Introduction

In this paper, we discuss unsupervised learning algorithms. These are split between feature transformation algorithms to reduce the dimensionality of the feature space and clustering algorithms to group together instances in feature space.

The feature transformation algorithms are Principal Component Analysis (PCA), Independent Component Analysis (ICA), Random Projection Filter (RPF), and Linear Determinant Analysis (LDA). The clustering algorithms are k-Means and Expectation Maximization (EM). The clustering algorithms are fully described in (Mitchell, 1997).

These algorithms are applied to two datasets, previously described in Assignments 1 and 2. The first dataset uses Dissolved Gas Analysis (DGA) from transformers to identify whether a short circuit occurred (Duval & dePablo, 2001). Preceding a short circuit, mineral oil insulation loses its dielectric strength, resulting in an internal arc. The resultant energy vaporizes the surrounding mineral oil, leading to a distribution of gases. Thus, our classification problem involves mapping the concentration of seven gases ($H_2, CH_4, C_2H_2, C_2H_6, CO, CO_2$) to a single binary indicator for arcing. The second dataset uses 13 patient health metrics to predict the presence of heart disease (Janosi, Steinbrunn, Pfisterer, & Detrano, 1988).

All simulations in this paper were performed using a modified version of the ABAGAIL Java library (Guillory, 2013).

In this paper, we first describe how the feature transformation algorithms alter our data. We follow by clustering our data. Afterwards, we cluster our filtered data. Finally, we determine whether any of the preceding data transformations are useful in order to train a neural network to predict arcing for the DGA dataset.

Feature Transformation

We begin the paper with a discussion of our feature transformation algorithms as applied to the datasets. For PCA, ICA, and RPF the data was transformed to a two dimensional dataset, whereas for LDA, the data was transformed to one dimension. The choice of two dimensions serves two goals: the classifier is binary and by using two dimensions we may get some visual intuition. In Figure 1, we see each algorithm applied to DGA, with the data plotted in the new feature space. The labels for each instance is coded as orange for positive and blue for negative.

One remarkable aspect of this Figure is that the RPF data appears to be nearly one dimensional, with the appropriate rotation. Indeed, the patterns look to be very similar to the LDA data. We may surmise that the random features selected were not linearly independent. Observing the PCA and ICA data, we see that the classification labels are roughly separable in these feature spaces. We therefore speculate that these feature transformations could assist in training a supervised learning algorithm such as an Artificial Neural Network.

