CS7641 Assignment 3 Jacob Kilver 9 Nov 2014

I. Introduction

Clustering and feature selection/transformation are explored using two data sets. Expectation Maximization and k-Means clustering are used on the original data. Then, feature selection/transformation algorithms are used to reduce the amount of data. Next, the reduced feature data is used to train a neural network and the results are compared to the results from training with the full data set. Clustering is then performed on the reduced feature data and the results are compared to clustering with the full data set. Finally, the clusters themselves are used as the features to train a neural network with the results being compared to the previous findings.

II. Clustering on original data

A. K Means

For k-means clustering, k=2 was chosen for both data sets because it was known in advance that each data set had only 2 classifications. It was assumed that this would lead to the best clusters that would correspond to the binary classifications of the instances.

1. Adult Income

Figure 1 shows what was observed when k-means clustering was performed on the adult income data. Clusters are on the y-axis and classes are on the x-axis. Each of the two clusters had instances from both classes. This is not a good split, and indicates that k-means had difficulty finding good clusters for this data.

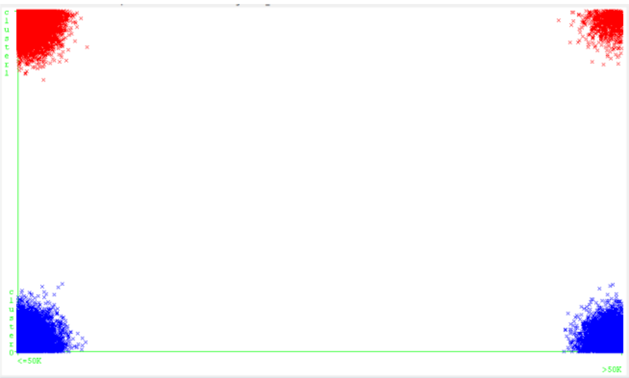

Figure 1: Adult Income Clustering

2. Bank Note Authentication

The k-means clustering algorithm with k = 2 was able to correctly cluster the two groups into authentic and inauthentic notes. One cluster had all the authentic instances and one cluster had all the inauthentic instances. This is in contrast to the adult income data set where each cluster had instances from both classes.

B. Expectation Maximization

Two runs were performed for each data set using expectation maximization. The first had no limit on the number of clusters it could find while the second was limited to only two.

1. Adult Income

EM was able to correctly pick two clusters when it was allowed to select the number of clusters. However, the clusters looked similar to Figure 1 above, where instances from both classes were in each cluster. This was unexpected. Since EM was able to pick 2 clusters, it was expected that the two clusters would fall primarily along the class division. Since this was not the case, it appeared that the data was not very consistent – similar data points according to EM (instances in the same cluster) did not necessarily belong to the same class.

When the number of clusters was restricted to two, the results were nearly identical to when EM was allowed to pick the number of clusters. This is reasonable, since EM first picked two clusters.

2. Bank Note Authentication

When EM was run without limiting the number of clusters, it resulted in 18 different clusters being generated even though the instances had only four features. Interestingly, the clusters still did a good job dividing the data according to the classification of the instances. Each cluster had instances from one of the two classifications, though each also had one instance from the other class. In other words, all but one instance in each cluster were of the same classification. This was a much better division of the data along class divisions than with the adult income data set.

Second, when EM was run with the number of clusters limited to 2, the behavior was similar. All the instances in each cluster were of the same classification except for one.

III. Dimensionality Reduction

A. Principle Components Analysis

1. Adult Income

PCA was performed twice on the adult income data set. The first time it was allowed to pick as many principle components as it saw fit. In this case, it selected 86 features that were linear combinations of the original features. At first this was not expected, but upon closer examination it was reasonable. The algorithm converted the nominal features (e.g. relationship status) to binary (e.g., relationship=married). When all the features were expanded as such, there are more than 100 distinct features. So, PCA was able to reduce the number of features.

The second time PCA was run, the number of features were limited to 10. The results were same as when PCA was able to select as many attributes as it liked with the difference being only the ten most important attributes (attributes with the largest eigenvalues) were retained. The value of 10 was selected because it was a significant reduction of the 86 features PCA selected on its own, and after the first 10 features, the eigenvalues levelled off, so choosing more features would not significantly help with training later.

2. Bank Note Authentication

Applying PCA to the bank note authentication data set resulted in a reduction of features from four to three. Each of the new features was simply a linear combination of the original four. Since PCA resulted in a reduction of features without restriction, it was only performed once.

