Abhay Dalmia
1st April 2018

CS 4641 – Project 3 Unsupervised Learning

1. Introduction

For this assignment, the spambase dataset from the first assignment was used and a new letter recognition dataset was used. This new dataset was used because it has specific properties that would lead to interesting results in this assignment (as described below).

1.1. Spambase Dataset

The spambase dataset from project 1 was used for this project. This dataset uses 57 attributes to predict whether an email is considered "spam" or not. With the hundreds of important emails that we get every week, it is important for email services to sort out the spam. This dataset counts the frequency of word considered to appear in abnormal frequency in spam emails (like "credit" and "money") along with other important features, like the frequency of special characters and the frequency of capital words, to decide if the email is considered spam.

Because this data set has a large number of attributes, it will provide an interesting opportunity to evaluate the feature reduction algorithms. Given that the data only has two classifications ("spam" and "not spam"), the clustering algorithms are likely to be *less interesting* and any inaccuracies in clustering are bound to be less obvious.

1.2. Letter Recognition Dataset

The letter recognition dataset contains an aggregation of 20 fonts and sets attributes based on these fonts. Attributes range from things like the height of the box of the letter, total number of pixels etc. Based on these, the letter is predicted. It has 17 different attributes and can be classified into 26 different letters. Due to the large number of clusters that are required to separate the data, this dataset should lead to *more interesting* clusters than the previous dataset and clustering algorithms can be sufficiently evaluated. The low number of attributes is likely to lead to indifference in the feature reduction (or at the very least, the results will not be as contrasting as those for the spambase dataset above).

2. Clustering Algorithms

Before performing clustering on both datasets, both the k-means and expectation maximization clustering algorithms used Euclidean distance as the default measurement to measure the similarity between the attributes of the datasets. The reasons for this were to allow for easier analysis as consistency is maintained between the two clustering algorithms. A Manhattan distance measurement could have been better for the letter recognition dataset due to its smaller number in attributes, therefore making it easier for clustering algorithms to detect similarity. For both K means and expectation maximization, the number of K was varied and the accuracy was detailed.

2.1. K-Means Clustering

In the K-means clustering algorithm, data points are classified into k clusters. In this iterative process, the data point is assigned to the "closest" cluster, and then a new centroid is decided for that cluster. The code was iterated over several K-values and the elbow method was used to pick the best K. (The best K would not necessarily be the most accurate because at that point, the data could be over fit. Additionally, after some K value, there is diminishing return in increased accuracy with increased K, in which case it is not worth it to consider a larger K.

