# Unsupervised Learning and Dimensionality Reduction

Rohan Ramakrishnan
Georgia Institute of Technology
rohanrk@gatech.edu

***Abstract*—In this analysis, clustering algorithms are applied to two datasets, a breast cancer dataset with 289 instances and 9 attributes and a dataset with pictures of handwritten digits. This dataset contains 5000 instances and 64 attributes. K-Means Clustering and Expectation Maximization were used to generated clusters while Principle Components Analysis (PCA), Independent Components Analysis (ICA), and Randomized Projections were used to show the effects of feature transformation and dimensionality reduction on the datasets.**

## I. Clustering

### A. *K-Means Clustering*

K-Means Clustering is a method that selects random data points and designates them as centers. Each center then 'claims' the closest data points using some similarity or distance metric. The centers are then recalculated by averaging the clustered points. This process is repeated until convergence and the corresponding partition is returned.

Figures 1 and 2 show the results of applying K-Means Clustering to both datasets. Euclidean distance was used as the distance metric to determine the distance/similarity of each data point to the centers. Note that sum of the distances between each point and the center of its designated cluster is known as distortion. Since this is a representation of error that should be minimized, it is possible to use randomized optimization techniques in conjunction with clustering techniques to minimize distortion. This process is not included in this paper for the sake of brevity.

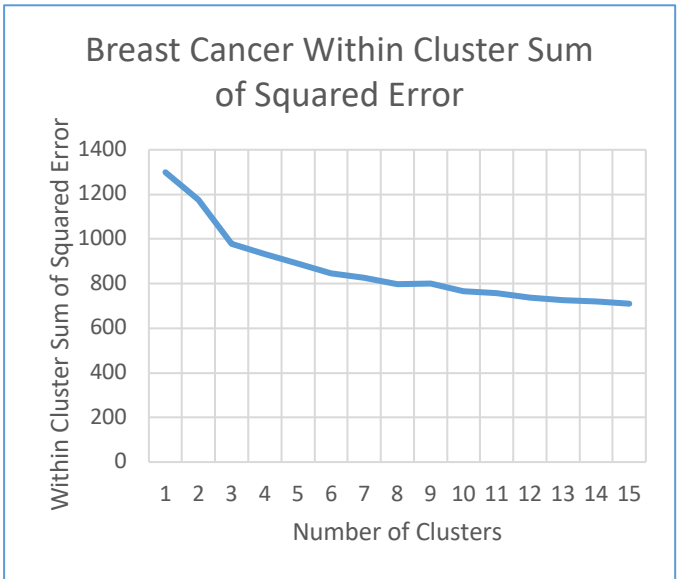


*Figure 1: K-Means Clustering Performance on Breast Cancer Dataset*

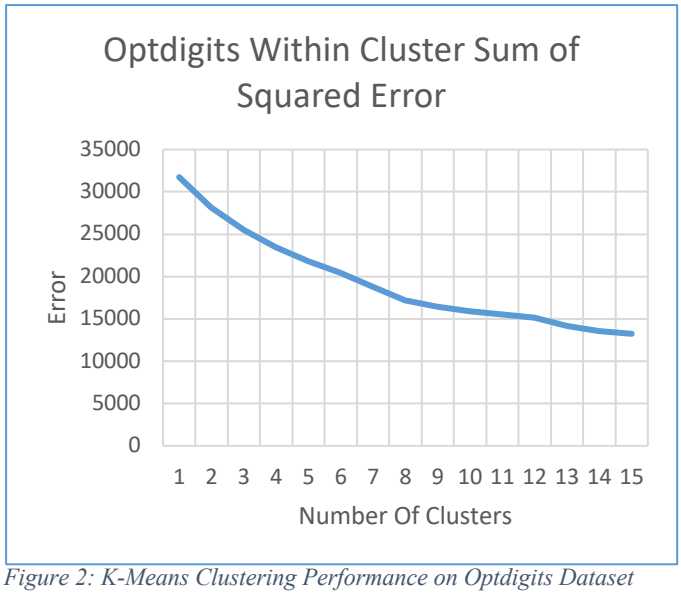


*Figure 2: K-Means Clustering Performance on Optdigits Dataset*

Figure 3 shows the class to cluster error for each dataset with varying number of clusters or k. Each cluster is mapped to a class based on the majority number of instances of a particular class in a cluster. If the majority class is already assigned to a cluster, the next major class is selected. Finally, once all classes are assigned, extra

clusters remain unassigned to any class. The class to cluster error is defined as all the instances that were mapped to the incorrect cluster based on their class. The results show that after the number of clusters exceeds the number of classes of both datasets, the error either slowly converges or increases. One cluster will always map exactly one class correctly and misclassify the rest of the instances. So for a binary dataset, one cluster generates very low error while for a multiclass dataset, one cluster will generate very high error.

