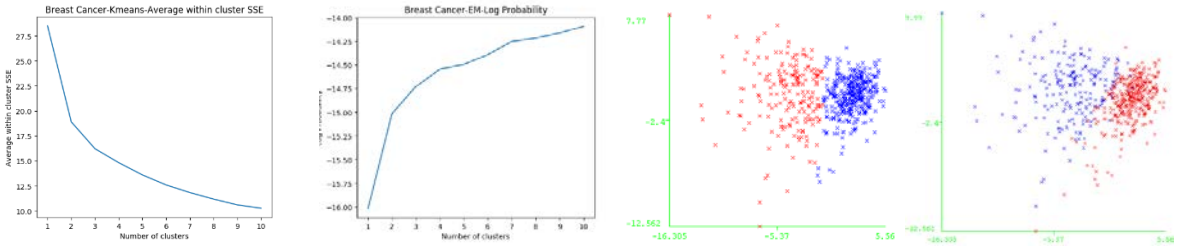


For both datasets, the eigenvalues for the last few components are relatively small, giving the possibility of removing them to apply classification.

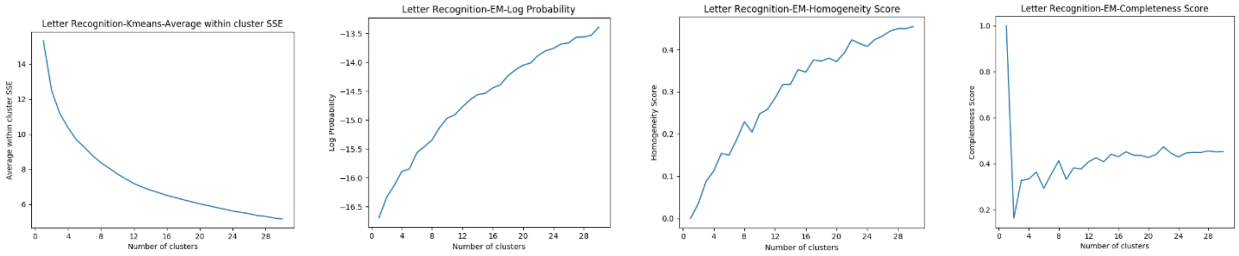
From the classification results, we can see for breast cancer dataset, the accuracy starts to drop after the number of principal components = 7, this suggests that the remaining principal components actually contain some noise that impacts the classification. For letter recognition dataset, the accuracy becomes flat when the number of principal components = 7. It suggests the remaining components do not contain worthy information that helps classification. Thus, we will select the 7 for either datasets as the choice of number of principal components

Clustering Analysis

Clustering algorithms are applied on the transformed data after PCA with number of components = 5 for breast cancer datasets and 7 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. The above right two figure shows the k-means and EM clusters based on the first and second principal components.

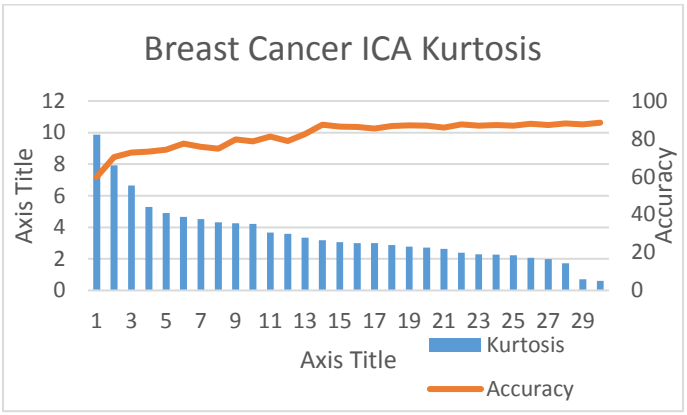


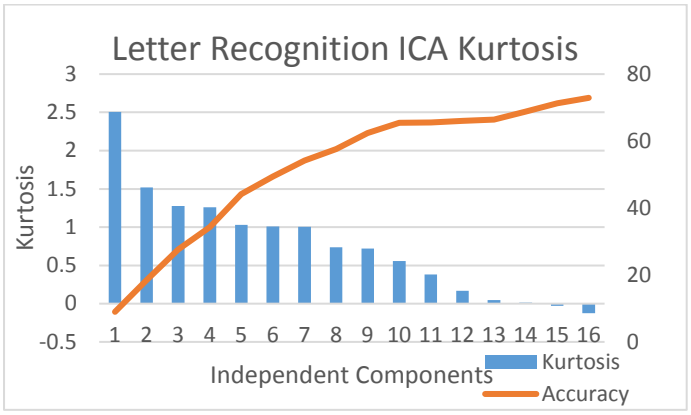
For letter recognition dataset, again the SSE decreases and log probability increases with PCA transformed data. Homogeneity and completeness score curves become smoother as well. It is not easy to identify the best cluster number by elbow method, but as we look at the spikes in the curve, we see big spikes when cluster = 23. This is smaller compared with non PCA EM clustering, which is possibly because the PCA removes some unworthy information and creates smaller clusters for letters.