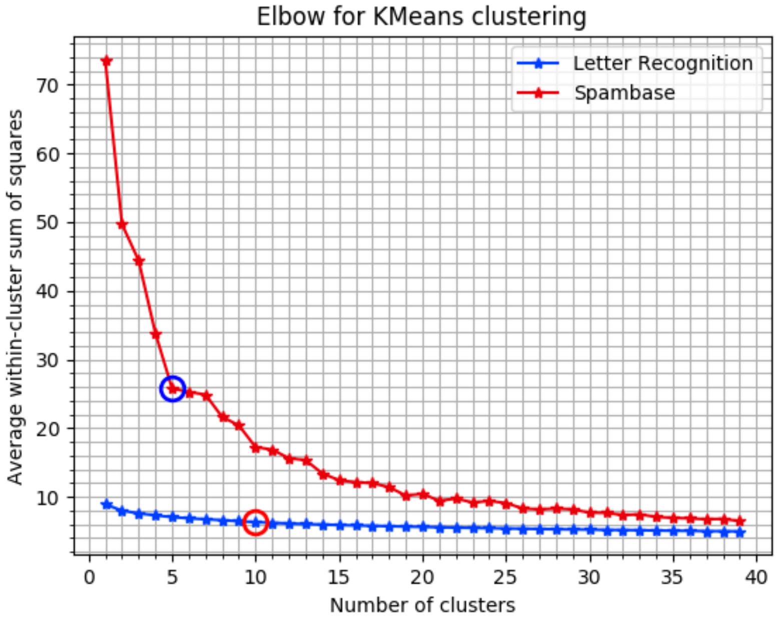


Figure 1: Accuracy of K-means clustering for the two datasets

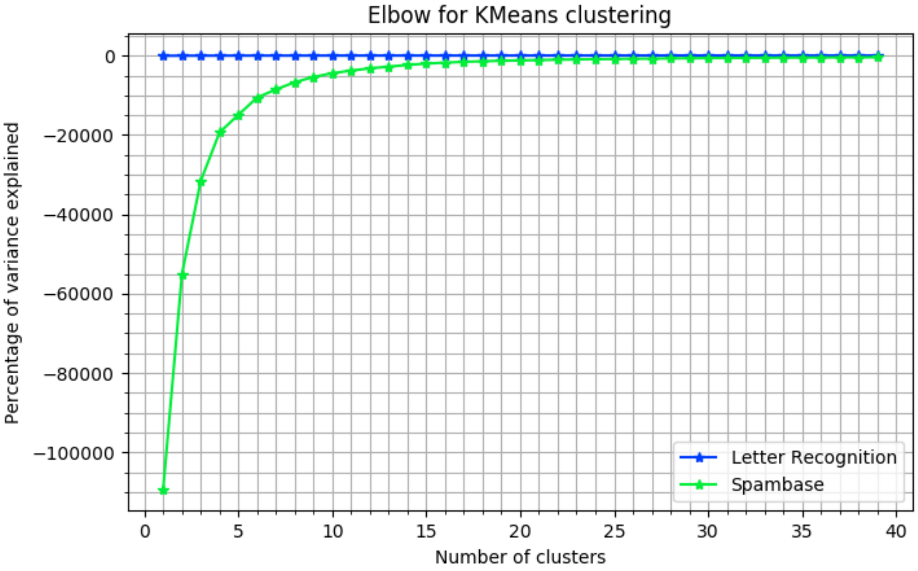


Figure 2: Variance of K-means clustering for the two datasets

Both the figures above show expected behavior, although the behavior when comparing spambase to letter recognition is unexpected. As the number of clusters is increased, accuracy increases (or in figure 1, error decreases), and variance decreases, as expected.

For the letter recognition dataset, we see smaller improvements in sum of squared errors and almost no improvement in variance as the number of clusters are increased. This can most likely be attributed to the fact that this dataset is has an almost uniform distribution – there are 700 to 750 examples for each letter label. Therefore, for increasing clusters from 1 – 40, there is almost constant variance for each model.

The results of the clustering for Spambase were quite unexpected. For instance, Spambase has a high kurtosis and the classification was skewed to the right as there may have been more redundant information in the attributes, however, compared to letter recognition, for both K-means and EM, it performed with much lower accuracy. The lower accuracy is surprising because it only classifies into two values (binary classification) as opposed to the letter recognition dataset that classifies 26 different letters. This means that the dataset has a lot of points intermingled with one another, and slight differences in some attributes can lead to completely different "spam" and "not spam" classifications. Hence, the clustering algorithms had a harder time with it.

2.2. Expectation Maximization

In this algorithm, the k-means approach is extended with the difference being that instead of trying to maximize the difference in means of the clusters (or decrease the sum of squares), this algorithm computes probability of cluster memberships and assigns the data such that the overall likelihood of the data falling in those clusters is maximized.