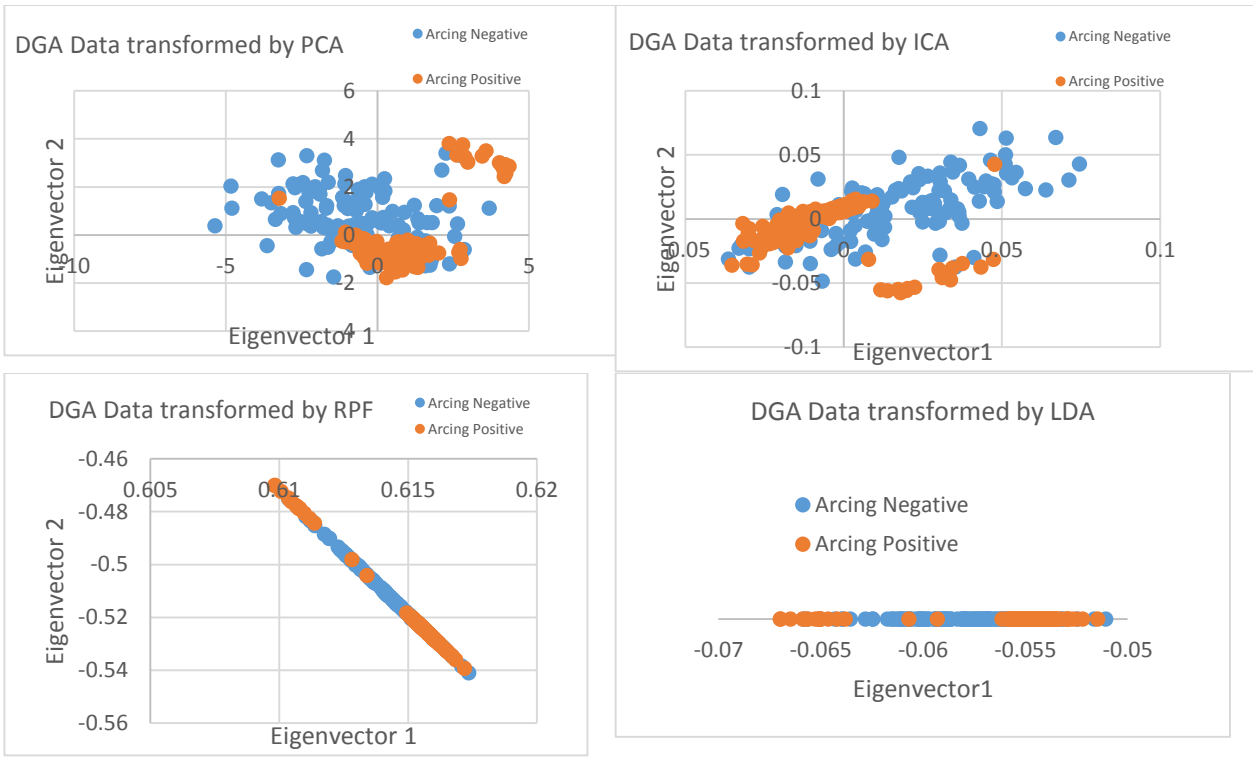

Figure 1: Feature Transformation for DGA Data

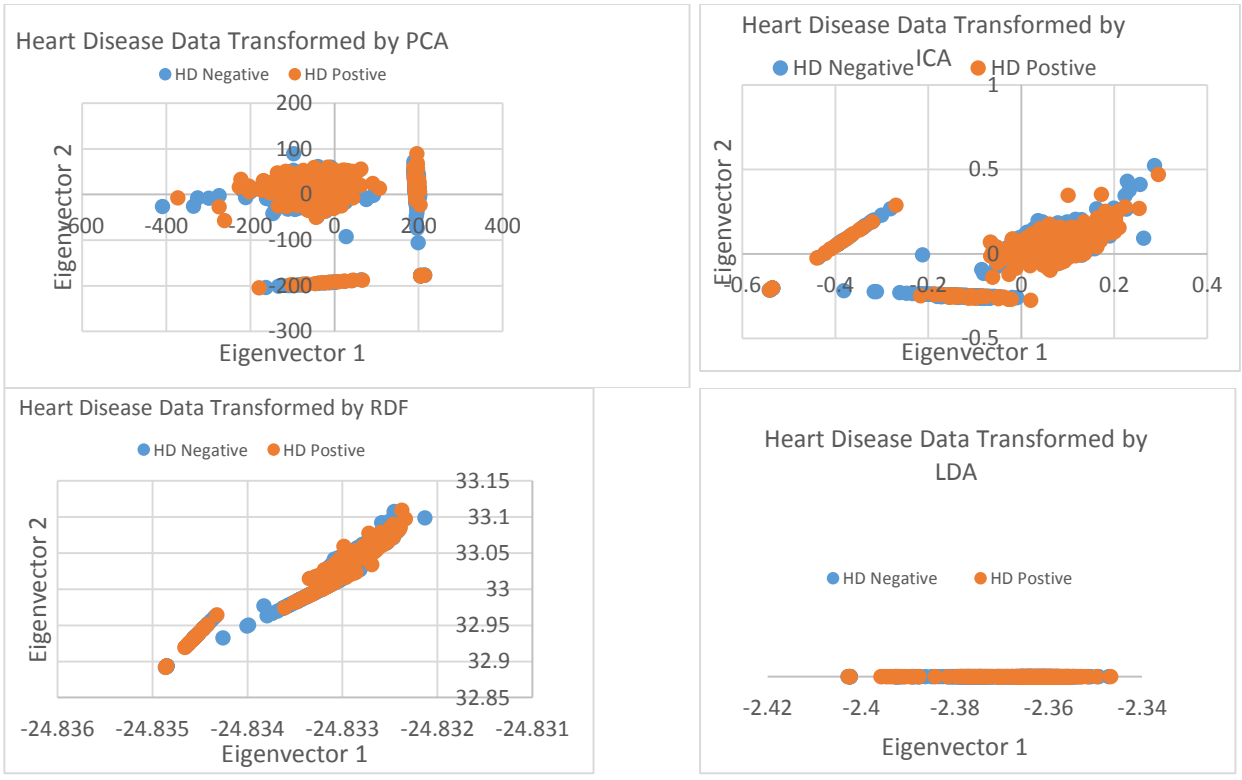

Figure 2: Feature Transformation for Heart Disease Data

In Figure 2, we see the same feature transformation algorithms applied to the HD dataset. There is more overlap between the feature labels, but there are clear patterns in separation between a positive

and negative heart disease diagnosis. Furthermore, we can clearly see a potential clustering of two to four clusters depending on the feature transformation algorithm. Finally, we see RDF appearing more one-dimensionally than PCA and ICA, although less so than with the previous dataset.

In Figures 3 and 4, we see the Eigenvalues of each Eigenvector graphed for both datasets. We note very sharp drop offs for both datasets, and particularly for the Heart Disease dataset which had to be graphed on a logarithmic scale. This implies that we may be able to reconstruct much of the information in our datasets with only a few dimensions. This is a very positive result as the number of instances needed to train a supervised learning algorithm scales exponentially with the dimension of the feature space. By reducing the number of features used, we may be able to dramatically speed up learning algorithms.

