ASSIGNMENT 3: Unsupervised Learning

Habeebuddin Mir
NUID: 002713929

Abstract

This report explores unsupervised machine learning algorithms as part of Assignment 3. The first two are clustering algorithms and the next three are dimensionality reduction algorithms. K-means and Expectation Maximization are the two clustering algorithms used. The three dimensionality reduction techniques are principal component analysis (PCA); independent component analysis (ICA and information gain (IG). The report is organised into three parts: part one explores clustering algorithms; part two applies the three-dimensionality reduction techniques and clusters the dimension-reduced data; part three applies both dimensionality reduction methods and clustering algorithms and uses the new data to train neural networks.

Datasets

Breast cancer Wisconsin diagnostic dataset and letter recognition dataset are used in this assignment.

Breast cancer dataset

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

The breast cancer dataset contains two classes of 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 is thus a good candidate for dimensionality reduction.

Letter recognition

Computer vision and image recognition are interesting filed in machine learning. Many industries use character recognition to help with process automation and improvement. The scanner is able to use letter recognition to convert text images to text.

The dataset has 26 classes, and each class is one letter in the 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 is thus a good candidate for dimensionality reduction and neural networks.

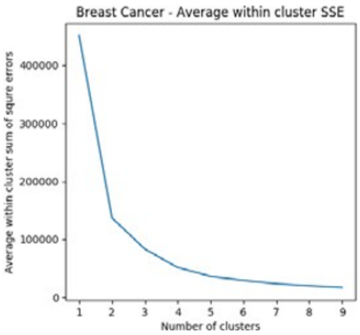
Part 1: Clustering

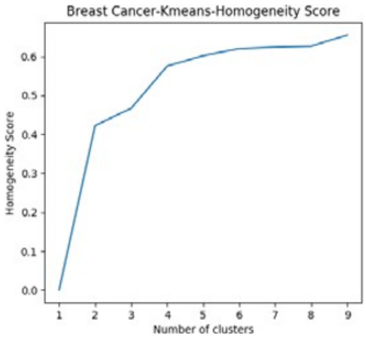
Clustering is a method of grouping the instances together such that instances which belong to the same cluster are more similar to each other than those in other clusters. In this section, K-Means clustering and Expectation Maximization (EM) algorithms are explored. In K-Means, Euclidean distance is 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 the centroid is equal to the sum of pairwise squared Euclidean distances divided by the number of points.

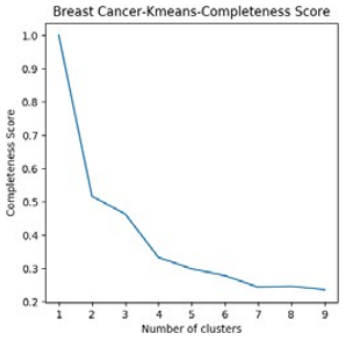
In contrast to K-Means, EM is structured with probability distributions. It uses maximum likelihood parameters. EM alternates between estimating the log-likelihood of current estimates (E step) and maximizing the likelihood based on the E step (M step).

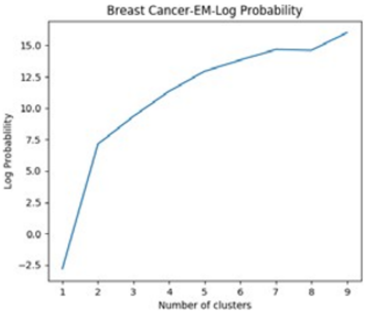
Both K-Means and EM are implemented using Scikit-learn. Clusters are evaluated using the average within-cluster sum of square errors for K-means and log-likelihood for EM. Homogeneity and adjusted RAND score are also used to evaluate the cluster. Homogeneity describes how each cluster contains only members of a single class, completeness describes the degree to 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.