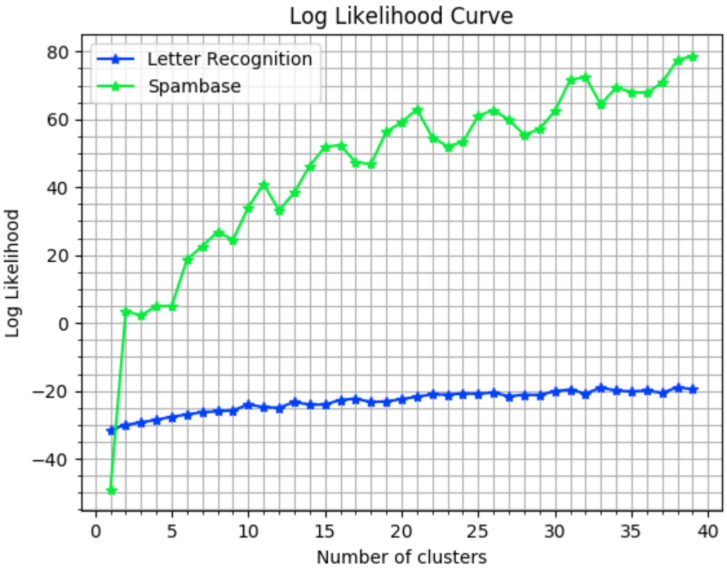


Figure 3: Log Likelihood curve of the two datasets

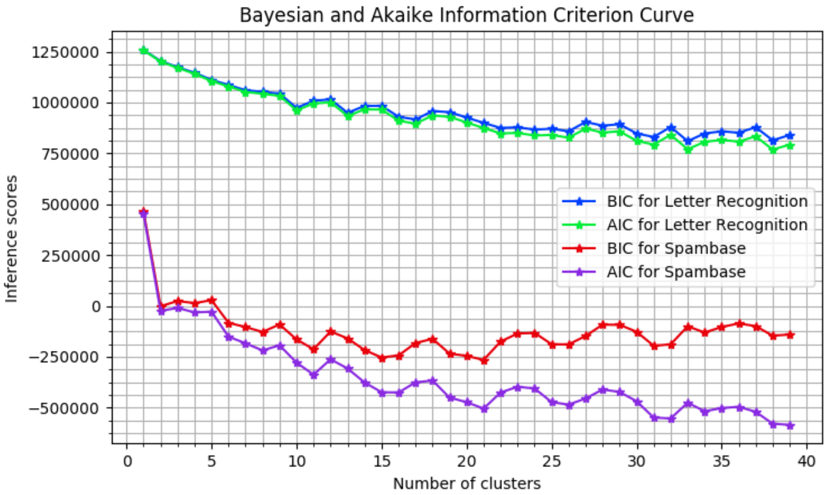


Figure 4: Information criterion curves as the number of clusters varies

The log likelihood curves show expected behavior. They generally increase, with an increase in the number of clusters. That means that the variance of each cluster is decreasing. This is likely due to the fact that as the number of clusters increases, a fewer number of data points are assigned to the cluster and the variance of the data points in relation to the centroid of the cluster decreases.

It is worth noting that the spambase curve is much steeper that the letter recognition one. Wine dataset shows diminishing improvements because small number of clusters can represent the data well and after that the improvements are marginal. It is important to note that Log Likelihood for spambase is much closer to 0 than Letter Recognition dataset. This means that expectation maximization performs better on the spambase dataset than on the Letter dataset for up to 40 clusters.

Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) are scores that measure the overfitting of the model, and punish increasing model complexity by lowering the score. Thus, the model with the highest BIC/AIC should be preferred. Both datasets show expected behavior where the BIC/AIC score reduces as the number of clusters increase. For letter recognition dataset, there isn’t much difference in the scores after 5 clusters, however according to Ockham’s razer model with 5 clusters should be preferred. But for the spam base dataset a sharp decline in Information Criterion score is seen as the number of clusters are increased.

3. Feature Reduction

The four algorithms that were used to evaluate feature reduction are: principal component analysis (PCA), independent component analysis (ICA), randomized projections (RP) and variance threshold reduction (VTR).

To test PCA, ICA and RP the following was done: Letter Recognition dataset was divided into 26 clusters, then each cluster was given the dominant label as its value. This was done because that cluster is probably representative of that label and would help provide some measure of accuracy to see if the same points when dimensionally reduced through various reduction algorithms would fall within the same clusters. The error is the sum of data points in a cluster that did not have the same label divided by the total number of data points. A similar test was then run for the spambase dataset, where the dataset was divided into 2 clusters and each cluster was given the label corresponding to the maximum data points.

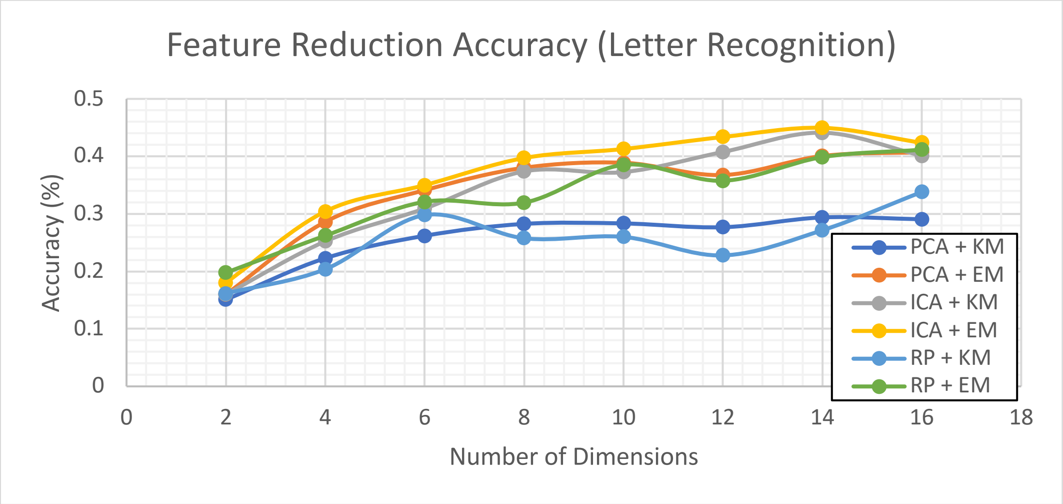


Figure 5.