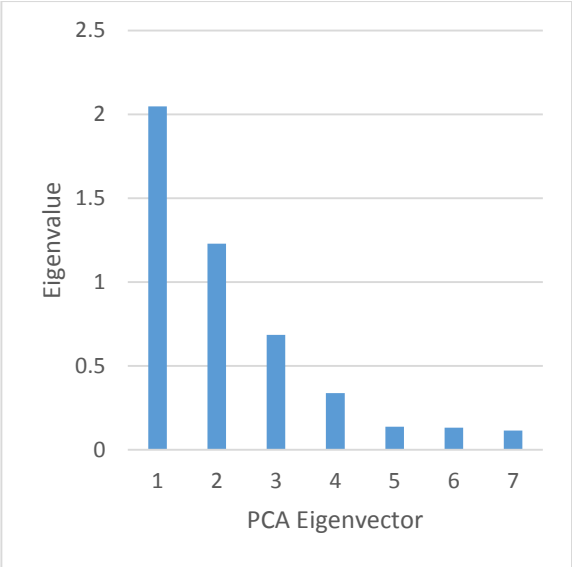

Figure 3: PCA Eigenvalues for DGA

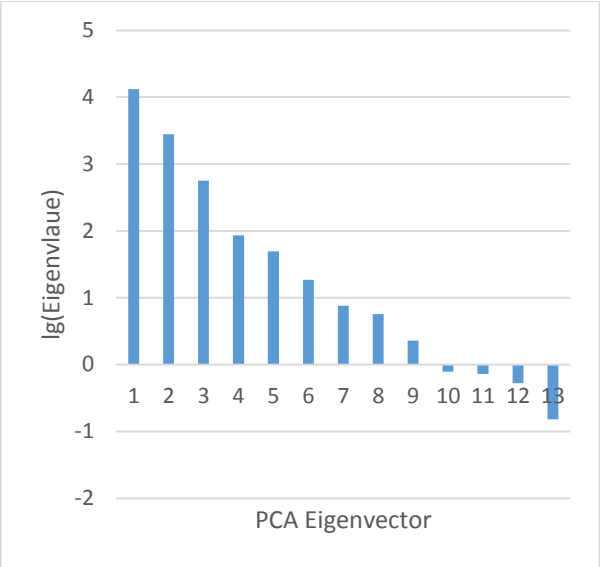

Figure 4: PCA Eigenvalues for Heart Disease

Clustering

In the next we describe some clustering results. In Figure 5, we see the comparison of k-Means and EM algorithms applied to the DGA dataset. We note that the data had to be normalized in order to reduce the likelihood that the algorithms would produce clusters of zero instances.

We choose the following performance metric: the mean distance of an in instance in a cluster to the cluster center. The first observation is that k-Means produces clusters that perform better in this metric. The second observation is that beyond approximately four clusters, there are rapidly diminishing returns for additional clusters. Finally, increasing the number of iterations for the EM algorithm beyond ~1000 does not produce appreciably better clusters.

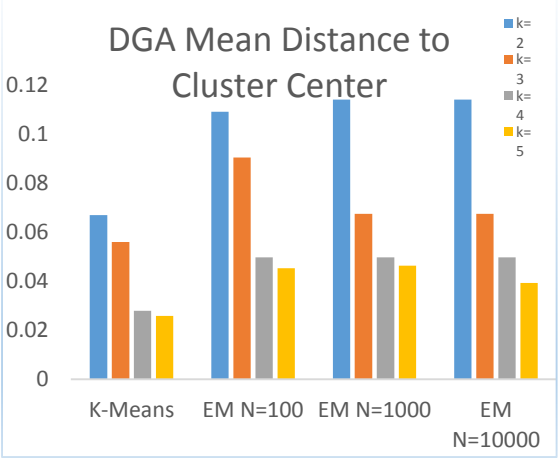

Figure 5: Cluster Performance for DGA

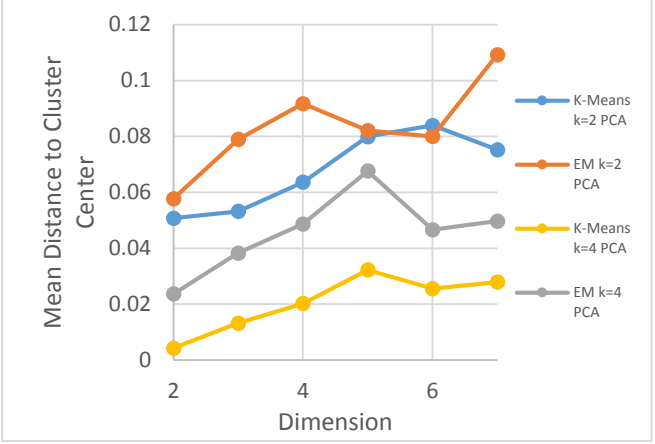

Figure 6: Clustering DGA Data Transformed by PCA

In Figure 6, we see the clustering algorithms applied to the DGA data after it is transformed by PCA. We note that our performance metric is of questionable validity in this context: if we suppose that features have roughly the same range of values after a transformation, any transformation that reduces the dimensionality of the feature space will tend to reduce the average L_2 distance between any two points. However, we will still posit that PCA using two eigenvectors is the optimal choice based on this measurement for either clustering algorithm. We note in addition that for this choice of PCA, for a clustering of k=2, EM performs better than k-Means, while for k=4 k-Means performs better than EM. Finally, we note that a larger number of clusters increases our performance metric, reinforcing the observation from Figure 3.

In Figure 7, we see the clustering algorithms applied to the DGA data after it is transformed by ICA. We note that the performance is far worse than the data filtered by PCA. In addition, we note that a monotonic decrease in our performance metric with increasing number of dimensions for the transformed feature space. We may surmise that is related to our observation above that increasing the dimensions will tend to increase distances. The relatively poor performance of ICA may result from a strong ordering of features that exists in the data that PCA is able to better represent. For example, the presence of certain subsets of gases such as Acetylene, may be more important indicators of arcing.

In Figure 8, we see the clustering algorithms applied to the DGA data after RPF transformations. We note very good performance particularly for k=4, even beating the PCA data in Figure 4. We surmise that this may be related to the projection of the data onto approximately one dimension as shown in Figure 1, which allows good performance for this metric for the same reasons as above.