B. Independent Components Analysis

1. Adult Income

ICA did not converge for the adult income data set. With the large number of features and the large number of instances, this is disappointing, but not unexpected. Finding independent components is not guaranteed with a given data set, as illustrated here.

2. Bank Note Authentication

Applying ICA to the bank note authentication data set resulted in four sources being found. This was as expected. There were four different sources of information, and ICA was able to identify them. The tool used to calculate ICA projections did not output the eigenvectors, so they were unable to be analyzed. However, by viewing the distributions of the original and the transformed data, it was observed that there was some correlation (similar distributions of the data) between the two data sets.

C. Random Components Analysis

1. Adult Income

The algorithm that was used for RCA required the number of dimensions to be input by the user. Since ten features were selected for PCA, RCA was limited to choosing 10 features. Using the same seed value, the distributions were identical, but when the seed value was changed, drastic shifts in the distribution could occur.

The tool used for RCA did not have the capability to reverse the projection directly. However, when random components data was converted back to the same number of features as the original data set, the distributions were remarkably similar suggesting that a large amount of the original data set was preserved.

2. Bank Note Authentication

When applied to the bank note authentication data set, RCA was set to reduce the dimensions from four to three. Distributions could vary significantly between various seed distributions. However, it seems that even though three dimensions were permitted, at most 2 dimensions were used. In other words, for all the distributions that were tested, one contained no information; one of the three remaining features was all zeros.

Projecting the data back to the original feature space did not work with the bank note authentication data set as it did with the adult income data set, most likely because one feature was always all zeros. As such, when trying to reproject back to the original space of four features, only one or two features contained information.

D. Information Gain

The information gain algorithm selected the features that provided the most information gain. As such, it only performed feature selection and no feature transformation.

1. Adult Income

For the adult income data set, information gain algorithm was run over all the features to see which provided the most information gain. Afterwards the top 10 features were selected to carry forward. As can be seen in Figure 2, the Relationship and Marital Status features were the most important, with almost one-third of all information between the two attributes. Following those, the information gain continued to decrease until the gain was actually zero. The Final Weight attribute provided no more information than all the preceding features combined did.

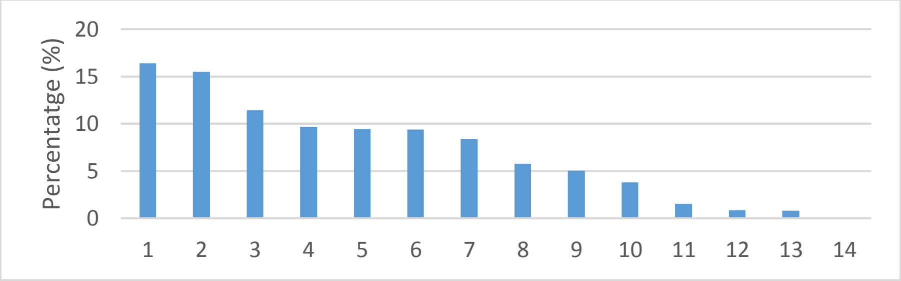

Figure 2: Percentage of information gain for adult income data set

2. Bank Note Authentication

As with the adult income data set, the information gain algorithm was run over all the features to rank them, then the top performers were selected. In this case, since there were only four features, 3 features were carried forward. Figure 3 shows the distribution of information gain percentages for the features. Entropy provided no information gain in conjunction with the other three features. As such, this feature was the one selected for removal.

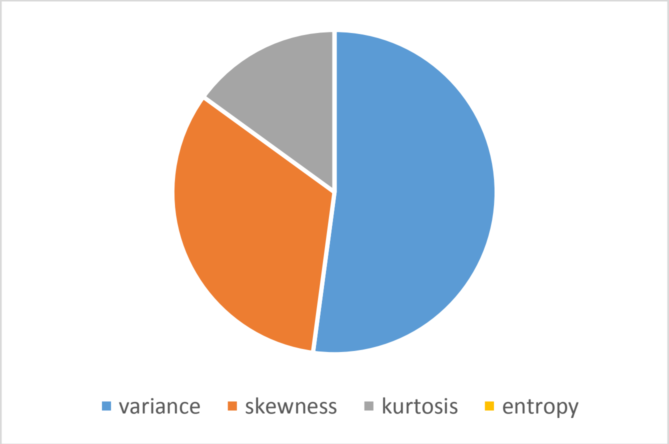

Figure 3: Information gain for bank note authentication data set

IV. Neural Network with reduced dimension data

Note: Testing error for all the reduced dimension data sets except those generated by information gain were not able to be tested using the testing data. The tool used was unable to correctly correlate the features between the reduced data set and the testing data set.