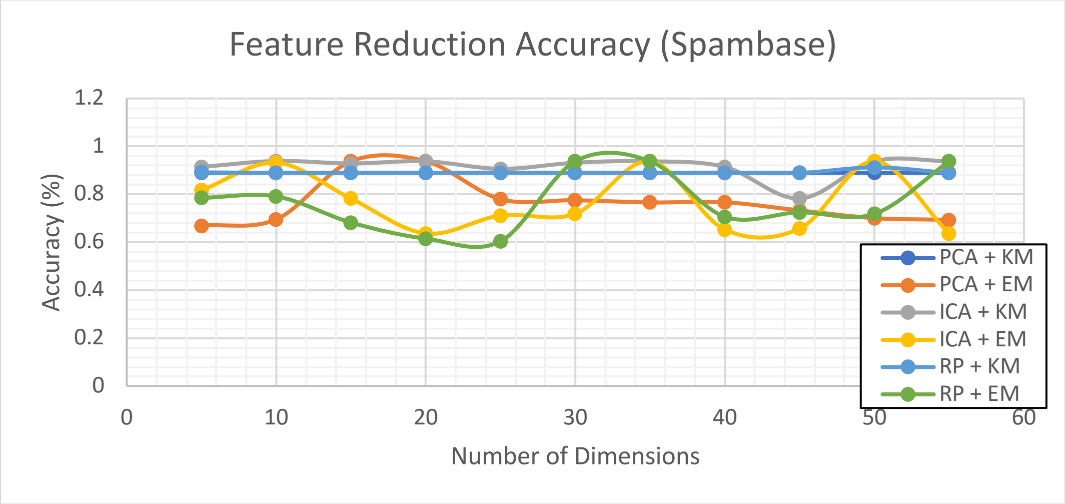
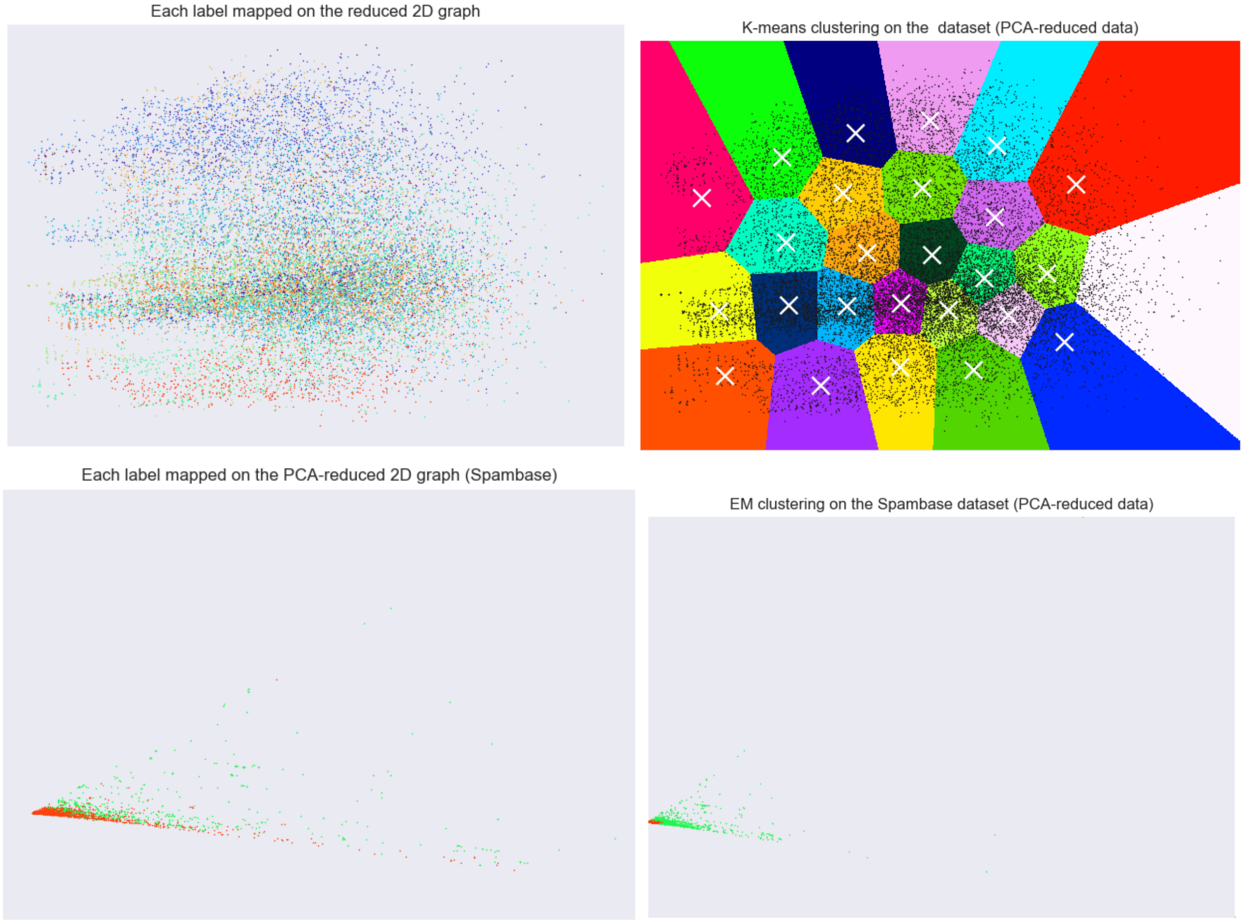


Figure 6.