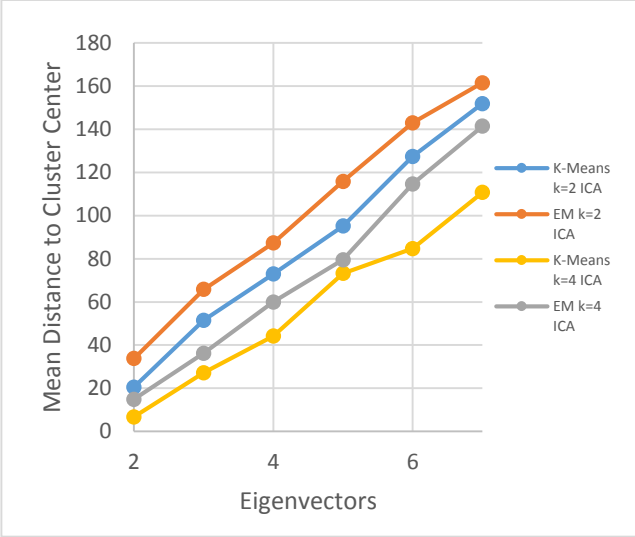

Figure 7: Clustering DGA Data transformed by ICA

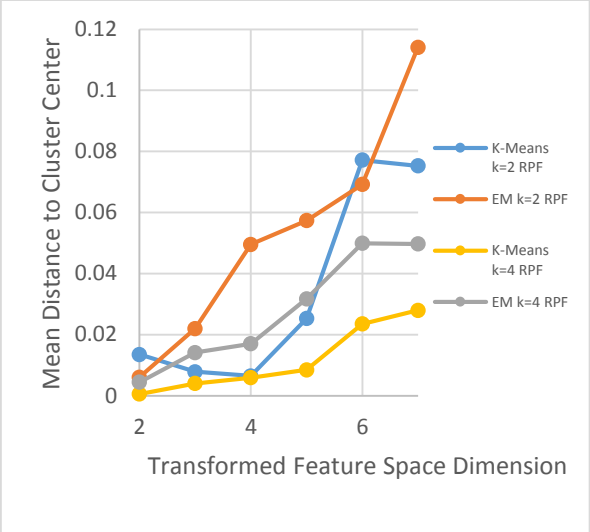

Figure 8: Clustering DGA Data Transformed by RPF

In Figures 9 and 10, we show the entropy associated with the DGA cluster distribution, as defined by $\sum_k -p_k log_2(p_k)$. A high entropy implies a more uniform split. We observe generally relatively low entropy for the ICA data which performs poorly, and generally higher lower entropy for the RPF data which performs well. However, this correlation is imperfect, we can't attribute the poor performance of ICA to an unbalanced distribution of instances between clusters.

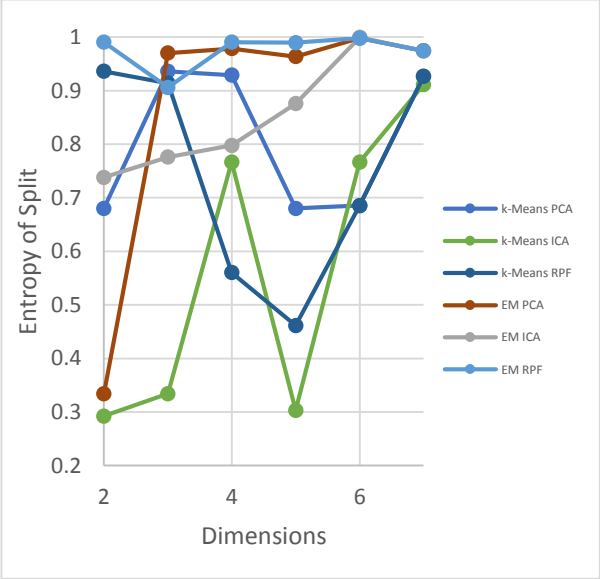

Figure 9: Entropy of DGA Clusters for k=2

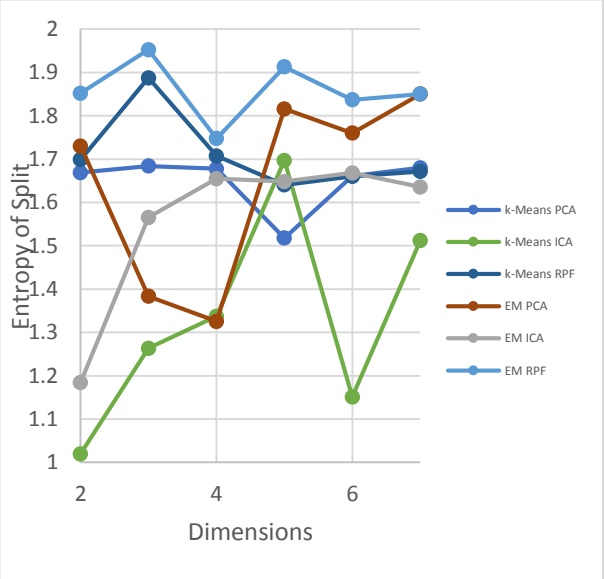

Figure 10: Entropy of DGA Clusters for k=4

In Figure 11, we see the general performance of all the algorithms for a clustering of two and a feature dimension of two where appropriate, with the exception of LDA. It is apparent that LDA is by far the best performing, possibly due to its ability to leverage the label to better cluster the data. We note that ICA is not shown on this graph due to its very poor performance. Indeed, it is much worse performing than clustering the untransformed data.