Independent Component Analysis

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

Dimension Reduction Analysis

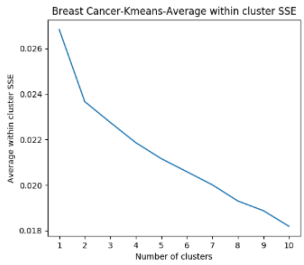


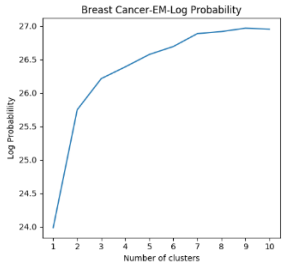


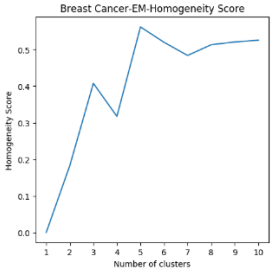
For breast cancer dataset, we can see the distribution of kurtosis values, which measures the degrees of the non-Gaussianity. The accuracy curve indicates that components after 14 whose kurtosis values are smaller than 3.17 do not contribute much worthy information to improve classification accuracy, as the accuracy starts to stay flat when more independent components are added, thus 14 is chosen as the independent component number . For letter recognition dataset, at the 10^{th} component, the accuracy starts to be flat and the remaining kurtosis start to drop to close to zero, though the accuracy increases a little bit in the end, but given their low kurtosis value, 10 is chosen as the reserved number of independent components

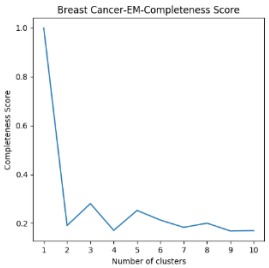
Clustering Analysis

Clustering analysis is applied on dataset where breast cancer has 14 components and letter recognition has 10 components.

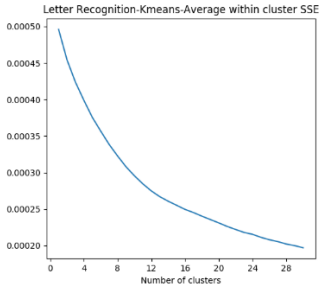


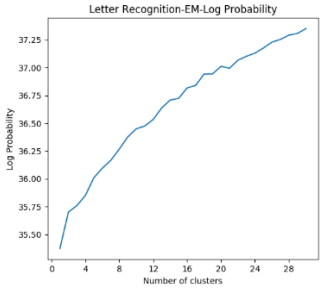


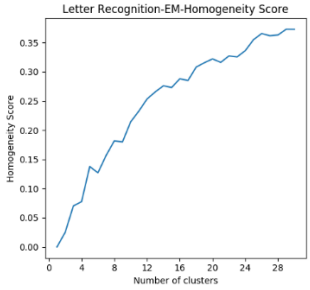


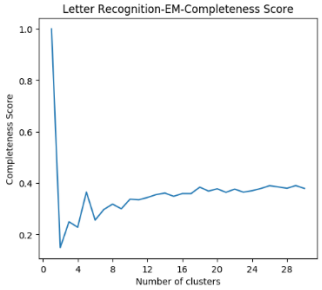


For breast cancer dataset, from the SSE and EM log probability plots, we use elbow method and cluster = 2 has the obvous angle. SSE further decreases and log probability increases for ICA in geneneral.