For the letter recognition dataset, it can be clearly seen that Expectation Maximization outperforms K-Means algorithm in all the tests. This is likely because K-Means strongly assigns data pointers to clusters which is not suitable for either dataset. EM weakly assigns each data point to a cluster which it turns out leads to higher accuracy, i.e. more points with same labels land in the same cluster. However, this trend is not similarly seen in the spambase dataset. For this dataset, they seem to be close enough to be indistinguishable. From closer inspection, it was discovered, that there were a few data points that were consistently being misclassified. This is due to the fact that the dataset has a large number of attributes and not as many data points. So the effect of some attributes cannot be seen from the data provided.

Another interesting trend is revealed by these graphs, for letter recogntion dataset there is significant loss of information as the total number of attributes are decreased the accuracy falls drastically. This is likely because each attribute in Letter dataset gives information about the physical features of a letter, thus intuitively making each feature important. However, for the spambase dataset we see improvement in accuracy as the number of attributes are reduced to 5, hinting that not all attributes provide meaningful information that is required to cluster similar labels together. This also reinforces the previous idea, that due to the large number of attributes but relatively small number of attributes, the importance of each attribute cannot be discovered in the dataset.

3.1. Principal Component Analysis (PCA) and Independent Component Analysis (ICA)



Each label mapped on the reduced 2D graph

K-means clustering on the dataset (PCA-reduced data)

Each label mapped on the PCA-reduced 2D graph (Spambase)

EM clustering on the Spambase dataset (PCA-reduced data)

Figure 7. Graphs for PCA for both datasets.