The breast cancer error continues to increase after two clusters. This result is expected as there are more clusters that are assigned to no class and several instances are assigned to these unassigned clusters. For Optdigits, the error decreases until around eight clusters and then evens out around 30%. More analysis on these results can be found after the Expectation Maximization section.

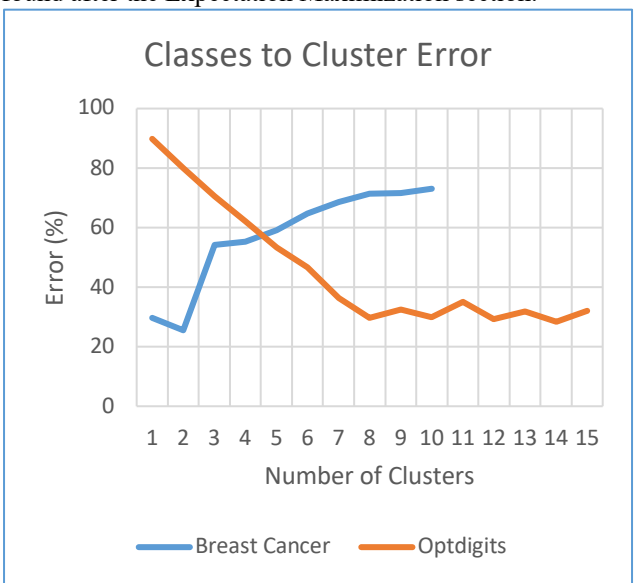


*Figure 3: K-Means Clustering Classes to Cluster Error*

Figures 4 and 5 show the original class distribution of each dataset. For the purposes of this report, The blue bar represents negative instances whereas the red bar represents positive instances for the Breast Cancer dataset.

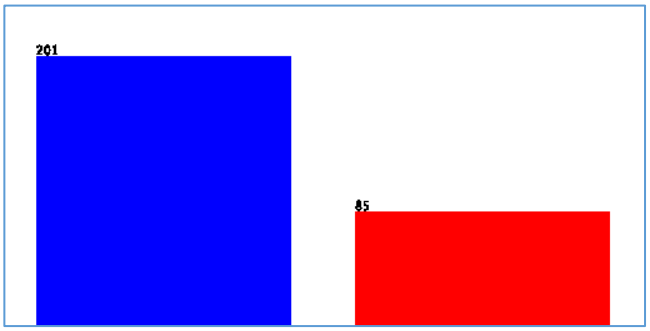


*Figure 4: Breast Cancer Class Distribution*

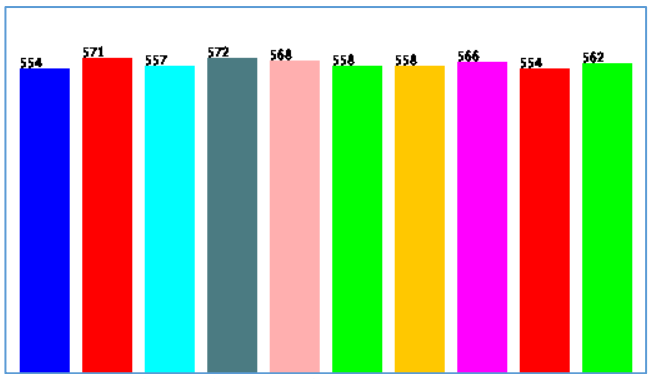


*Figure 5: Optdigits Class Distribution*

Finally, figures 6 and 7 show the clustering visualizations of both datasets.