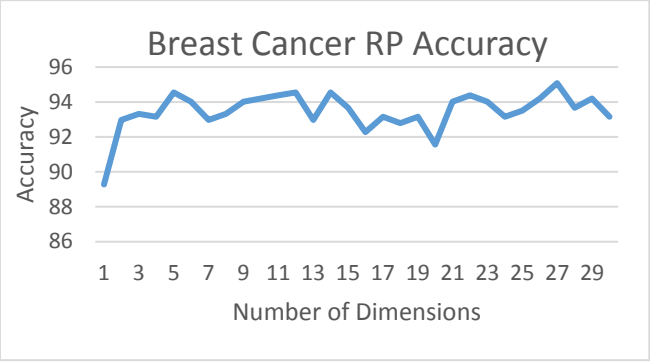


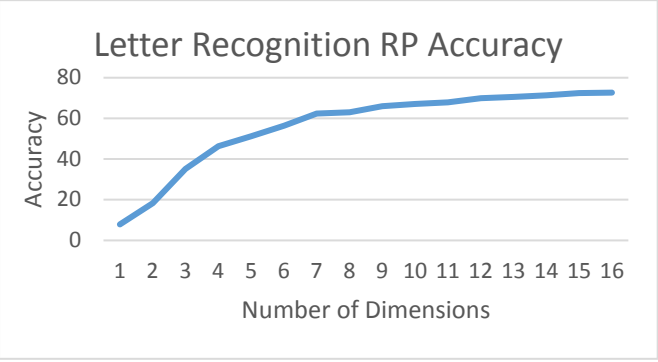
For letter recognition dataset, it is not very obvious to tell the good cluster number using elbow method for SSE and EM log probability plots. From homogeneity and completeness curves, when cluster number is 26, the curve starts to stay flat. This is consistent with the 26 alphabetical letters. It also 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. Random projection is implemented in Weka. We first set the dimension number to the number of attributes, then remove the attribute one by one to get the classification accuracy.

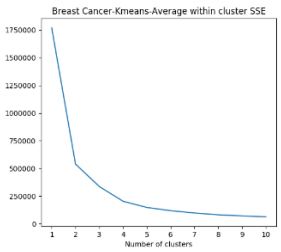
Dimension Reduction Analysis

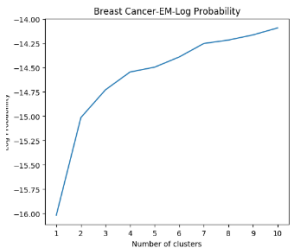


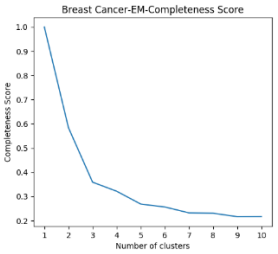


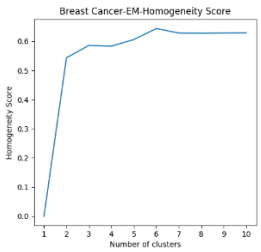
From accuracy plots, for breast cancer, dimension = 5 achieves 94.55% accuracy and the remaining spikes have similar accuracy number. Thus we pick dimension = 5 for this dataset. For letter recognition, when dimension = 12, the accuracy starts to be flat, thus we pick 12 as the dimension number for this dataset.

Clustering Analysis

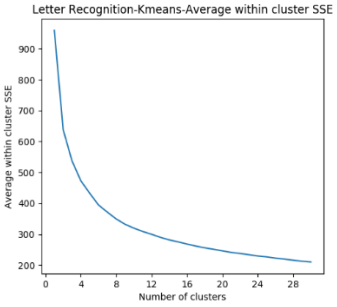


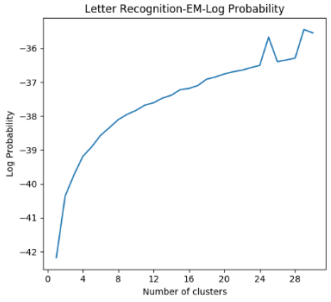


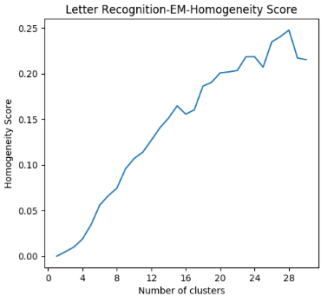


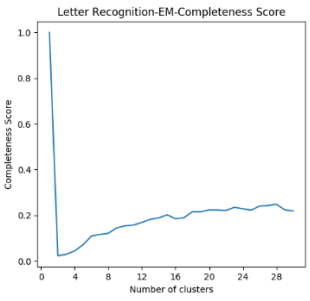


For the breast cancer dataset, we can see from the SSE and EM log probability curves that cluster = 2 is the obvious angle by using elbow method. The SSE increases a lot than PCA and ICA and log probability decreases a lot too. This is because in random projections, the projected vectors are chosen randomly,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