A. Original Data

For comparison, the results using the original full data set are presented in the tables below.

Table 1: Full Adult Income Data Set Neural Network Training results

Training Error	16.27%
Training Time (s)	217.52
Testing Error	16.25%

Table 2: Full Bank Note Authentication Data Set Neural Network Training Results

Training Error	.4049%
Training Time (s)	2.17
Testing Error	1.5%

B. Principle Components Analysis

1. Adult Income

Neural network training was performed once for each data set produced via PCA feature transformation. First, the data set in which the number of features was selected by PCA was used. The results can be found in Table 3. Surprisingly, the training error actually decreased. Since the amount of data was reduced, the training time decreased as expected.

Next, the data set where PCA was permitted only 10 features was used, with the results in Table 4. The training error was virtually unchanged as compared to the original data set even though there was significantly less data in this data set. As expected, the training time improved significantly.

Table 3: Neural Network Training Results with PCA reduced data - unlimited features

Training Error	15.0282%
Training Time (s)	184.78

Table 4: Neural Network Training Results with PCA reduced data - 10 features

Training Error	16.533%
Training Time (s)	36.19

2. Bank Note Authentication

The results from neural network training with the PCA transformed data are presented in Table 5. The training error increased dramatically as compared to the full data set while the training time was cut approximately in half.

Table 5: Neural Network Training Results with PCA transformed data

Training Error	5.668%
Training Time (s)	1

C. Independent Components Analysis

1. Adult Income

Since ICA did not converge for this data set, the neural network using ICA reduced data could not be trained.

2. Bank Note Authentication

The results from neural network training with ICA transformed data are presented in Table 6. The training error doubled but is still insignificant. Meanwhile, the training time was reduced by more than half as compared to the full data set.

Table 6: Neural Network Training Results with ICA transformed data

Training Error	0.8097%
Training Time (s)	1.02

D. Random Components Analysis

1. Adult Income

The results from neural network training with RCA transformed data are presented in Table 7. The training error did not increase significantly while the training time was reduced dramatically.

Table 7: Averaged Neural Network Training Results with RCA transformed data

Training Error	18.8455%
Training Time (s)	39.31

2. Bank Note Authentication

The results from neural network training with RCA transformed data are presented in Table 8. The training error increased significantly while the training time was reduced by half.

Table 8: Averaged Neural Network Training Results with RCA transformed data

Training Error	26.99056667%
Training Time (s)	0.97

E. Information Gain

Note: Testing error is included in the tables below because the tool used was able to correctly map the features of the training set to the features of the testing set.

1. Adult Income

The results from neural network training with Information Gain reduced data are presented in Table 9. As compared to the full data results, the training error was actually less and the testing error was approximately the same. Furthermore, the training time was significantly reduced. Feature reduction was able to reduce the amount of data needed for this data set while leaving the results unchanged or improved.

Table 9: Neural Network Training Results with Information Gain reduced data

Training Error	15.1646%
Training Time (s)	115.31
Testing Error	16.4312%

2. Bank Note Authentication

The results from neural network training with Information Gain reduced data are presented in Table 10. Training and testing error increased, but not significantly while training time was reduced by more than half. As with the adult income data set, information gain feature reduction was able to reduce the amount of data required for classification while not significantly affecting accuracy.

Table 10: Neural Network Training Results with Information Gain reduced data

Training Error	0.5668%
Training Time (s)	1
Testing Error	2.1898%

V. Clustering reduced/transformed data

A. K Means

As before, the k-means algorithm was restricted to having 2 clusters.

1. Adult Income

a) PCA

The k-means algorithm was run for both PCA data sets. The first one, where PCA was allowed to select as many features as it saw fit, resulted in clusters as with the original full data set. The clusters did not fall along class divisions. Rather, there were instances from each class in both clusters. However, cluster 1 appeared much smaller with this data than with the original data.