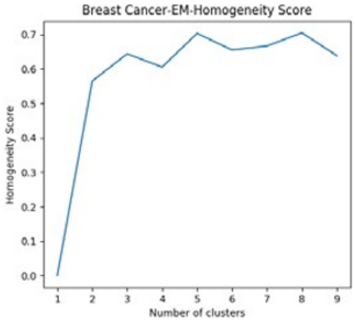
Breast Cancer

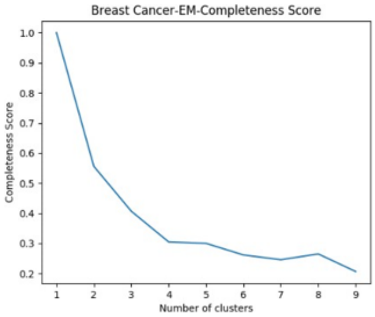


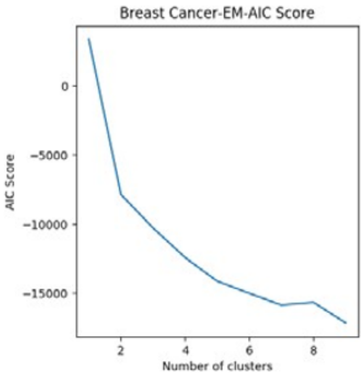


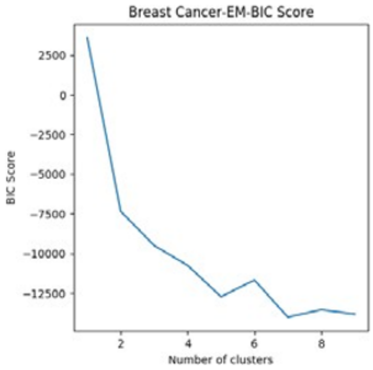






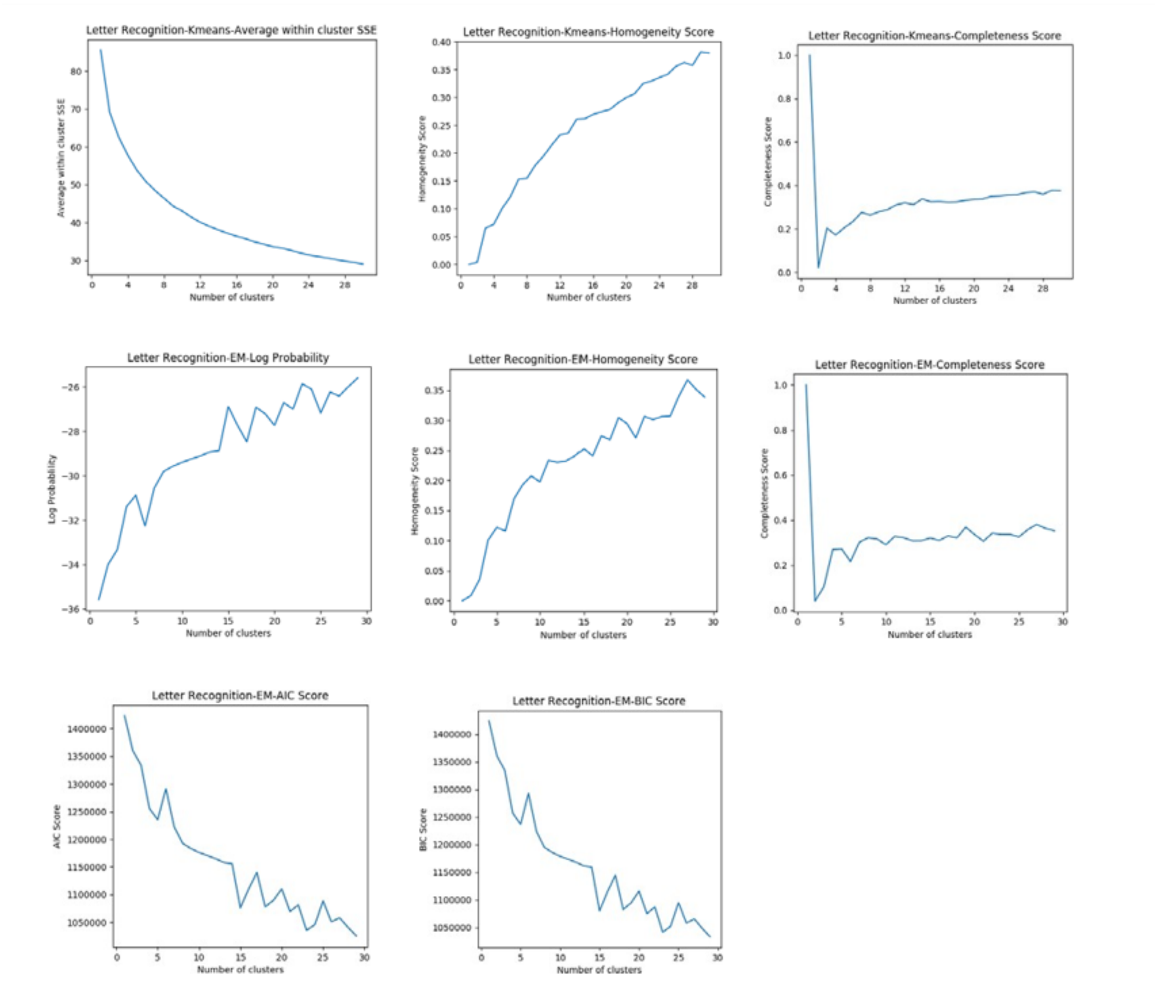






From the above plots, we can use the elbow method to evaluate the cluster. For almost all the plots, the elbow methods indicate that cluster = 2 seems to be the best choice, that is because when cluster number = 2, we can see the angle in the SSE and log probability curves and after that, the curve starts 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 the completeness score, we actually see the score improves as the number of clusters increases. This is because the clustering algorithms recognize 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 