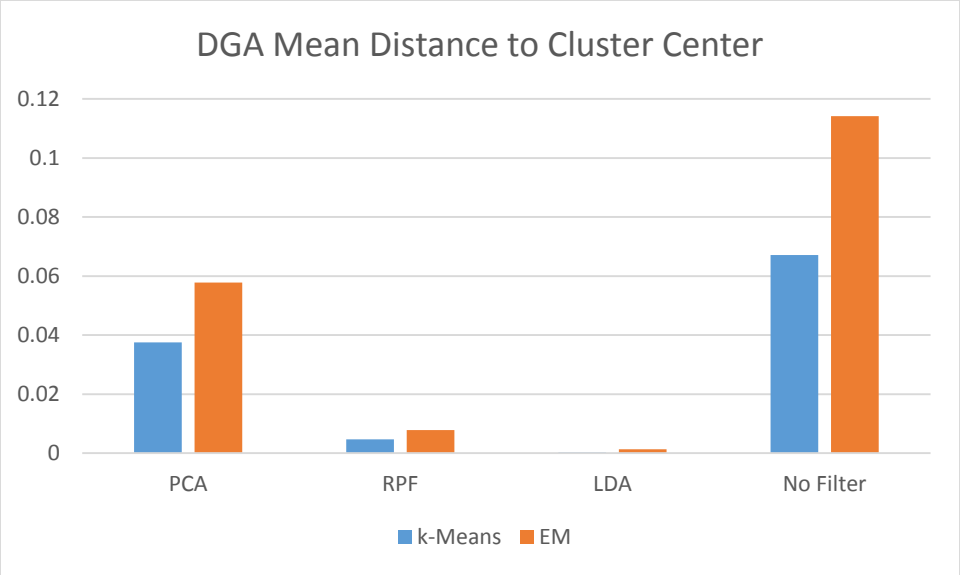

Figure 11: Performance of Clustering Algorithms Applied to Filtered DGA Data

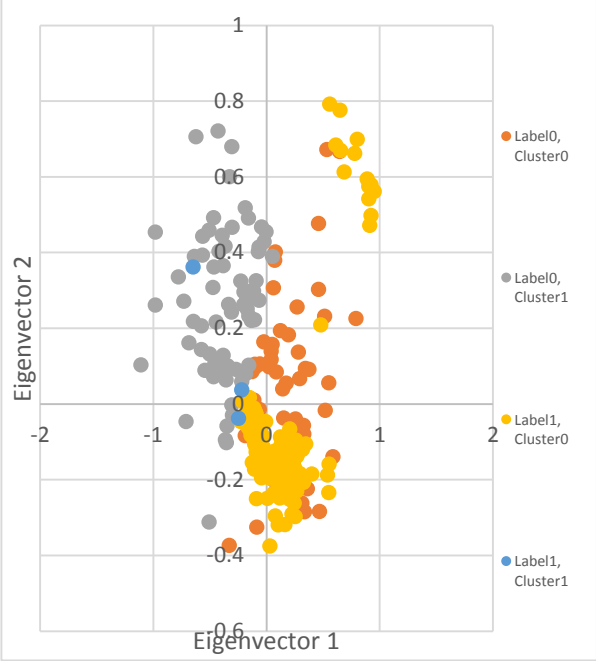

Figure 12: k-Means k=2 Clustered DGA PCA Data Projected onto PCA Eigenvectors

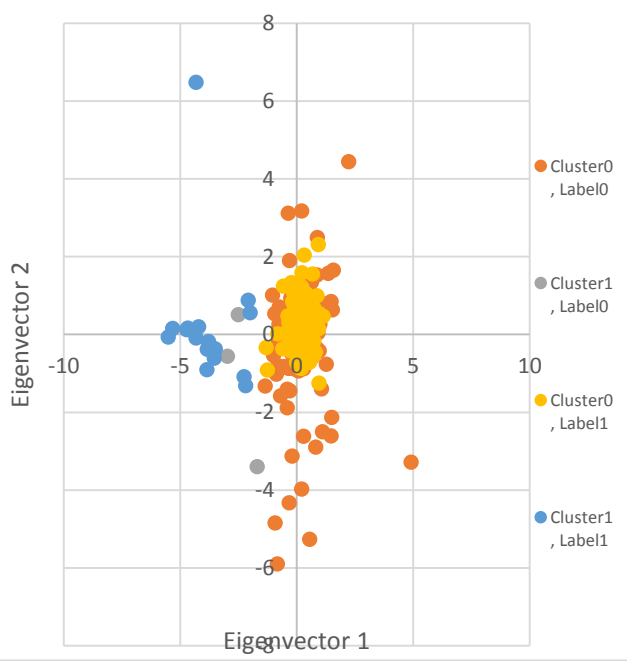

Figure 13: k-Means k=2 Clustered DGA ICA Data Projected onto ICA Eigenvectors

In Figures 12 and 13 we see a visual representation of the PCA and ICA filtered data after it has been clustered by the k-Means algorithm into two clusters. It is clear that the PCA filtered data is able to better separate into two clusters. Furthermore, this separation has significant overlap with the labels, as can be seen from the large numbers of instances either in Label0, Cluster1 or Label1,Cluster0 sets.

Now that we have some intuition for our clustering algorithms, we move onto the somewhat more complex heart disease dataset. In Figure 10, we see the clustering performance applied to the heart

disease dataset. In contrast to the DGA dataset, the increasing numbers of clusters dramatically decreases mean distance to cluster centers, particularly with the Expectation Maximization.