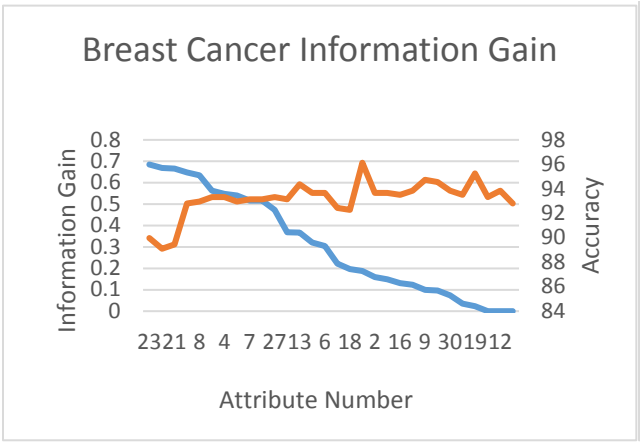


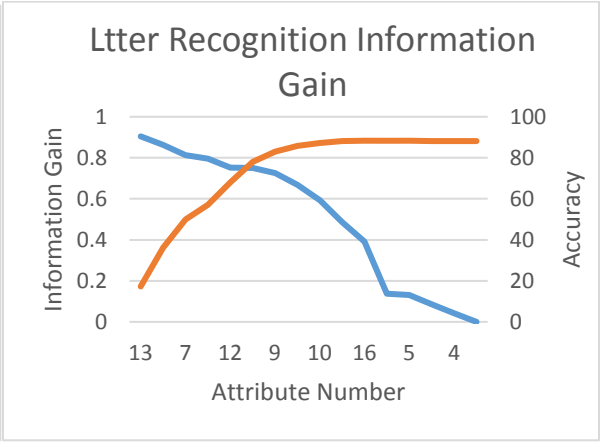
Likewise, the SSE increases for letter recognition dataset and log probability decreases. We saw peaks at cluster = 25 in log probability and dip at homogeneity score. Give then class number is 26, the cluster = 25 is very close to the class number and thus we pick 26 as the cluster number for this dataset.

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 one my one and evaluate the performance.

Dimensionality Reduction Analysis

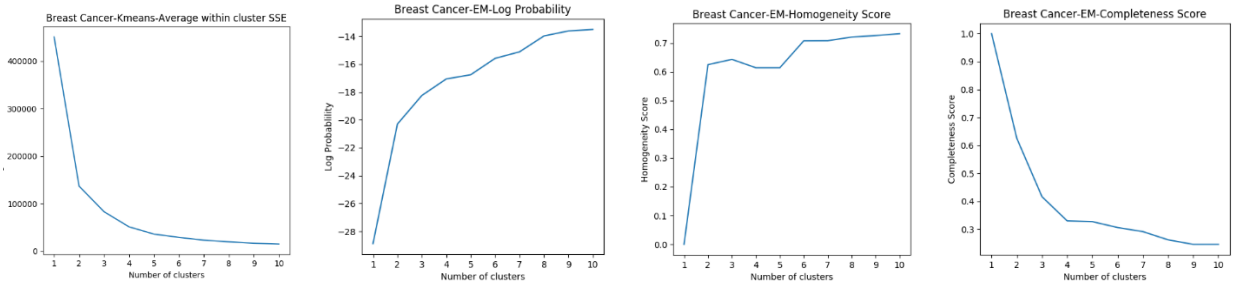




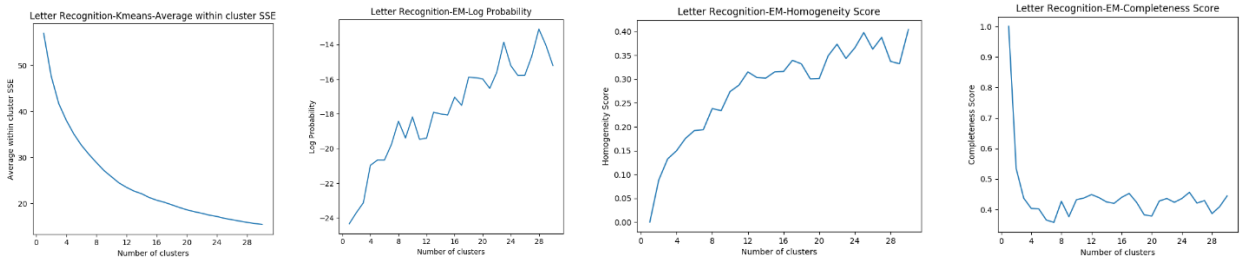
For breast cancer dataset, the last three attributes have information gain = 0, and we see the peak classification accuracy 96.13% when the cutoff information gain is 0.1881. The remaining attributes contribute either positively or negatively to

the classification accuracy, but does not surpass the peak accuracy, thus we select 18 attributes out of 30. For letter recognition dataset, only attribute 2 has 0 information gain. We find out when attribute number = 11, the classification accuracy reaches its peak at 88.33, the remaining classification performance stays flat. Since the remaining attributes 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. We also note that the SSE and log probability values are close to original dataset. That is because we do not transform the data but only select datasets.