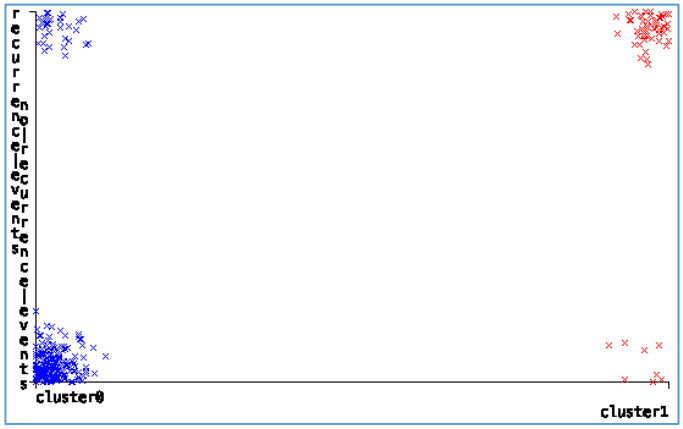


*Figure 6: Breast Cancer K-Means Cluster Visualization*

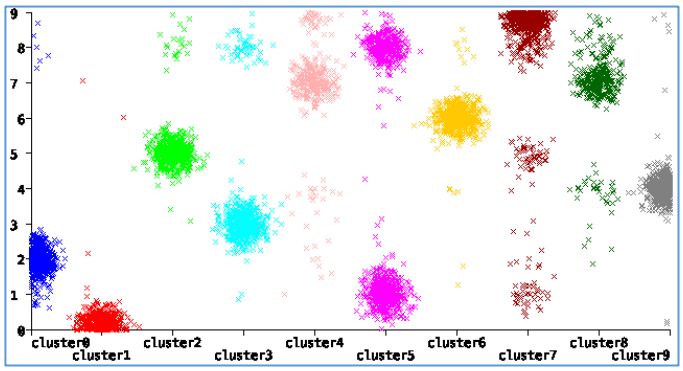


*Figure 7: Optdigits K-Means Cluster Visualization*

There are two particular objectives of clustering algorithms that will be used to evaluate these visualizations: Homogeneity and completeness. Homogeneity of a clustering assignment is where each cluster contains only members of a single class. Completeness of a clustering assignment is where all members of a given class are assigned to the same cluster.

For both visualizations, k was selected to be the number of class labels as that value of k best display K-Means

progress on achieving homogeneity and completeness on both datasets.

Although Weka does not display numerical scores, the visuals can be evaluated on how well each cluster represents a particular class. K-Means does well in classifying the breast cancer dataset. Most of the negative instances are located in the first cluster and the positive instances are located in the second cluster. The clusters also have a similar degree of unevenness in the number of instances assigned. The first cluster clearly contains most of the instances while the second contains significantly fewer. This is also represented in the class distribution in figure 4.

Figure 7 shows that most of the clusters contain a majority of instances in a particular class. The one exception is cluster 5 where there are several instances of class 1 and 8 that belong to that particular cluster. The class to clusters errors greatly exceeds the error of supervised learning techniques applied to both datasets.

*B. Expectation Maximization*

K-Means is strict in its assignment of points to clusters, so if data points are close to two different clusters, this attribute is not reflected when running K-Means. Expectation Maximization is a soft clustering technique that attempts to show points' similarities with different clusters. Soft clustering assigns probabilities to each point based on their likelihood of belonging to a particular cluster. Expectation Maximization alternates between two probabilistic calculation to calculate the expectation of each distribution for each point and then attempt to sample to maximize that expectation.

Figures 8 and 9 show the results of applying Expectation Maximization to both datasets.

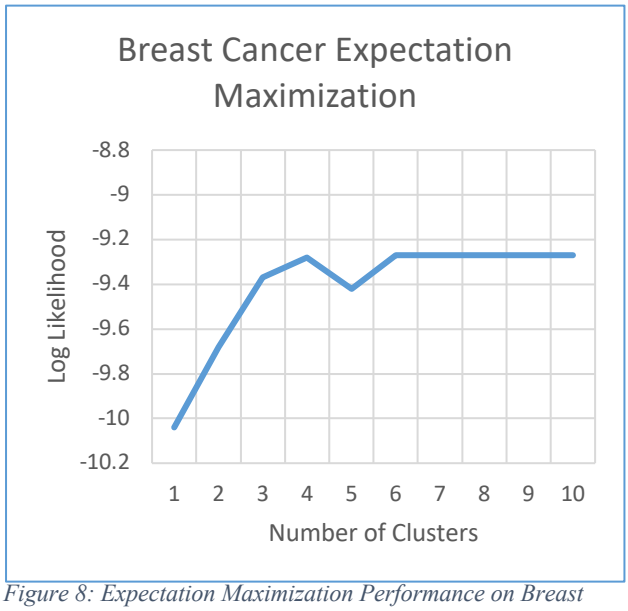


*Figure 8: Expectation Maximization Performance on Breast Cancer Dataset*

Log Likelihood is defined by the natural logarithm of the likelihood of seeing a particular clustering assignment. Since Expectation Maximization is attempting to find the "most likely" clustering (maximum likelihood hypothesis), the log likelihood is a score with a max value of 0 because the likelihood is defined by probabilities and the values of the the attributes. The attributes of both datasets are either numerical or nominal such that their values are greater than 0. This means that the likelihood is a value between 0 and 1. Therefore the natural log of the likelihood has an upper bound of 0 and no lower bound.