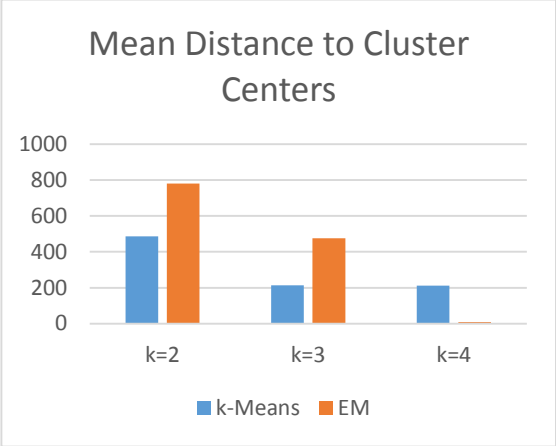

Figure 14: Clustering Performance of Unfiltered Heart Disease Data

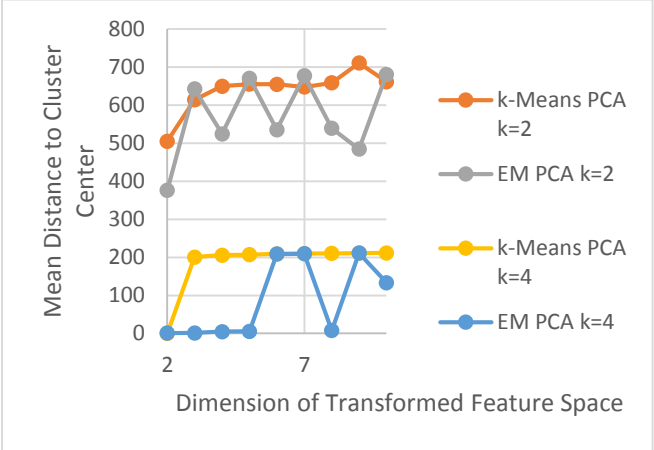

Figure 15: Clustering of Heart Disease Data Transformed by PCA

In Figure 14, we show clustering results from using the Heart Disease dataset transformed by PCA. In contrast to the DGA data, the difference between clustering performance of 4 clusters compared to 2 clusters is enormous. This is consistent with our visual indication of four different groups of instances as seen in Figure 2. We also note that EM nearly always performs better than k-Means.

In Figure 16, we show the results where the data is transformed by ICA. We generally see that the clustering performs better than PCA, with the disputable exception of EM k=4. As with the DGA dataset, the mean distance to the cluster center monotonically increases with the number of dimensions.

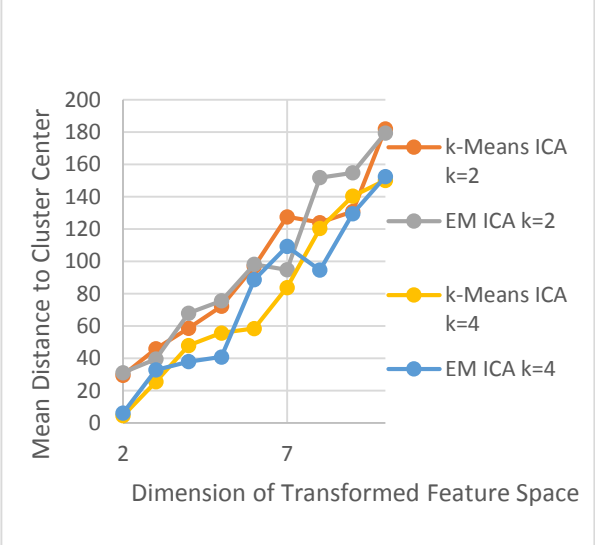

Figure 16: Clustering of ICA Heart Disease Data

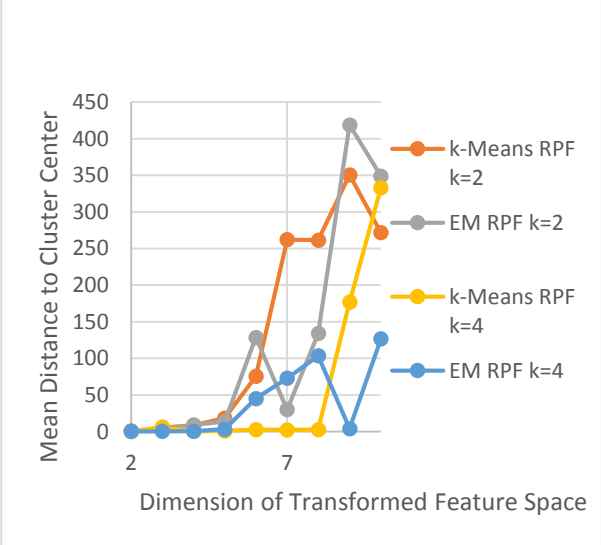

Figure 17: Clustering of RPF Heart Disease Data

In Figure 17, we see the clustering algorithms applied to the heart disease data transformed by RPF. Greater than a dimension of approximately five, we generally see much worse performance. In Figure 18, we again see LDA outperforming the other feature transformation algorithms, although the ICA does not fare quite as poorly as in the DGA dataset. Instead, it is the unfiltered data which performs

sufficiently badly as to not be represented on the same scale. In Figures 19 and 20, we see a visual representation of our data for clustering of two and four using k-Means, confirming that k-4 is more appropriate for this dataset.

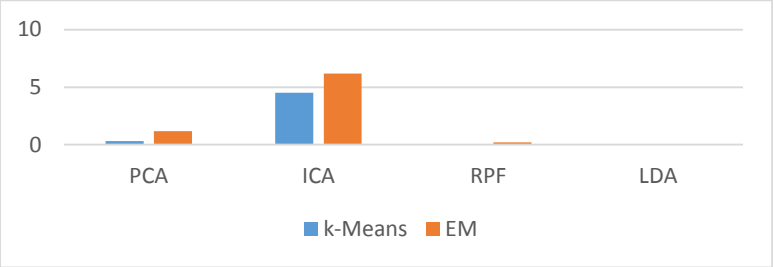

Figure 18: Clustering Performance for all Filtering Algorithms on Heart Disease Data, k=4

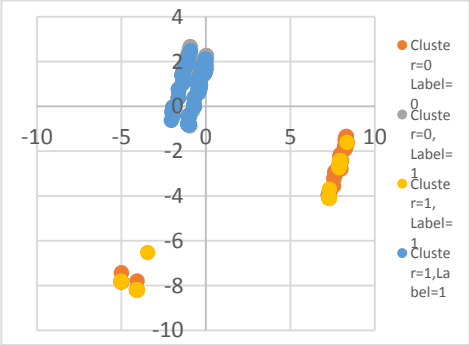

Figure 19: Distribution of Heart Disease Instances Visualized on PCA Eigenvectors k=2