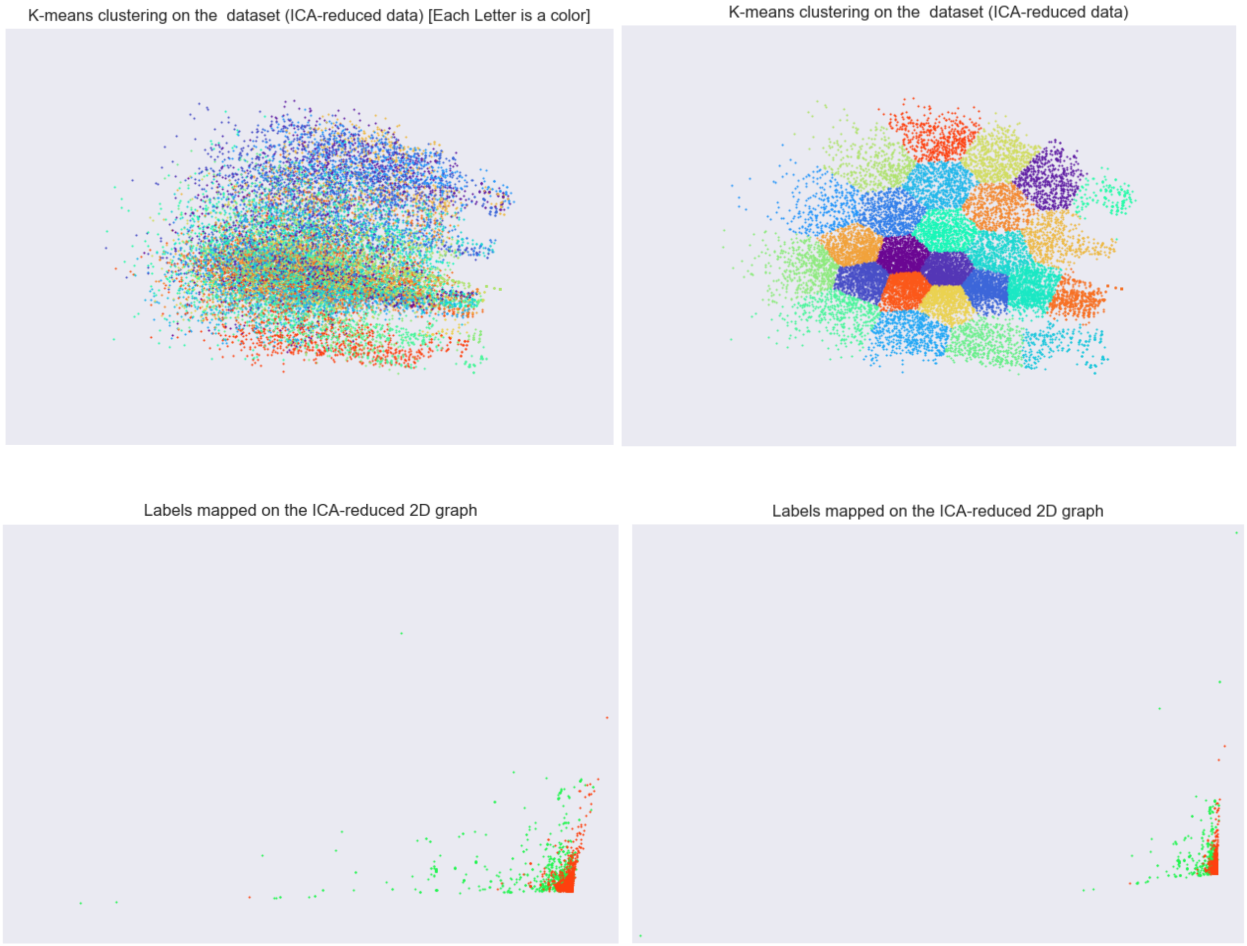


Figure 8. Graphs for ICA for both datasets

For Spambase, the performance of clustering after applying PCA generated eigenvalues that were not very accurate. In fact, they were skewed with 3 linear combinations of attributes in the 2-6 range, while the rest were between 0 and 1. As a result, we can conclude that the dataset shows variance for only a few axes and the rest shows no variance at all. This is explained by the redundant and irrelevant attributes present in Spambase as the classification is skewed to the right. Furthermore, while K-means, after performing PCA on it, performed better on average than without dimensionality reduction, we can see that the dimensionality reduction algorithm did attempt to remove those redundant and irrelevant attributes before clustering the relevant information. However, in contrast, EM performed significantly worse. Because PCA generates attributes on the dataset where most variance occurs, the effect of extraneous attributes will get reduced, therefore making it harder for EM, with the use of probability, to assign correctly each instance to their “correct” cluster.

For both PCA reduced and ICA reduced Letter dataset a clear trend can be seen. The letter classes fall in somewhat horizontal lines, explaining the low accuracy achieved in the previous tests. Euclidean distance is not able to capture the shape of the clusters for each label and to improve the measured accuracy in previous test a new form of distance metric is required that is able to capture the similarity represented by each label.

3.3. Random Projections (RP)

This section refers to figures 5 and 6 above.

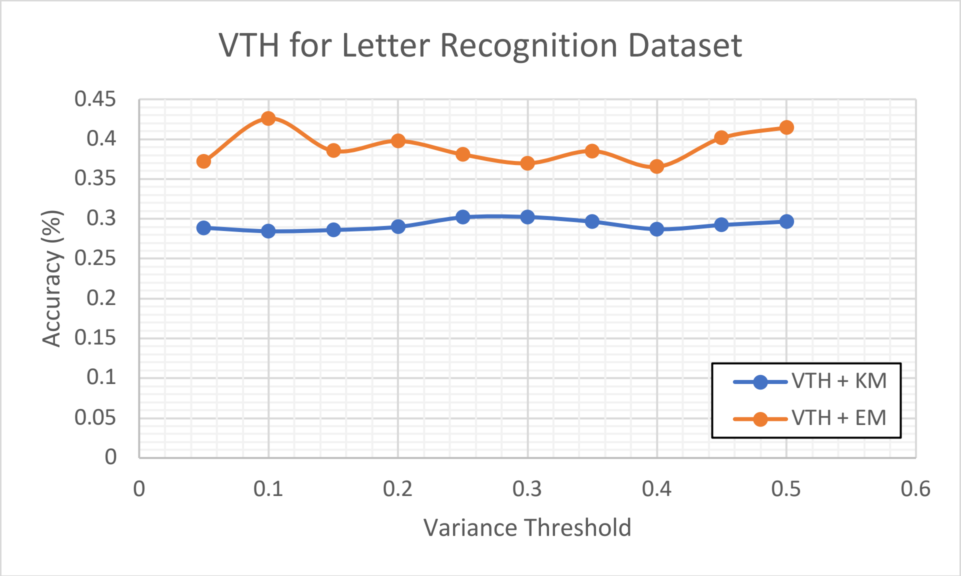
It is interesting to see how much more consistent EM is in Spambase than K-means after performing Random Projections with different seeds, due to how it performs its clustering with probability, and at the same time, makes a better performance than K-means. This EM, in contrast with PCA, is actually producing less error than K-means, but at the same time, worsening the original accuracy without dimensionality reduction in the first place.