clustering = 22, 25, and 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 information gain (IG).

Methodology

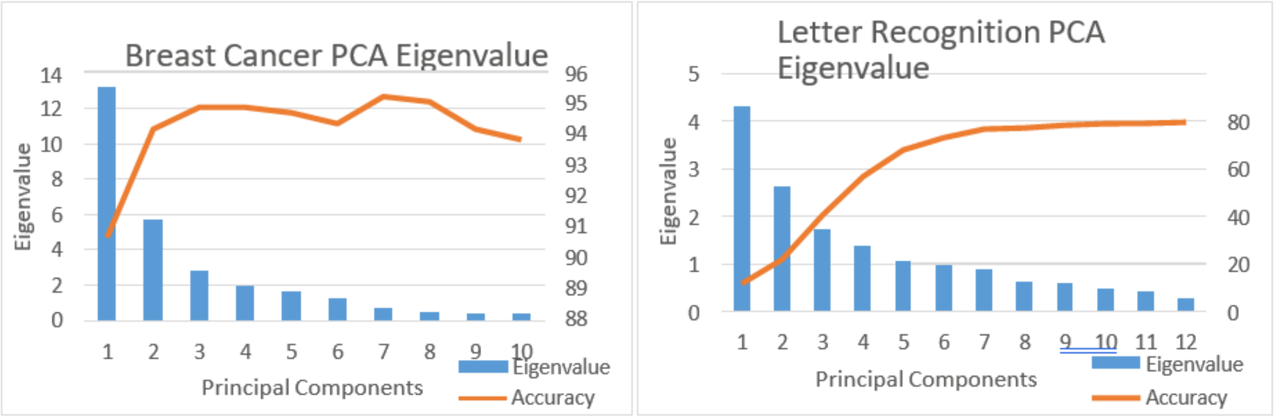
All dimension reduction algorithms use Weka and are applied to both datasets. The procedure is as follows:

1. Apply the dimension reduction algorithm, and get the newly transformed dataset.
2. Apply the J48 classifier to get the optimum choice of the 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 principal 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

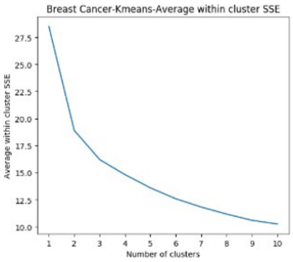


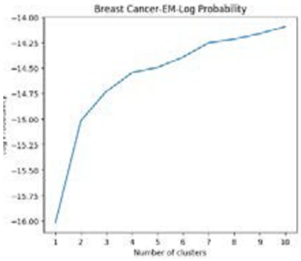
For both datasets, the eigenvalues for the last few components are relatively small, giving the possibility of removing them to apply classification.