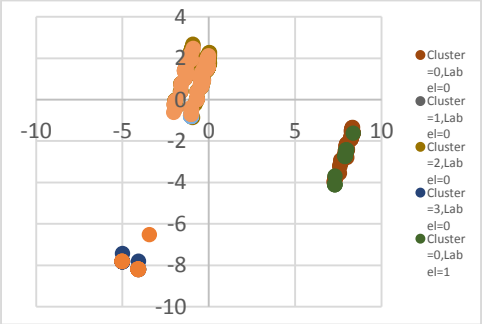

Figure 20: Clustering of Heart Disease Instances Visualized on PCA Eigenvectors k=4

Neural Network Training

Finally, we feed our transformed and clustered DGA data into an Artificial Neural Network. We begin by investigating the training and testing accuracy of our DGA data subject to our feature transformation algorithms. The PCA, ICA, and RPF dimensions were chosen to be 2 since those inputs lead to relatively good clustering results as seen in the previous section. The data is shown in Figures 21 and 22. It is notable that even LDA does not stand out as a strong performing algorithm, even though it uses label information to cluster data. Furthermore, we note that over-fitting becomes prominent after ~2.5e-3 iterations. Up to this number of iterations, PCA performs relatively well compared to the other algorithms.

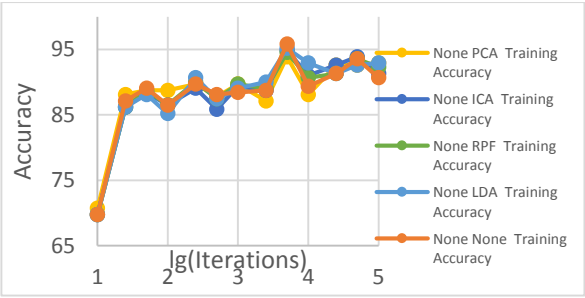

Figure 21: DGA Neural Net Training Accuracy, No Clustering

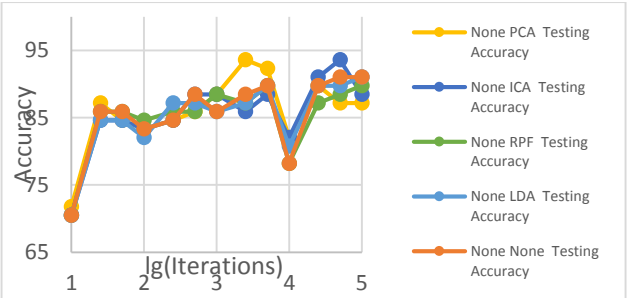

Figure 22: DGA Neural Net Testing Accuracy, No Clustering

In Figures 23 and 24, we first apply the k-Means clustering algorithm with two clusters, define the cluster as an additional feature, and then apply the five feature transformation algorithms (including a lack of filter). We get significantly better performance compared to the lack of clustering. It becomes harder to identify a filtering algorithm that leads to particularly good performance if Figure 24, but we can identify RPF and PCA as doing relatively well for a small number of iterations.

In Figures 25 and 26, we first apply EM clustering with two clusters, define the cluster as an additional feature, and then apply the five feature transformation algorithms. We observe similar results as for k-Means clustering, but none of the filtering algorithms perform appreciably better than a lack of filtering.

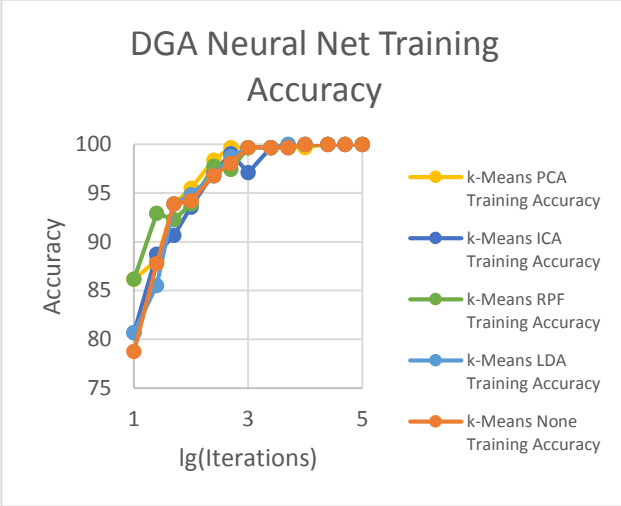

Figure 23: DGA Neural Net Training Accuracy k-Means Clustering k=2

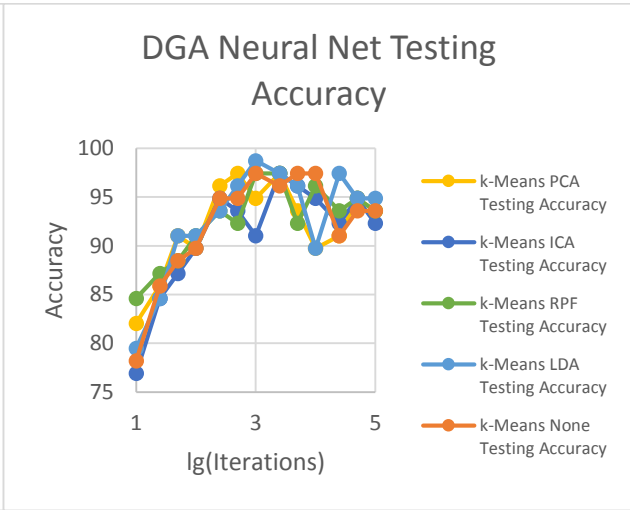

Figure 24: DGA Neural Net Testing Accuracy k-Means Clustering k=2

In Figure 27, we show the time cost of training the neural net with the transformed data, after clustering. There is no significant improvement compared to the unfiltered data.