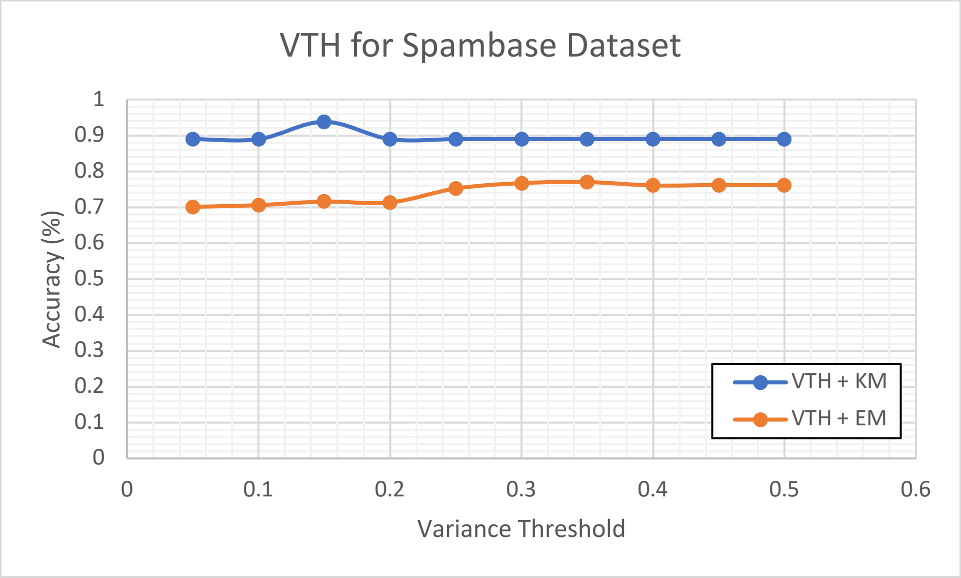
It is a simpler algorithm that trades accuracy for simpler models. Hence, the accuracy from this algorithm was lower than PCA/ICA as can be seen in the figures above.

3.4. Variance Threshold Reduction (VTR)

This algorithm is very similar to RP. Basically, it removes features below a certain threshold of variance. In other words, it removes complexity in the model for attributes that only affect the output marginally. It is similar to RP in that it trades in accuracy for simpler complexity.

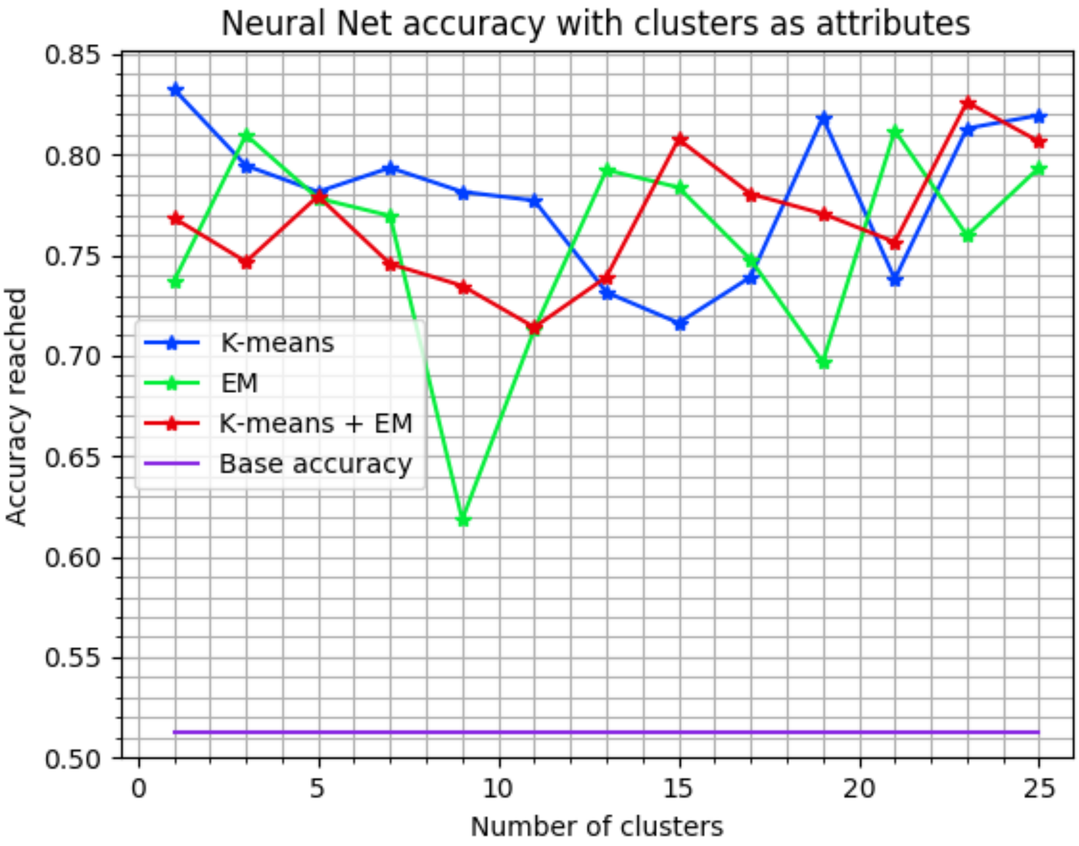
To test this algorithm, it was evaluated for various thresholds of variance. The results are similar to that of RP. What is interesting to note is that the accuracy achieved for the letter recognition dataset is half of that for the spambase dataset. This means that removing any attribute whatsoever in the letter database has a great affect on the accuracy. In other words, most if not all attributes are very important. For the spambase dataset on the other hand, the accuracy is still >80% even with the highest threshold. This means that it possess many attributes that don’t affect the result as greatly (or conversely, that there are a few attributes that are very important, and as long as those are included, the accuracy of the classification will still be high).