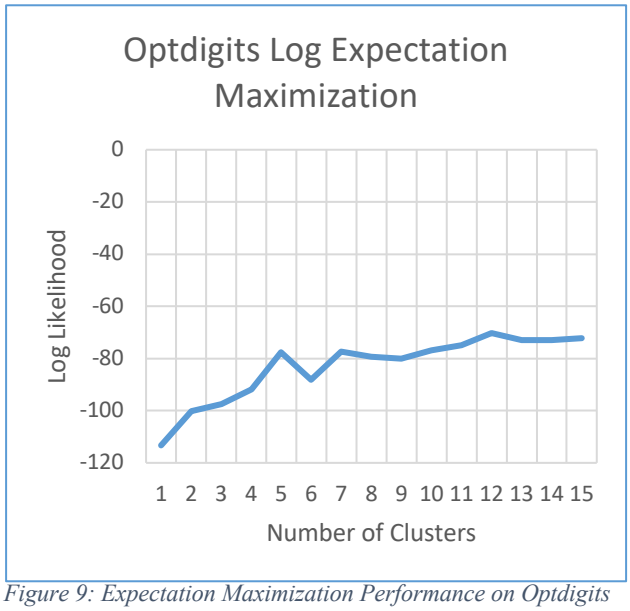


*Figure 9: Expectation Maximization Performance on Optdigits Dataset*

Figure 10 shows the class to cluster error for each dataset. The Breast Cancer dataset was run only till 10 instances since the error was monotonically non-increasing.

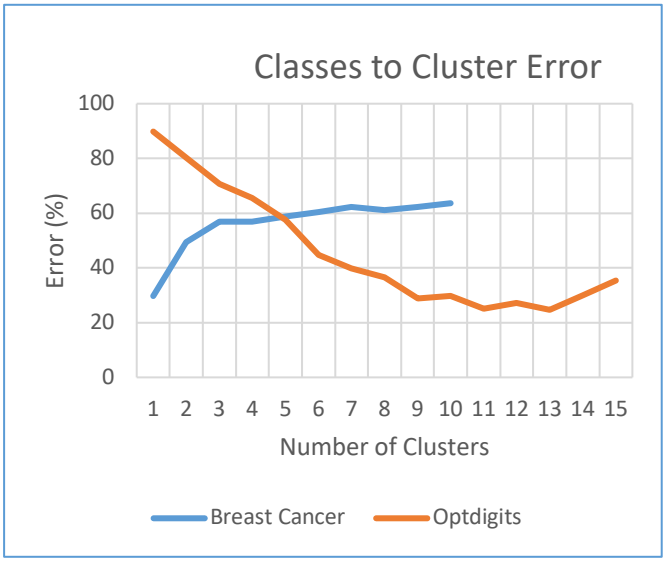


*Figure 10: Expectation Maximization Classes to Cluster Error*

*1) Classification Using Clustering*

A cursory glance at the class error vs number of clusters for both methods shows that both algorithms' performance start to decline after the number of clusters exceeds the number of features. This is intuitive as having fewer clusters causes clustering algorithms to underfit since there aren't enough clusters to cover each class. However having more clusters causes the algorithms to overfit. With more clusters, there are several clusters that do not correspond to any class. This means each instance has a higher chance of being misclassified. For this analysis, a misclassified instance is an instance that has been assigned to a cluster with a different majority class. The classes to cluster error for EM is significantly smoother than that of K-Means. This fits as EM is a soft clustering algorithm and provides cluster assignment based on the maximum expectation which requires several iterative probabilistic calculations. Another interesting note is that K-Nearest Means provides significantly better results for the Breast Cancer dataset. For 2 cluster, K-Means reaches a low of 25% error whereas the erro with EM is monotonically increasing as the number of clusters increases. This makes sense based on the class distribution of the dataset so K-Means works better as it forces several instances to be negative instances while EM's soft clustering doesn't take into account the class distribution in such a direct manner.

