Joseph Cantrell

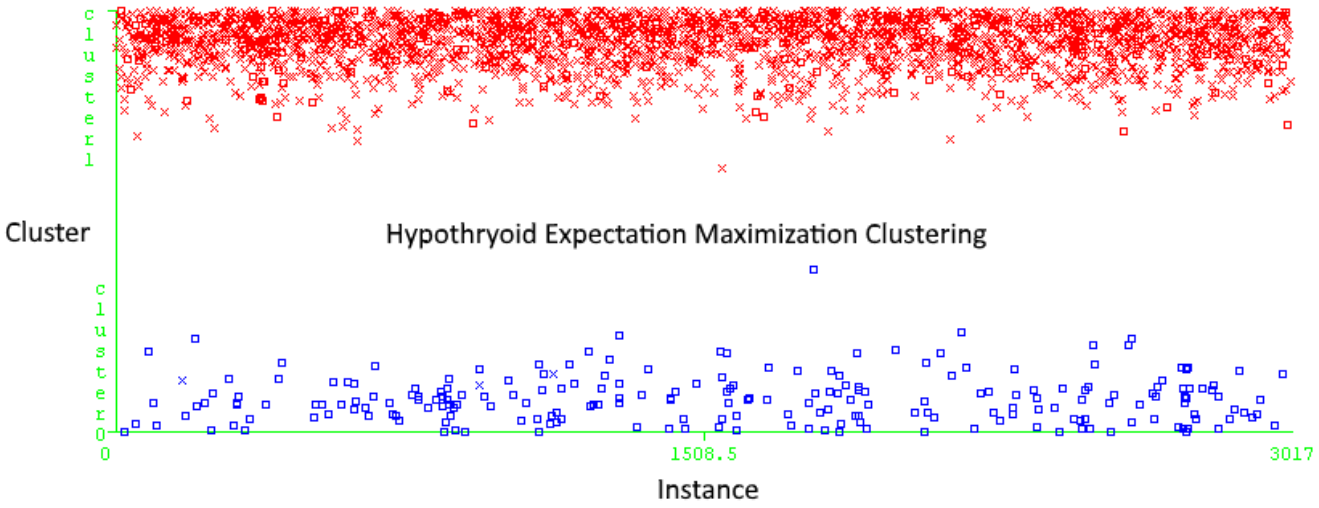
CS 4641

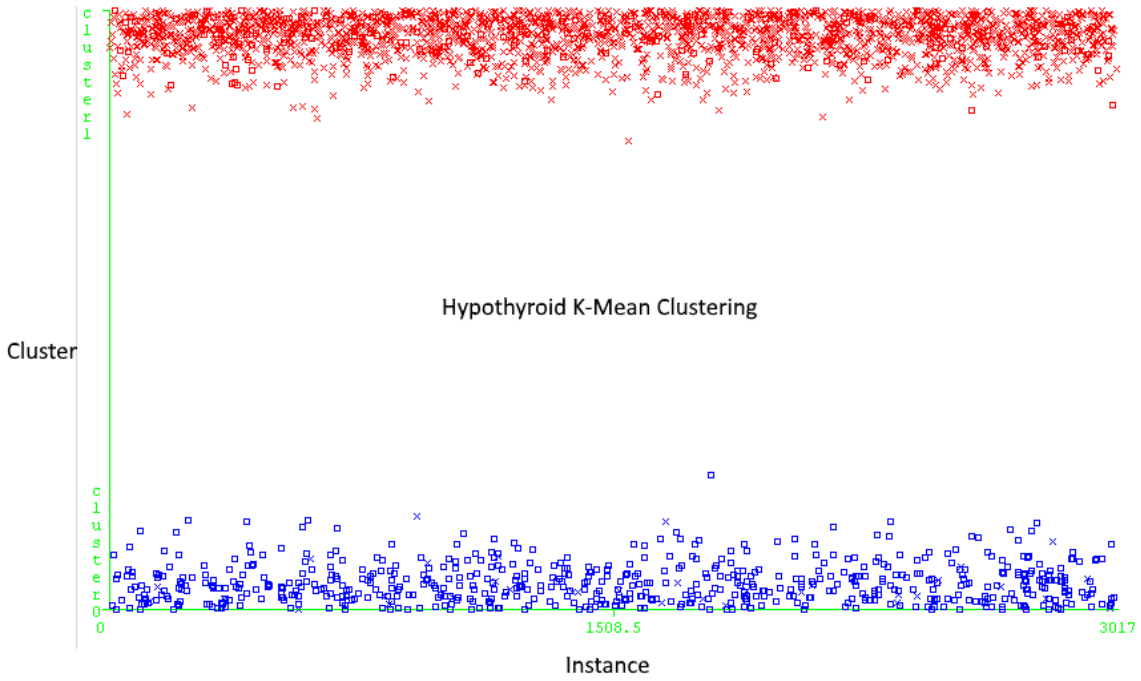
# Unsupervised Learning Algorithms

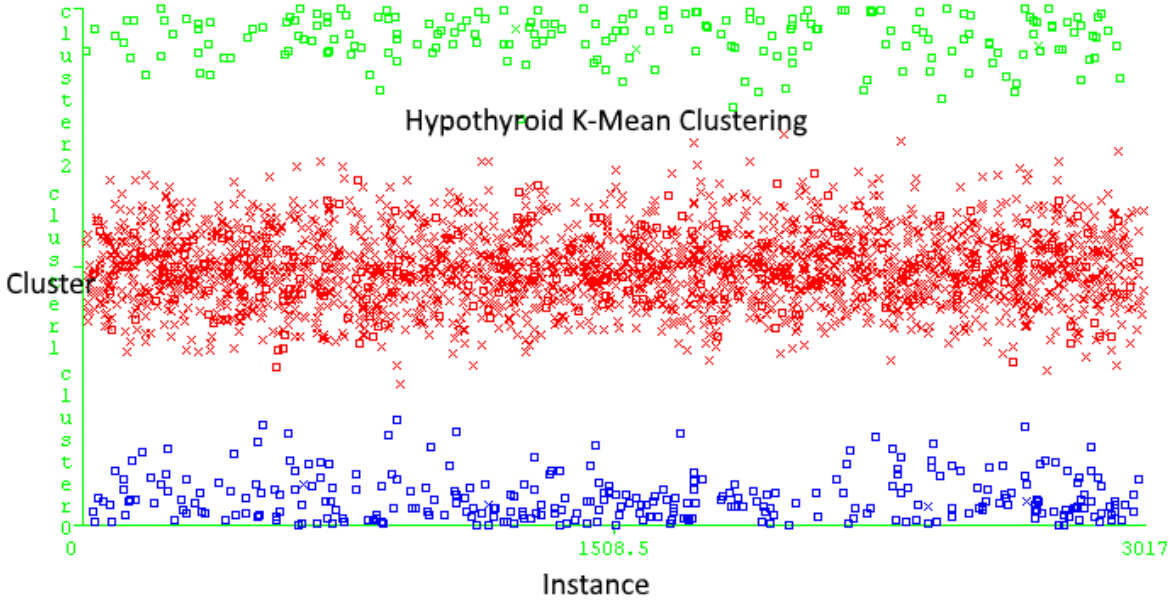
This experiment tests the performance of two clustering algorithms and four dimensionality reduction algorithms. Clustering algorithms divide the dataset's instances into similar categories. Dimensionality reduction algorithms reduce the number of attributes in the dataset. The two clustering algorithms tested are k-means and expectation maximization. The four dimensionality reduction algorithms are the principle components analysis (PCA), the independent components analysis (ICA), randomized projections (RP), and support vector machine (SVM). These algorithms are applied to two different datasets. The first is a German credit rating dataset with 20 classifying attributes and 1,000 records. It has two labels: good and bad. The second dataset is a hypothyroid disease data set from Australia with 3,772 records. It has thirty classifying attributes and four classes: negative, compensated, secondary, and primary.