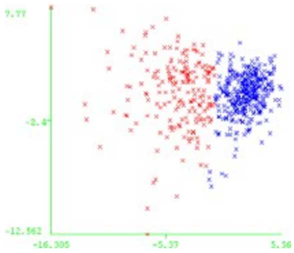
From the classification results, we can see for the breast cancer dataset, the accuracy starts to drop after the number of principal components = 7, this suggests that the remaining principal components actually contain some noise that impacts the classification. For the letter recognition dataset, the 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 dataset as the choice of the number of principal components.

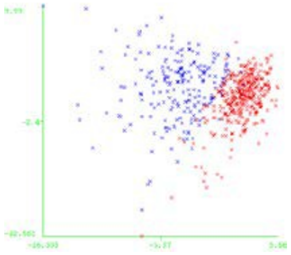
Clustering Analysis

Clustering algorithms are applied to the transformed data after PCA with the number of components = 5 for breast cancer datasets and 7 for letter recognition datasets.

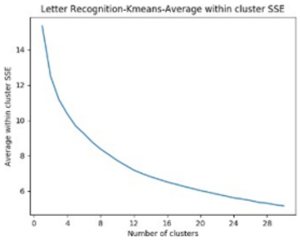


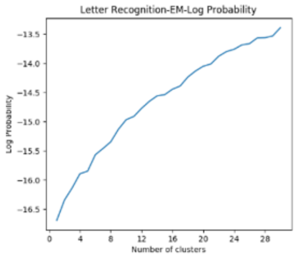


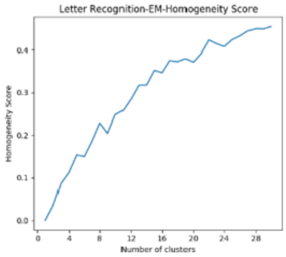


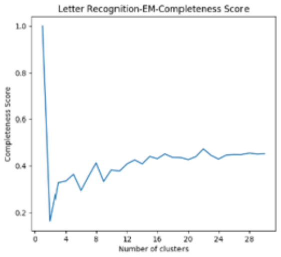


For the breast cancer dataset, as we can see from SSE and log probability, the curve has its angle when clustering number = 2. PCA-transformed data has similar performance curves as the original dataset. However, with PCA, SSE is lowered and log probability is increased. This indicates that PCA makes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







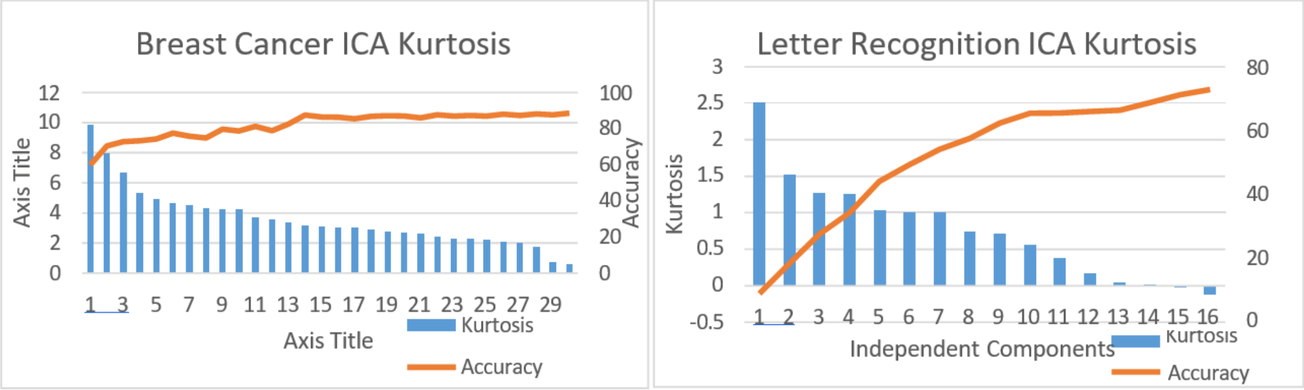


For the letter recognition dataset, again the SSE decreases and log probability increases with PCA-transformed data. Homogeneity and completeness score curves become smoother as well. It is not easy to identify the best cluster number by the elbow method, but as we look at the spikes in the curve, we see big spikes when cluster = 23. This is smaller compared with non-PCA EM clustering, which is possible 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 finding independent components of the original data. We use fast ICA in Weka. The independent components are sorted by kurtosis values from highest to lowest