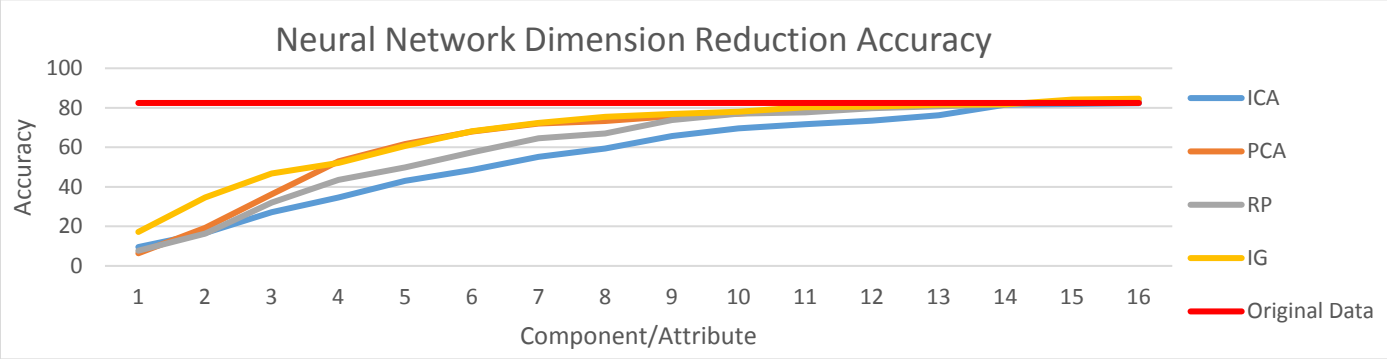


For the letter recognition dataset, we can also see the SSE and log probability values are close to the original dataset. There are a few spikes in log probability and homogeneity score curves, indicating there are more than one choices of cluster number. Based on the position of the spike, we pick cluster number = 25.

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. We 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 is using default (attribute + class) / 2 = 21 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 do 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. The baseline is same neural network classifier with original dataset, and the accuracy is 82.455%.



	Training Accuracy Percentage	Training Time in seconds
Original Dataset	82.455	138.751
PCA	80.46	129.721
ICA	82.315	137.331
Random Projection	83.38	141.634
Information Gain	84.595	144.589

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 note 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 good 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 cases 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 Percentage	Training Time in seconds
Original Dataset Only	82.455	138.751
Cluster Number = 26, K-Means	88.26	300.82
Cluster Number = 26, EM	87.27	303.86
Cluster Number = 2, K-Means	83.385	142.74
Cluster Number = 2, EM	83.445	137.65
Cluster Number = 15, K-Means	86.07	242.24
Cluster Number = 15, EM	86.325	249.54

We can see that adding the cluster helps increase the accuracy. We pick 2, 15 and 26 as the cluster number. Meanwhile, the training time also significantly increases as the cluster number increases. K-means and EM has similar performance and training time.

Cluster As Only Attribute	Training Accuracy Percentage	Training Time in seconds
Original Dataset Only	82.455	138.751
Cluster Number = 2, EM	74.525	392.27
Cluster Number = 26, EM	4.375	26131.04

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 as alphabetical letters, if we choose cluster number = 2, we will have 52 attributes. If we have 26 clusters, there will be 676 attributes. This indeed will dramatically increase the computation complexity due to the curse of dimensionality. In the experiment, it takes more than 7 hours to get the result. 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 = 2, we have lower accuracy and much longer training time. This is because the cluster itself is not giving enough information. When cluster number = 26, the accuracy is only 4.375%, 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 find out the low ranking components (by eigenvalue or kurtosis) do not have worthy information and can be discarded for further dimension reduction. Information gain helps identify the more important attributes, also achieving very good performance. The Random projection performance varies but it actually shows some good accuracy results. We also find out PCA and ICA transform the data such that it has lower K-Means SSE and higher EM log probability. The clustering added as an additional attribute generally helps achieve better performance as it provides extra information at the 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.

References

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