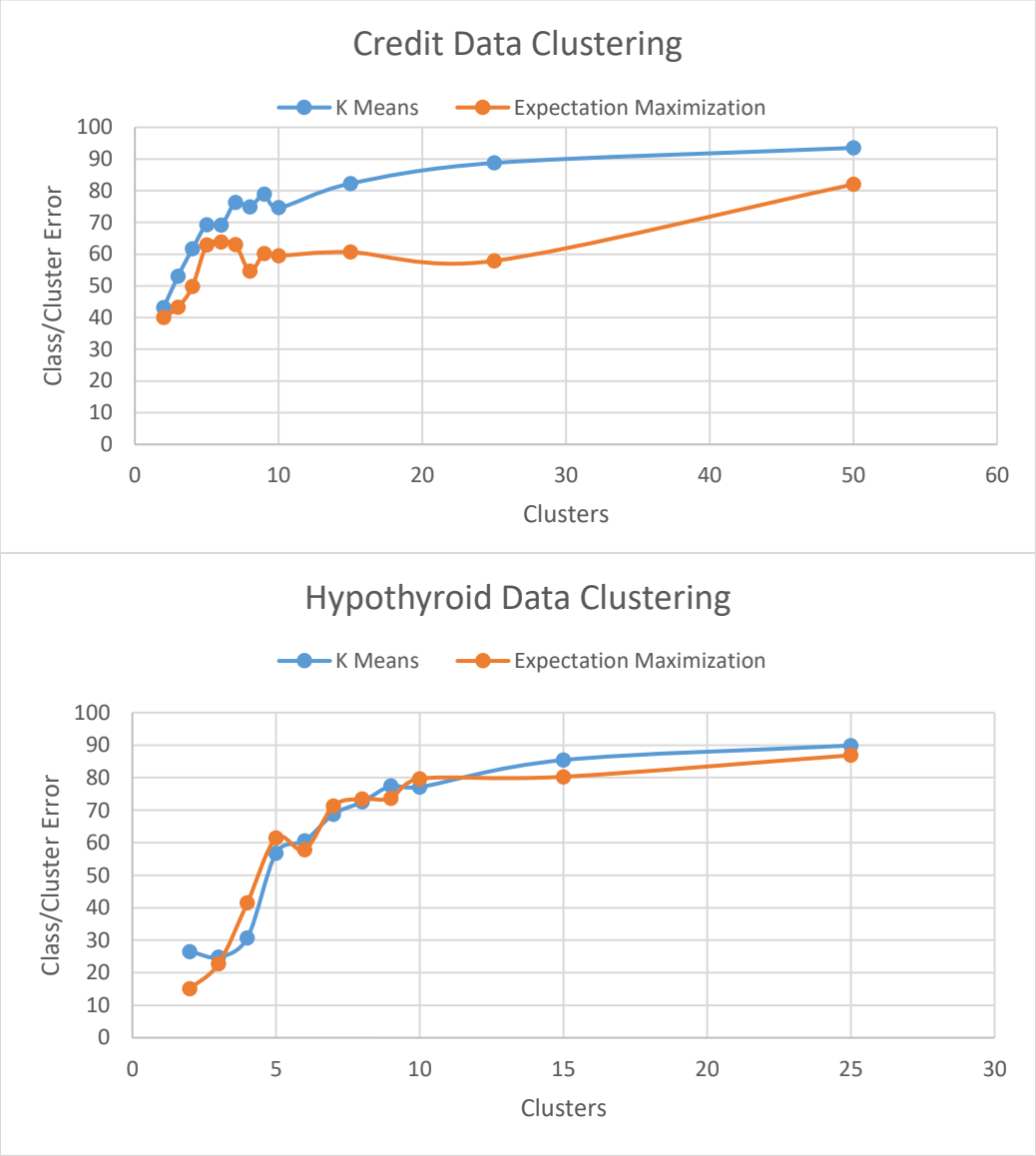
First, the two clustering algorithms were applied to the two datasets. These algorithms arrange the data instances to group the same classes within the same cluster. The class/cluster error is the percent of class labels that did not match up with intended cluster label. The plots below show how the hypothyroid data was divided into clusters. Both algorithms put a large majority of the instances into the red cluster, making this cluster a good indicator for the negative label given this label applies to 92% of the data. This trend continues when the number of clusters is raised to three as well. K, the number of clusters, was tested with several different values. The results of these tests are in the charts below.

For the credit dataset, the k-means algorithm performed worse than expectation maximization in all cases. The algorithms begin with very similar errors but k-means has a much more rapidly growing error than expectation maximization. Both algorithms have an optimal k of two clusters; this is intuitive given the clusters are aligned to the two data labels. The error curves for the hypothyroid dataset are much closer together. The algorithms are very consistent with each other except for the very beginning of the curve. The expectation maximization algorithm outperforms k-means here with an optimal k of two and fifteen percent error. The k-means algorithm has an optimal k of three and an error over fifty percent greater than that of the expectation maximization. Again, the optimal number of clusters for this data set is very close to the number of labels. Keeping the number of clusters close to the number of labels improves the ability of the clustering algorithms to group the labels.




Hypothyroid K-Mean Clustering
Cluster
cluster1
cluster0
0
1508.5
3017
Instance


Hypothyroid K-Mean Clustering
Cluster
cluster2
cluster1
cluster0
0
1508.5
3017
Instance