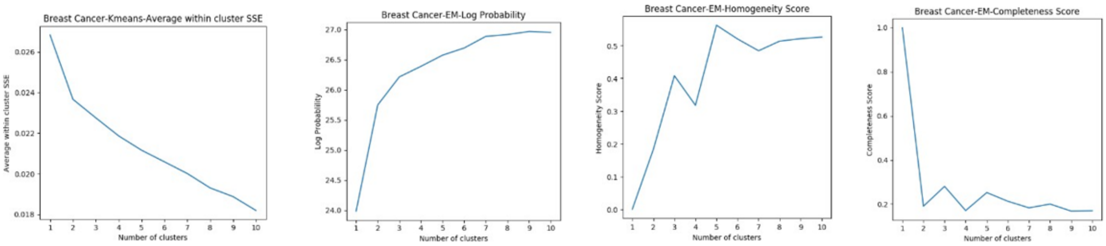
Dimension Reduction Analysis



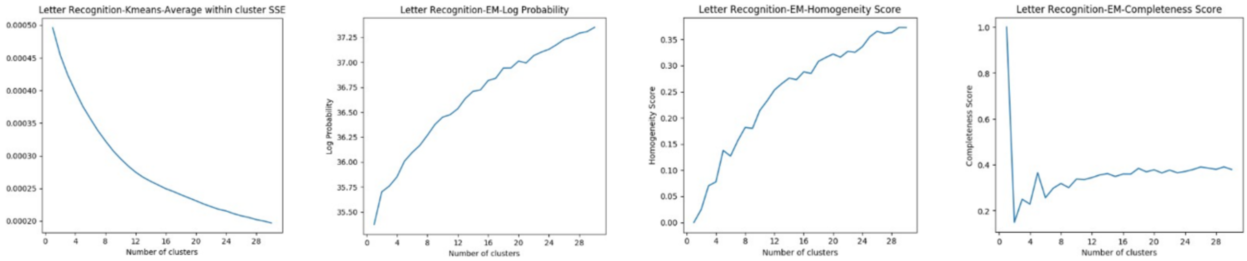
For the breast cancer dataset, we can see the distribution of kurtosis values, which measures the degrees of the non-Gaussianity. The accuracy curve indicates that components after 14 whose kurtosis values are smaller than 3.17 do not contribute much worthy information to improve classification accuracy,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

Clustering Analysis

Clustering analysis is applied on a 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 obvious angle. SSE further decreases and log probability increases for ICA in general.

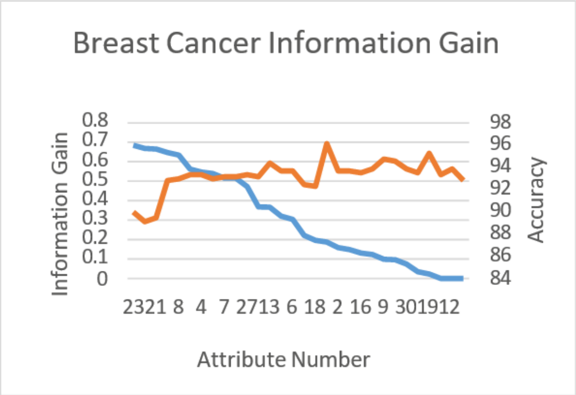


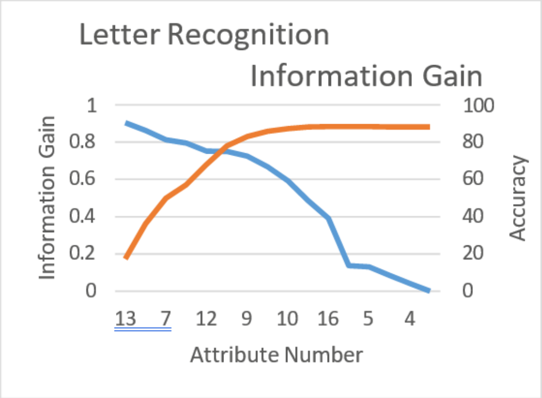
For the letter recognition dataset, it is not very obvious to tell the good cluster number using the elbow method for SSE and EM log probability plots. From homogeneity and completeness curves, when the cluster number is 26, the curve starts to stay flat. This is consistent with the 26 alphabetical letters.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

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 attributes one by one and evaluate the performance.