The results with the data set with only 10 features were virtually identical to those with the full PCA data set.

b) ICA

ICA did not converge for the adult income data set, so this test could not be performed.

c) RCA

K-means clustering was performed on each RCA transformed data set individually. The results were mostly similar, though at times the cluster assignments were swapped. The clusters were similar to the original in that there were instances from each class in each cluster. However, the size of the clusters differed from the clusters of the original data.

d) Information Gain

The clusters generated using information gain were virtually identical to those generated from the original data. This is not surprising since information gain only removed features without performing any transformation. However, it is interesting that even with less data, the same clusters could be generated.

2. Bank Note Authentication

a) PCA

The clusters using the PCA reduced data were virtually identical to those using the original data set. Although the centroids were in different locations, the clusters still contained instances from only a single class. Even though there were fewer features, each feature was a linear combination of the original features. Thus, the same information was available to both, but PCA presented it in a more compact manner.

b) ICA

The clusters using the ICA reduced data were nearly identical to those using the original data set. Although the centroids were in different locations, the clusters still contained instances from only a single class.

c) RCA

The clusters using the RCA reduced data were nearly identical to those using the original data set. Although the centroids were in different locations, the clusters still contained instances from only a single class.

d) Information Gain

The clusters using the Information Gain reduced data were nearly identical to those using the original data set. The only difference was that the reduced data had one less dimension than the original data. Otherwise the centroids were in the exact same location. As such, each cluster contained instances from only a single class.

B. Expectation Maximization

The EM algorithm for each data set was run twice. The first time it selected the number of clusters on its own via cross validation. The second time the number of clusters was fixed at two.

1. Adult Income

a) PCA

The EM algorithm selected 11 clusters when permitted to select the number of clusters on its own. However, each cluster contained instances from both classes. The results were the same when the number of clusters was restricted to two: both clusters had instances from both classes.

With the PCA data with only 10 features EM selected 20 clusters. While several clusters had instances from only a single class, most had instances from both classes. When EM was limited to two clusters the results were similar to those generated using the original data sets; each cluster had instances from each class.

b) ICA

There was no ICA data to work with since ICA with the original data did not converge.

c) RCA

When EM was allowed to choose the number of clusters on its own, it selected between 10 and 21 clusters depending on the data set. However, the clusters did not fall along class divisions; each cluster had instances from both classes. The results were the same when the number of clusters was restricted to two.

d) Information Gain

The EM algorithm output the same results both times it was run. Both clusters had instances from each class.

2. Bank Note Authentication

a) PCA

With the number of clusters selected by PCA, 18 clusters were generated. This was the same result as with the original data, so it appeared the information was preserved. The clusters also generally fell along class divisions, although this time there appeared to be a handful more cases where one instances of each class belonged to the same cluster.

The results when the number of clusters was restricted to two were identical to that of the original data. All but one instance in each cluster belonged to the same class. This was further evidence that the information content was the same.

b) ICA

With the number of clusters unbounded, EM selected 22 clusters. Once again, the clusters generally fell along class divisions, but there seemed to be more cases where one cluster had instances from both classes than in the PCA case.

With the number of clusters restricted to 2, the results were different from those with the original data. Each cluster contained many instances from each class. It appeared that some of the data was lost projecting the data to the independent components.

c) RCA

The results using EM on the RCA data were not consistent. Depending on the RCA data, the number of clusters was anywhere from 12-25. Furthermore, the division of these clusters along class divisions was not clear. Many clusters had instances from both clusters.

With the number of clusters restricted to two, the results were still not consistent. For some RCA data sets, the clusters were exactly as with the original data – all but one instance in each cluster was from the same class. However, with other RCA data sets, the clusters were more like ICA where each cluster contained many instances from both classes.

d) Information Gain

When EM was allowed to select the number of clusters for the information gain reduced data, it selected 28 clusters. This is more clusters than the original data produced, but the information gain data had fewer features. Furthermore, the clusters fell exactly along class divisions, just as the original data did.

When EM was restricted to two clusters, the results were exactly the same as those with the original data and the PCA data. The clusters fell along class divisions with the exception of one instance of the opposite class in each cluster.

VI. Neural Network training with clusters membership

A. K Means

1. Adult Income

a) PCA

The results were for all purposes the same between the two PCA data sets with kMeans clustering. Both put all the instances into a single class. The only advantage was that the training time was approximately 22 seconds as opposed to nearly 200 seconds on the original data.

b) ICA

ICA did not converge for the adult income data set.

c) RCA

All the RCA data with k-means clustering behaved similarly. All the instances were placed in a single class within about 17 seconds of training.

d) Information Gain

The information gain data set with k-means clustering behaved like the data sets above. All the instances were placed in a single class, only this time it took a few seconds longer, 24.5 seconds.