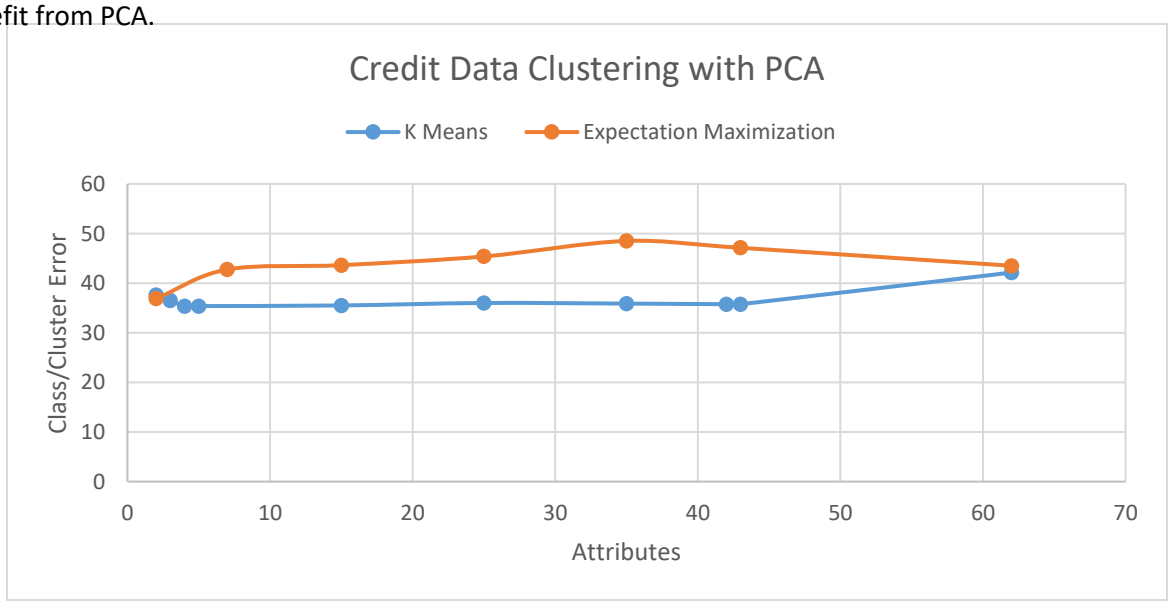


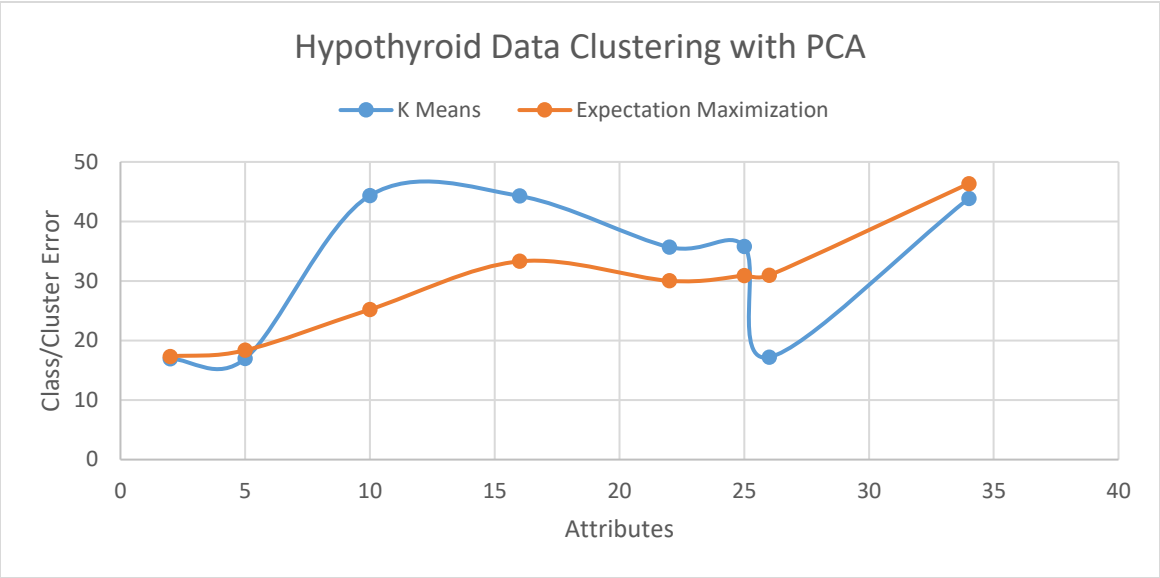

The four dimensionality reduction algorithms were applied to the dataset. These algorithms rank attributes to discard unimportant ones and create an attribute space gleaning the most information using the fewest attributes. These algorithms can drastically decrease machine learning algorithm computation time by reducing the amount of attributes they must process. The first algorithm analyzed was the principle components analysis. The resulting eigenvalues of the transformed hypothyroid dataset are below. The algorithm reduced the dataset from 34 attributes to 25. These eigenvalues are representative of the information the

attribute carries. An eigenvalue greater than one means an attribute has great weight in defining the dataset. The smaller eigenvalues mean these attributes do not capture much variance in the data and have less utility in classification.

| Eigenvalues | | | |
|---|---|---|---|
| 3.71661 | 1.23477 | 0.96244 | 0.78607 |
| 2.82754 | 1.15557 | 0.92957 | 0.78238 |
| 1.91545 | 1.07245 | 0.91274 | 0.7494 |
| 1.80226 | 1.04227 | 0.88998 | 0.72715 |
| 1.66252 | 1.02618 | 0.86451 | 0.64594 |
| 1.26325 | 1.01144 | 0.84057 | 0.6076 |
| | | | 0.45914 |