Dimensionality Reduction Analysis

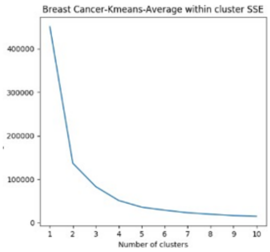


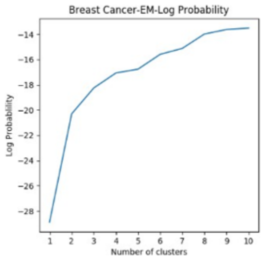


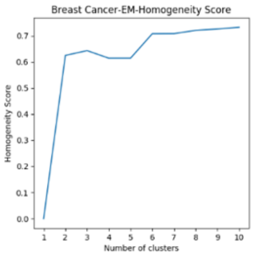
For breast cancer dataset, the last three attributes have information gain = 0, and we see the peak classification accuracy 96.13% when the cutoff information gain is 0.1881. The remaining attributes contribute either positively or negatively to

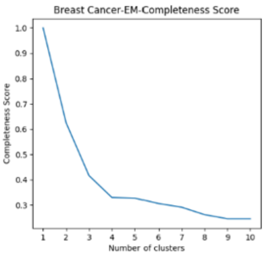
For the breast cancer dataset, the last three attributes have information gain = 0, and we see the peak classification accuracy of 96.13% when the cutoff information gain is 0.1881. The remaining attributes contribute either positively or negatively to the classification accuracy but do not surpass the peak accuracy, thus we select 18 attributes out of 30. For the letter recognition dataset, only attribute 2 has 0 information gain. We find out that when attribute number = 11, the classification accuracy reaches its peak at 88.33, and 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

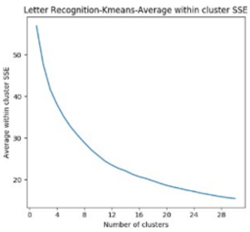


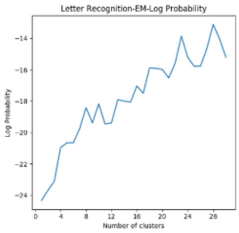


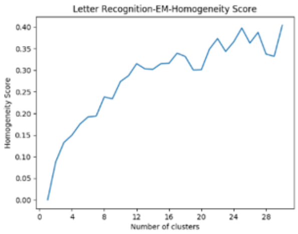


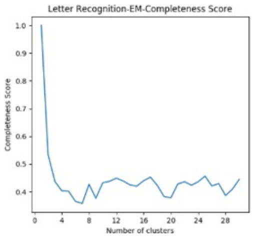


As we can see for the breast cancer datasets, the elbow is obvious at cluster = 2. We also note that the SSE and log probability values are close to original dataset. That is because we do not transform the data but only select datasets.