## C. *Manhattan Distance*

The distance/similarity metric utilized in K-Means is a representation of domain knowledge with respect to the data. For the sake of comparison, K-Means was run using Manhattan Distance to compare against its performance using Euclidean Distance. The Breast Cancer dataset had identical class to cluster errors given both distance metrics. This means that taking the average distance with respect to the attributes, and individually taking the distances of its attributes yield the same result. This implies that the attributes of the dataset are sufficiently independent when assigning instances to clusters.

The Optdigits' dataset shows slightly worse performance with respect to class to cluster error when K-Means uses Manhattan distance to assign instances to clusters. Figure 12 show these results.

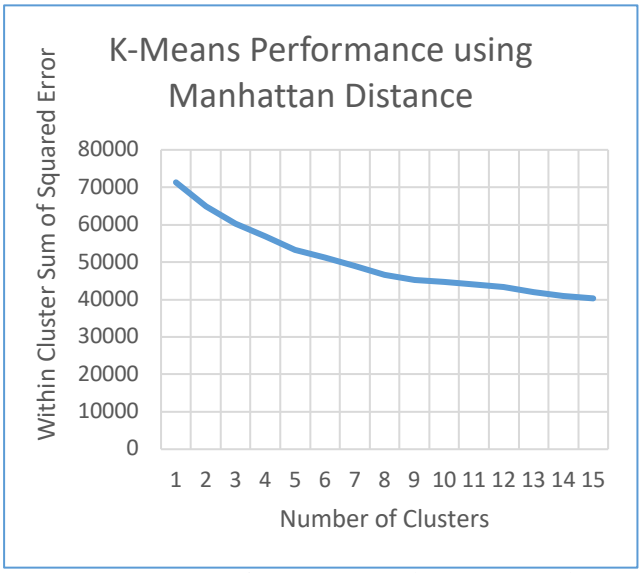


*Figure 11: K-Means Performance on Optdigits using Manhattan Distance*

Figure 11 shows the results of applying K-Means to the Optdigits dataset. The squared error is significantly larger than the Euclidean Distance error but this is not an indicator of poor performance since Manhattan distance is always greater than Euclidean distance since it takes the independent distance between each dimension. The error curve is also monotonically decreasing like its Euclidean Distance counterpart.

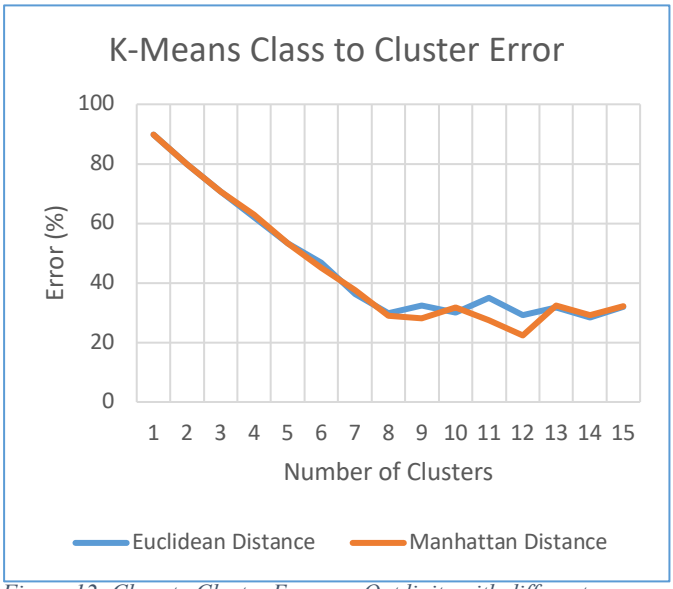


*Figure 12: Class to Cluster Error on Optdigits with different distance metrics*

The class to cluster error is very similar and Manhattan Distance even dips below the Euclidean Distance error. This is most likely due to the random initialization of the k centers since it is impossible to have a set of points such that the pair with the minimum Euclidean distance is

different from the pair with the minimum Manhattan distance.