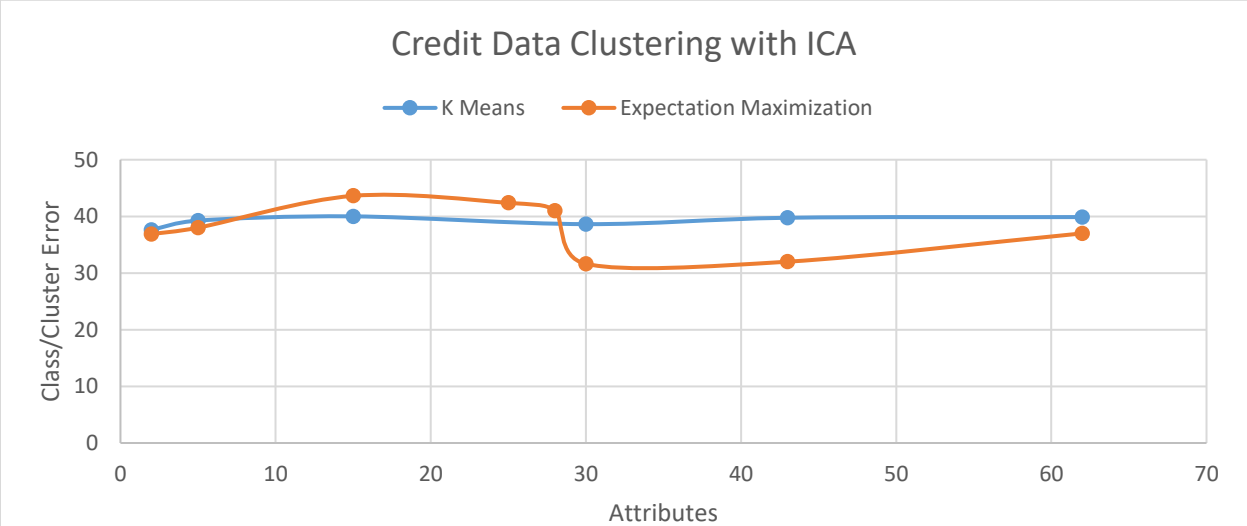
The number of attributes was varied to affect the class/cluster error. PCA tends to decrease the clustering error of the datasets as the number of attributes is decreased more and more. At two attributes (including the label), the clustering algorithms perform best. The credit data is not very responsive to PCA. However, the hypothyroid data shows a massive decline in error by dimensionality reduction. K-means receives a very large performance boost here with PCA; its lowest error dropped from 26% to 17%, a 35% decrease in error. Expectation maximization suffered a performance loss due to PCA; without PCA, the algorithm managed an error rate of 15% instead of 17%. Both datasets can be better clustered by applying PCA first, but k-means and expectation maximization do not always both benefit from PCA.

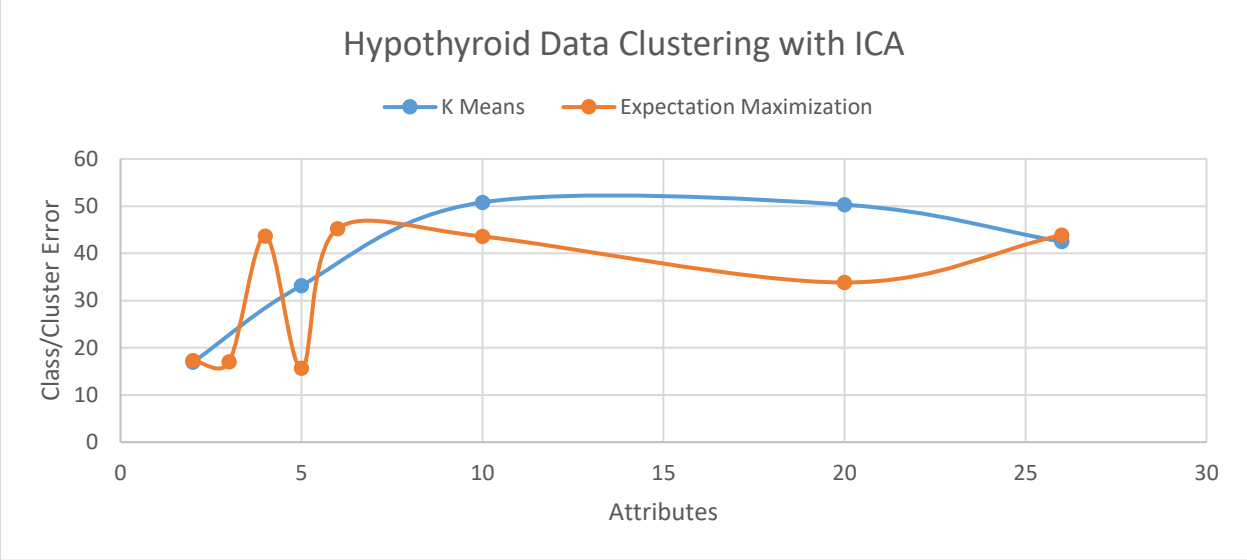





The independent components analysis algorithm was applied to the two datasets as well. The credit data benefitted greatly using expectation maximization with ICA. At 30 attributes, its clustering error fell from 40% with no modification to 32% using this method, a 20% decrease in error. However, the k-means function was much more unresponsive to the algorithm, lingering near the 40% mark before decreasing to 37% at 2 attributes.