4. Neural Net Training

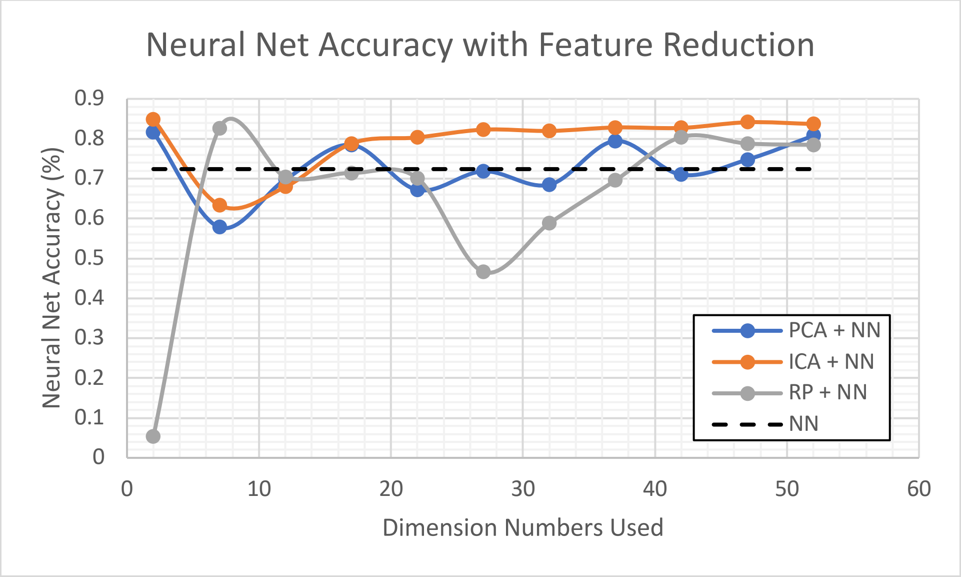
4.1. Using Clustering Algorithms



The spambase dataset was clustered using K-means and Expectation Maximization where the number of clusters was increased from 1 to 25 in increments of 2. Then the training data was given an additional attribute 'kmeans' which listed the cluster number each data point belonged to for K-mean clustering algorithm. Similarly, for EM, a column called 'EM' was added to training data which listed the cluster each data point belonged to. Then as a third test, both columns 'kmeans' and 'em' were added to the training data. Then finally, the neural nets were trained on these new datasets and tested for changes in accuracy to see if the clustering information reveals something new to the neural net.

As seen from the graph, the clustering improves the neural net quite significantly. Also of interesting note however, is that the algorithms seem to lead to very similar results, where the K-means algorithm actually has a slight advantage to the EM algorithm, which would seem a little counter-intuitive. This is counter-intuitive because EM can be seen as almost an improved extension of k-means.

4.2. Using Feature Reduction Algorithms



Next, feature reduction algorithms were run on the spambase dataset before running it through the neural net. This should reduce the effect of “the curse of dimensionality”. The number of features for the spambase dataset was varied from 5 to 55 attributes, and the accuracies are plotted above.

As can be seen, the accuracy generally increases with an increased number of dimensions, but after some point, it either does not make much of a difference, or actually reduced the accuracy. This reveals that there were some additional attributes that made the data more confusing/incorrect and harder to classify rather than revealing any important information about the data.

We can conclude from this however, that generally the feature reduction algorithms were useful in that most of the time, their accuracies were better than just the plain neural net.