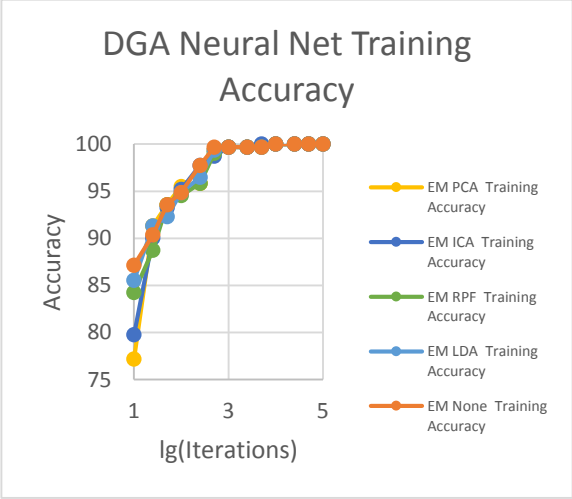

Figure 25: DGA Neural Net Training Accuracy, EM Clustering k=2

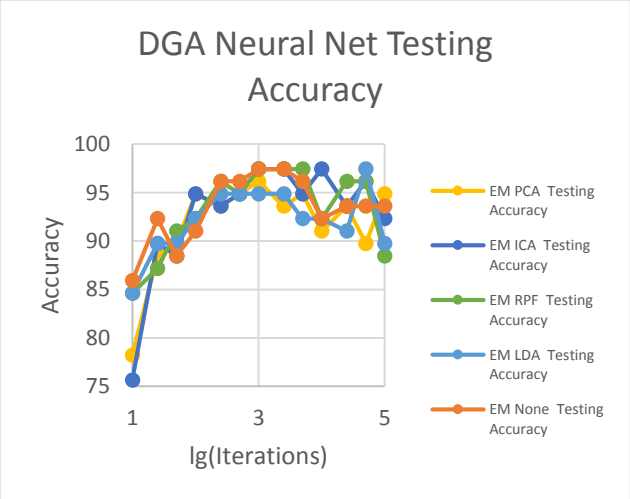

Figure 26: DGA Neural Net Testing Accuracy, EM Clustering k=2

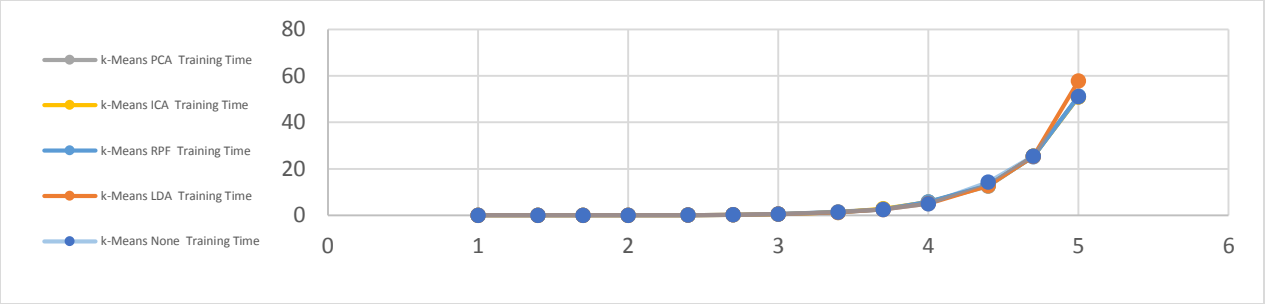

Figure 27: Neural Net Training Time with k-Means Clustering

Conclusions

In this paper, we have investigated unsupervised learning and dimensional reduction algorithms. These algorithms were applied to two datasets: one using patient information to predict heart disease, and a second using Dissolved Gas Analysis to predict whether internal arcing occurred in a transformer.

In the first section of the paper, four feature transformation algorithms were applied to each dataset. These are Principal Component Analysis, Independent Component Analysis, Random Project Filter, and Linear Discriminant Analysis. Of the four, PCA grouped data in a distribution most consistent with the labels. In addition, based on the PCA transformation, we observed that only a few eigenvectors are important in terms of representing information in each dataset.

Armed with this knowledge, we proceeded to cluster the data. For the DGA dataset, increasing the number of clusters much beyond two didn't produce large decreases in the mean distance to the cluster center. This was not the case with the heart disease dataset, where a clustering of four produced dramatically better results. Both observations were consistent with the number of clusters appearing in the PCA transformed data.

We then clustered the transformed data. We observed that LDA was most effective in terms of minimizing our performance metric. This is consistent with the observation that the labeled data can form spatially distinct clusters, particularly for DGA.

Finally, we used a neural net to train the data that was subject to feature transformations, and then data that was subject to clustering as well as the feature transformations. We observe no significant improvement in terms of the feature transformations, but the clustering allows us to produce an accurate neural net in fewer iterations.

References

Duval, M., & dePablo, A. (2001). Interpretation of gas-in-oil analysis using new IEC publication 60599 and IEC TC 10 databases. *IEEE Electrical Insulation Magazine*.

Guillory, A. (2013). *ABAGAIL*. Retrieved from Pushkar's Github Repository: https://github.com/pushkar/ABAGAIL

Janosi, A., Steinbrunn, W., Pfisterer, M., & Detrano, R. (1988). *Heart Disease Data Set.* Retrieved from UCI Machine Learning Repository: http://archive.ics.uci.edu/ml/datasets/Heart+Disease

Mitchell, T. (1997). *Machine Learning.* McGraw Hill.