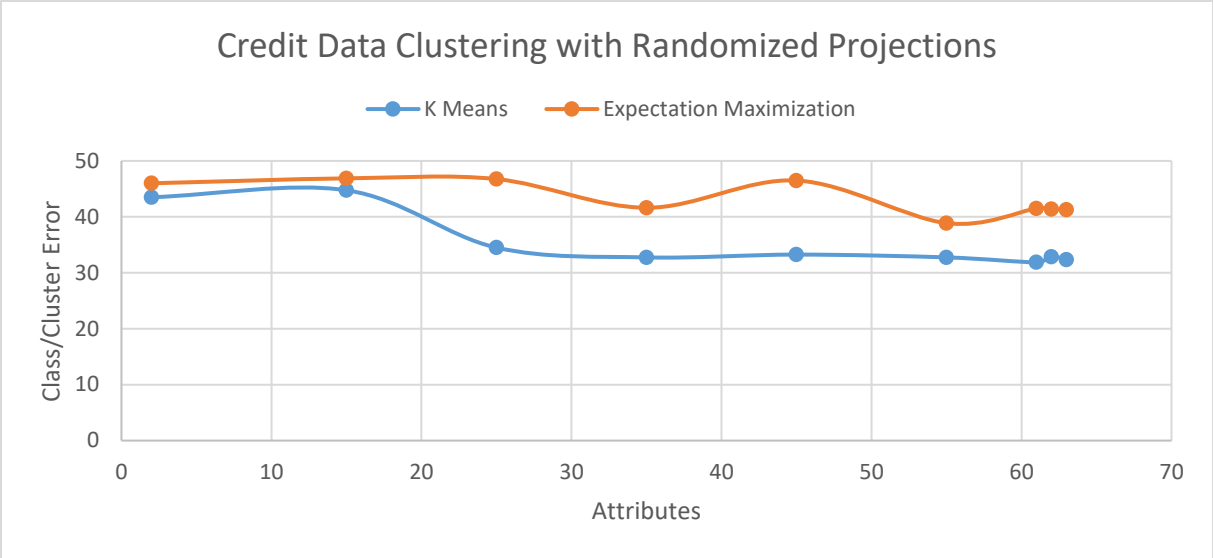
For the hypothyroid data, the behavior of the k-means and expectation maximum was quite different. The k-means algorithm begins to increase in error with a decrease in attributes, but at ten attributes the error begins to decrease to about half its original value and reaches that point, 17%, using two attributes. This is a 32% improvement over the unaltered data's error of 25%. On the other hand, expectation maximization decreases at first, then increases before finally erratically converging to the same place as k-means. This point has more error than the unaltered data using expectation maximization of 15%. Using the ICA dimensionality reduction algorithm does not improve the best clustering error performance with the hypothyroid data set but it worked well with the credit dataset.