2. Bank Note Authentication

a) PCA

Using the clusters generated by PCA did not affect the output as much as some of the other reduced data did. Training error increased significantly as compared to the original data set to 10.6%, but was significantly less than the other data sets using cluster membership. Meanwhile, training time was reduced by half.

b) ICA

The ICA data set performed the worst. The neural network put every instance into a single class, increasing the training error to 44.7%. This did not warrant the relatively insignificant training time reduction to 0.8 seconds.

c) RCA

The various data sets generated from RCA on average performed poorly. Most of them placed all the instances in a single class while offering an insignificant reduction in computation time. See Table 11.

Table 11: Averaged RCA data sets with k-means clustering neural network training performances

Training Error	42%
Training Time (s)	.867

d) Information Gain

It was expected that the information gain data set would have performed similarly to the PCA data set, but this was not the case. The training error increased significantly (38.3%) while the training time was similar to the other data sets (0.84 seconds).

B. Expectation Maximization

Since EM was run twice for each data set, neural network training was also performed twice.

1. Adult Income

a) PCA

The training error for the PCA data set where PCA selected the number of features was the same for both EM clustering data sets. All the instances were placed in a single class. The difference was in the training time; the set where the number of clusters was set to two took half the time to train as the set where EM selected the number of clusters (20 and 40 seconds, respectively).

The data set where PCA only selected 10 features did have different results. With EM able to select the number of clusters, neural network training did not place everything in a single class. Training error was 19%, which was not significantly worse than with the original data set. This was highly unexpected, especially since EM clustering did have clusters containing instances of only a single. However, training took just over a minute, longer than with any other PCA data set but still less than one-third of the amount of time as the original data set. With the number of clusters restricted to two, the training results were virtually identical to the PCA data set above with the number of features selected by PCA. All the instances were placed in a single class.

b) ICA

ICA did not converge for the adult income data set.

c) RCA

The RCA data sets with EM clusters did better than expected. Unlike some of the other data sets, not every neural network classified all the instances the same. Nevertheless, the average performance was not all that better, (22.6% and 23.25% training error for the EM selected clusters and the 2 cluster limited data sets, respectively). The training times were typical of the other data sets as well (30 and 18 seconds, respectively).

d) Information Gain

Since EM chose two clusters for the information gain data set, neural network training was only performed once. While it did not put everything in a single class, it actually performed worse than if it had (29% vs 24%). The training time was about average (20 seconds).

2. Bank Note Authentication

a) PCA

When EM was allowed to choose the number of clusters, the results did not offer any advantage over the original data. The training error increased several fold (6.5%) while the training time remained about the same (2.53 seconds).

When EM was restricted to 2 clusters, neural network training was not nearly as successful. Although the training time was significantly reduced (0.81 seconds), the training error increased significantly (37.6%).

b) ICA

Neural network training with the ICA data using the number of clusters selected by EM was not very successful. The training time remained virtually unchanged as compared to the original data set (2.6 seconds) while the training error increased significantly (38.8%).

The only advantage of restricting the number of clusters to two was a significant reduction in training time (0.75). However, the cost was even greater training error (44.7%).

c) RCA

Using the clusters selected by EM on the RCA data resulted in moderate neural network performance. The training error did not increase as much as with some other data sets (22.45%) while the training time had a slight decrease (2 seconds).

However, when the number of clusters was restricted to two, the performance was similar to the other data sets (training error: 39.7%, training time: 0.84%).

d) Information Gain

The information gain data set with the number of clusters selected by EM resulted in the smallest increase in training error (2.35%). However, the training time actually increased rather than decreased (2.83 seconds). This was not expected because the amount of data was smaller than with the original data set, and even more so than with the actual information gain reduced data set.

Once again, when the number of clusters was restricted to two, the training error increased significantly (39.7%) while training time decreased as a much slower rate (0.75 seconds).