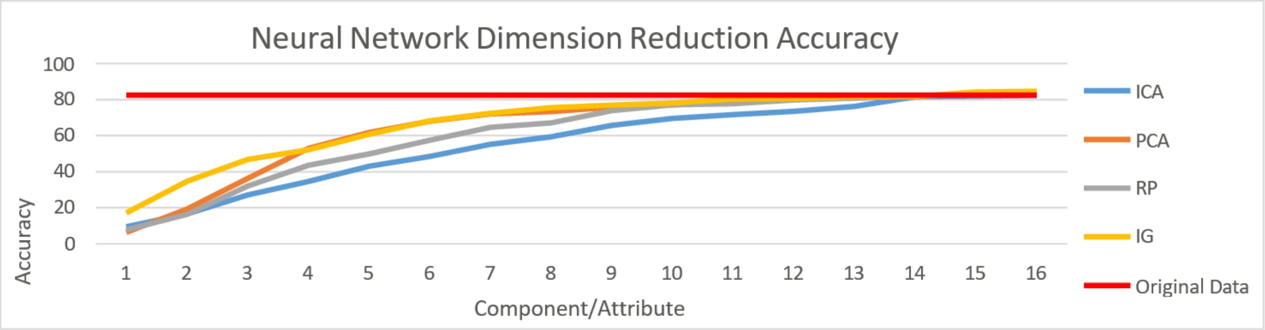


For the letter recognition dataset, we can also see the SSE and log probability values are close to the original dataset. There are a few spikes in log probability and homogeneity score curves, indicating there is more than one choice of cluster number. Based on the position of the spike, we pick cluster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

Part 3: Neural Network Performance

Dimensionality Reduction and Neural Network

In this part, we pick the letter recognition dataset to train a neural network with the three dimensionality reduction algorithms. We apply different algorithms to the dataset, then do a forward search from only one component or attribute to all 16 components or attributes with a neural network classifier. The classifier is using the 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 the 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 focus here. The baseline is the same neural network classifier as the original dataset, and the accuracy is 82.455%.

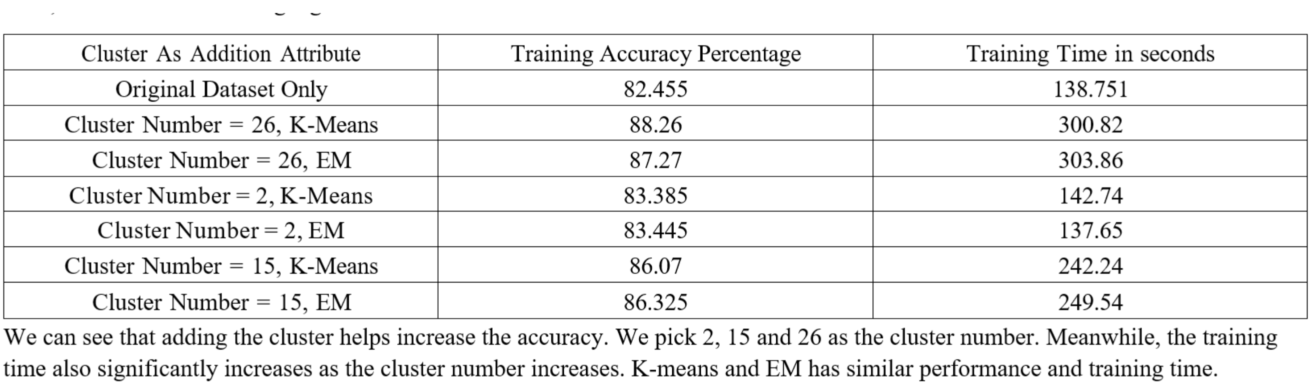


	Training Accuracy Percentage	Training Time in seconds
Original Dataset	82.455	138.751
PCA	80.46	129.721
ICA	82.315	137.331
Information Gain	84.595	144.589

The plot and table indicate that the information gain algorithm has the best classification performance, but a shorter training time. PCA has lower accuracy but also shorter training time. ICA has comparable performance against the original dataset and 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 good performance faster. PCA can be used as a trade-off for shorter training time though it has a 2% performance decrease.

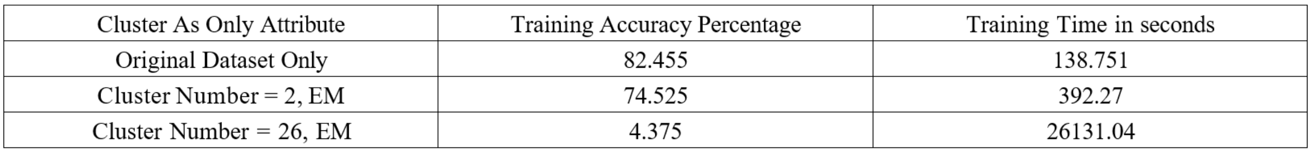
Clustering and Neural Network

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 attributes 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 a shorter training time. For PCA and ICA, we also find out that 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. We also find out that 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 depend on the number of classes. If the number of classes 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