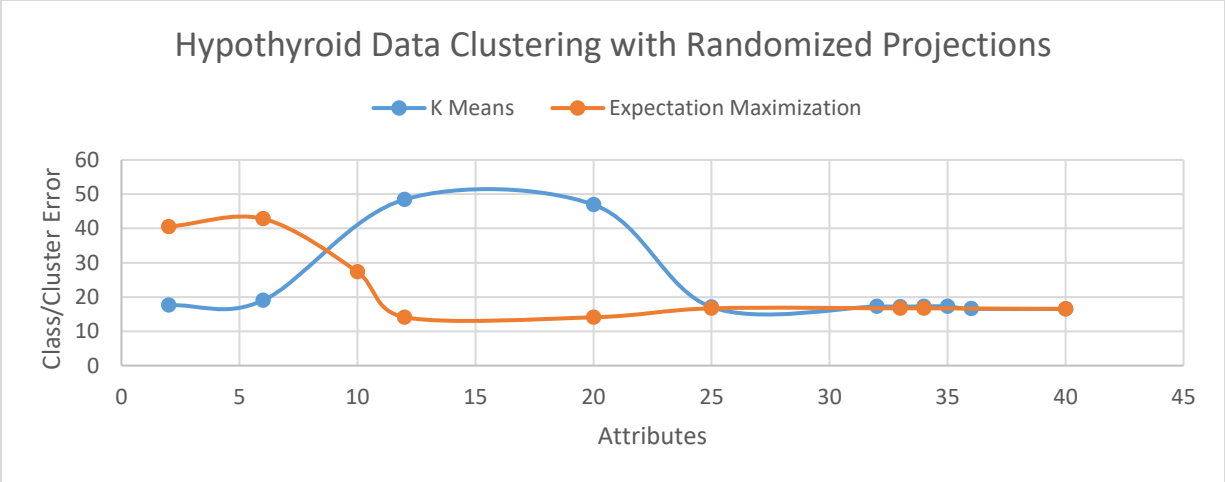


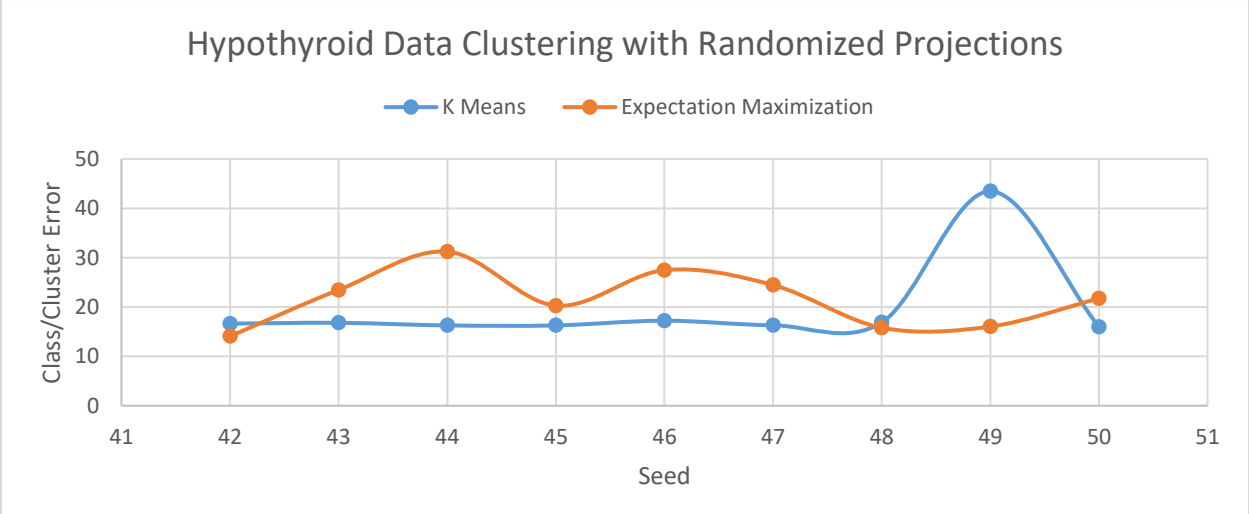




The random projection algorithm was applied to the datasets next. The credit data's k-means error has much better performance with more projections. Decreasing the number of attributes to zero yields a much higher error than the original number of attributes. This algorithm also clusters the hypothyroid data very well with a larger number of projections. K-means and expectation maximization have almost equivalent clustering error from twenty-five to forty attributes. The expectation maximization function with randomized projections achieves the lowest clustering error for the hypothyroid dataset with 14% at twenty attributes. Changing the seed yields drastically different clustering errors. The percent range for k-means

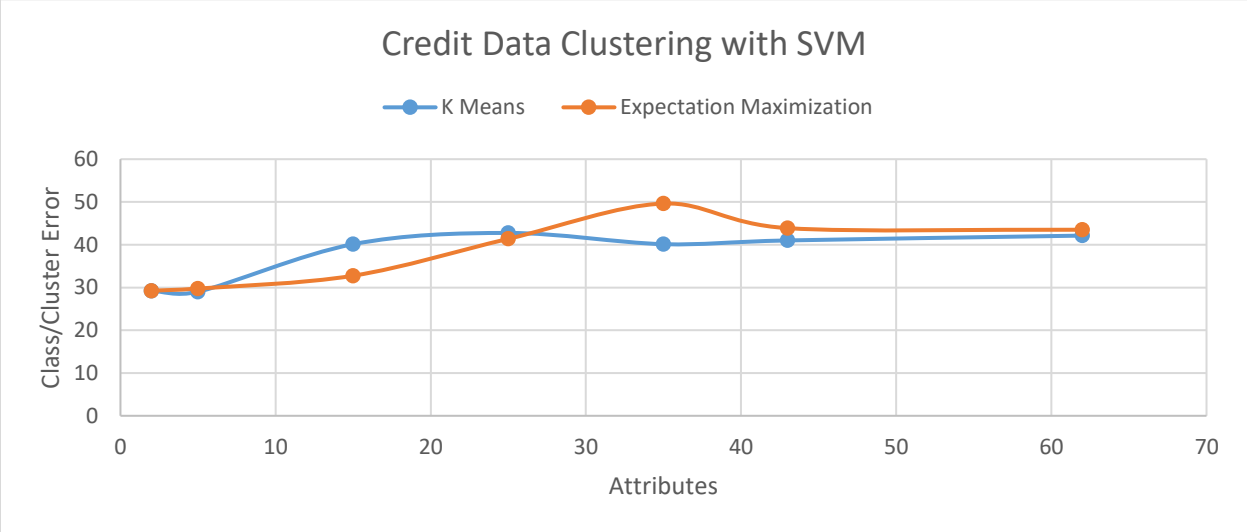
is 27%, and the range for expectation maximization is 17%. This large variance makes this method very unreliable for repeat experiments.