VII. Conclusions and Future Work

Clustering behaved quite differently for the two data sets, both with the original data and the reduced data. With the original data sets, clustering never fell along classification division for the adult income data set as it did for the bank note data set. None of the clustering algorithms with any of the adult income data sets were able to cluster the data according to classification while nearly all the clustering algorithms did so with the bank note data set. This was most likely the fault of the adult income data set more than the algorithms themselves. In previous analysis, it was thought that this data set might not be very consistent. In other words, data points that were "close" according to the definition used for distance were not of the same classification. Meanwhile, the distance metric for the bank note data set seemed to work surprisingly well. Thus, the domain knowledge of the adult income data set was lacking. A future avenue of research should be exploring how the various distance metrics affect clustering behavior.

Dimensionality reduction algorithms can sometimes improve neural network performance, but most of the experiments conducted showed that there was at least some tradeoff between speed and accuracy when performing dimensionality reduction. However, typically this tradeoff is not very significant. For instance, with the adult income data set, the training error was never more than 3 percentage points worse than the original data set. Meanwhile, even the slowest neural network training time was still 30 seconds faster than with the original data, and most of the other training times were significantly smaller. The one troublesome case with the adult income data set was that of dimensionality reduction using ICA. It did not converge even though it was allowed to run for more than a day with 20 000 iterations. As such, it was determined that convergence within a reasonable amount of time was not possible for this data set using ICA, especially since independent components are not guaranteed for any given data set.

The neural network training results across the reduced bank note data sets are similar to those from the adult income data set. Training time always decreased as expected since there was less data to process. However, the change in the training error was much more erratic. RCA had the worst training error probably because choosing random components with this very consistent data set did not show good domain knowledge. PCA had a surprising increase in training error, especially since there was almost no change with the training error on the adult data set. The data reduced by information gain had the best performance. It had the least increase in training error with a training time no worse than the other data sets.

Clustering the reduced data had some interesting results. For the adult income data, the results using k-means did not differ greatly from the original clustering. As expected, the reduced data was not able to cluster according to the class divisions. This showed a lack of domain knowledge, specifically with the distance metric used to cluster similar instances. In contrast, the reduced bank note data was always able to cluster along the class divisions using k-means. This was surprising since some of the data sets had 25% less data in an already narrow data set.

Clustering with EM with the reduced data was in most respects the same as compared to the original data. While EM selected many clusters for the adult income data set, very few of these clusters corresponded to a particular class. Most of the clusters had instances from both classes. In contrast, EM with the reduced bank note data almost always had clusters that only contained instances from a single class. Once more, this showed a lack of domain knowledge for the adult income data set. Nevertheless, this showed that given the right data set and domain knowledge, dimensionality reduction truly can achieve results similar to the original data with fewer features.

Using the k-means generated clusters as the features for neural network training did not produce acceptable results. For the adult income data set, neural network training placed all instances in a single class. While the training time was greatly reduced, these neural networks were unusable. It was expected that the results would be better with the bank note data set since the clusters corresponded to classes almost exactly and that results would be similar for all data sets. This was not the case, and the reasons are unknown. Since only cluster membership was considered for the input features and since cluster membership was approximately the same for all the data sets, it was expected that all the neural networks would be trained similarly and thus have similar results.

When EM clustering was restricted to two clusters, the neural network training results for the adult income data sets were similar to those using k-means. On average the neural networks performed no better than placing all the instances in a single class while training time was reduced significantly. Since classification was so poor, these neural network were not helpful. When EM was allowed to select the number of clusters, neural network training sometimes did surprisingly well, specifically with the PCA reduced data limited to 10 features, especially since the clustering did not fall along class divisions.

Neural network training using the EM clusters with the bank note data sets had similar results to the adult income data sets. When EM was restricted to two clusters, training error was never less than 37%, which is much worse than the less than 0.5% with the original data. Furthermore, the training time does not significantly decrease. Therefore, this does not represent a tradeoff so much as a general decline in performance. When EM was allowed to select as many clusters as it saw fit, performance improved for some reduced data sets, specifically PCA and Information Gain. Given the above results, it can be concluded that while cluster membership can be used as a feature reduction algorithm, in general its performance is not acceptable.

There are a number further research areas. One important avenue that was not explored was how different distance measurements for clustering affected the clustering results. The value of k for k-means clustering was assumed to match the number of classifications, but as EM showed, this is not always the case. The value of k should be varied. There was not sufficient time to explore how changing the size of the neural network affected training performance, but would be an interesting topic to explore further. Lastly, determining a way to test the trained neural networks using the testing set would be an important step to study next.