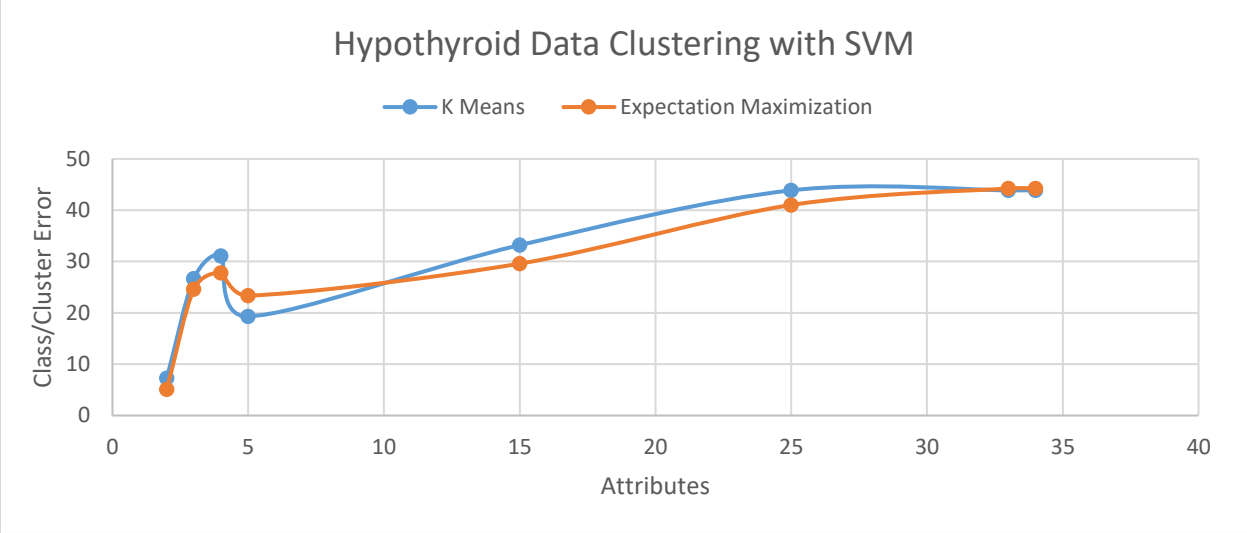


Lastly, the support vector machine algorithm was applied to the datasets. Decreasing the number of attributes rendered by this algorithm tended to decrease the error. There was a very low amount of variance between k-means and expectation maximization. However, both algorithms achieved their lowest yet error on both datasets using SVM. K-means achieved 29% clustering error with the credit data, and expectation maximization achieved 5% error. This algorithm performed best out of all four dimensionality reduction algorithms.



These reduced and clustered datasets were then run through the MultilayerPerceptron neural network in Weka. The data was trained through the neural network using cross validation to mitigate overfitting due to modeling the training data too well. This trained network was then tested using an unseen testing set which is 20% of the original data size. The results are recorded below.

| | Unclustered | | K-Means | | Expectation Maximization | |
|---|---|---|---|---|---|---|
| | % Error Train | % Error Test | % Error Train | % Error Test | % Error Train | % Error Test |
| PCA | 2.8827 | 2.558 | 3.4791 | 1.2922 | 3.2803 | 0.7952 |
| ICA | 2.8164 | 2.1432 | 2.949 | 42.7056 | 3.0152 | 44.9602 |
| RP | 2.949 | 2.122 | 3.0152 | 2.4964 | 2.8164 | 3.4483 |
| SVM | 4.0093 | 3.3639 | 4.0093 | 3.577 | 5.6991 | 4.2011 |

Previously, an unaltered set trained with 1.86% error. PCA worked best, achieving an error of .8% on the test set using expectation maximization and 1.3% using k-means. Clustering greatly lowered the error of PCA. ICA suffered a large performance hit from clustering. Overfitting occurred here due to the low training error and the very high testing error. Random projections had the lowest error without clustering, but also suffered a performance hit from clustering. SVM, though it had the best impact on clustering error, performed the worst of the dimensionality reduction algorithms in the neural network. Clustering the SVM reduced data only decreased the performance more. Here, information was lost when the data was reduced, and then lost again when the data was clustered. Reducing the dimensions resulted in a noticeable speedup in the neural network due to a lower network complexity from a smaller amount of inputs. PCA combined with expectation maximization was found to be the best algorithm for preprocessing the credit and hypothyroid datasets for neural networks.