## II. Dimensionality Reduction

Datasets with many features may provide many relevant and useful features, but they fall prey to the curse of dimensionality which states that the number of samples required to generate a consistent hypothesis is exponentially proportional to the number of features. Dimensionality Reduction attempts to resolve this issue by reducing the number of features considered. It attempts to transform the current features into a different and more useful set of features. For the purposes of this analysis, the transformed features represent linear combinations of the original features. For the purposes of this analysis, the algorithms analyzed are using filtering methods meaning that they do not account for model bias of any particular supervised learner.

### *A. Principal Components Analysis (PCA)*

Principal Components Analysis attempts to transform the features into a a dimension that maximizes the variance of the data. Furthermore, it selects dimensions that are mutually orthogonal. Weka's Principle Components filter was applied to both datasets with a default variance of 0.95. Then both clustering algorithms were run on the datasets once their attributes were modified.

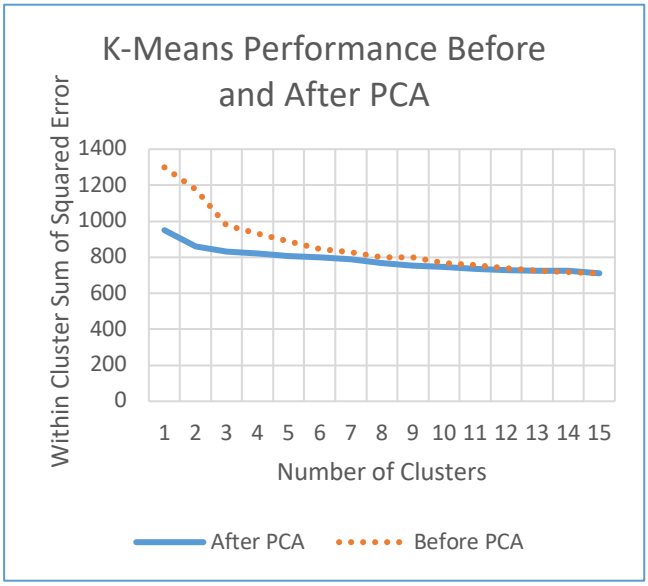


*Figure 13: K-Means Performance on Breast Cancer Dataset after PCA*

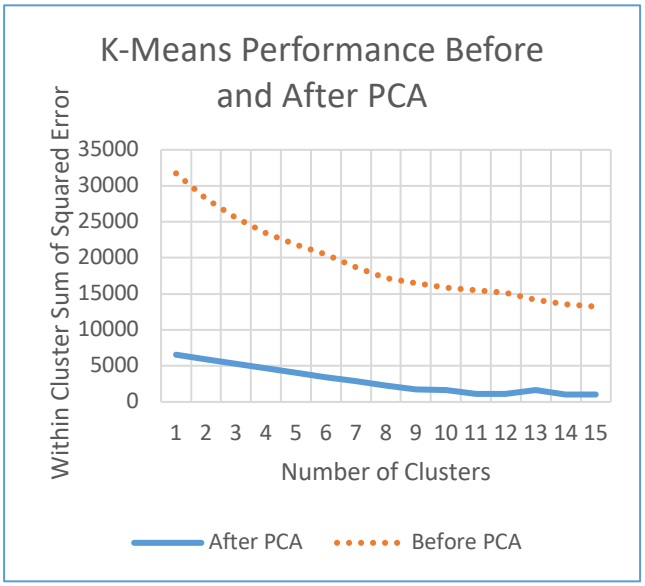


*Figure 14: K-Means Performance on Optdigits Dataset after PCA*

Note that PCA actually increased the number of features in the Breast Cancer Dataset which means it increased the dimensionality (9 features to 23) rather than reducing it.

However, applying PCA to both datasets reduces the distortion of K-Means when applied to both datasets which can be seen in figures 13 and 14. For the Breast Cancer dataset, the error eventually converges and decreases at similar rates with both the original and the reduced dataset. However, since there are more features or dimensions, the PCA must be performing better as it gives a similar error with more dimensions.

With the Optdigits dataset, the reduced dataset performs significantly better in terms of squared error. This is because the number of dimensions are actually reduced meaning the samples provide more information and allow for better clustering. Furthermore, since PCA maximizes the variance, the features can be separated into clusters more easily such that the instances are closer to the center of their cluster.

B. *Independent Components Analysis (ICA)*

Unlike PCA which maximizes reconstruction, Independent Components Analysis attempts to maximize independence of the features. This allows ICA to avoid the pitfalls of PCA and ensures that the features aren't simply being added together to form some Gaussian (Central Limit Theorem).

Figures 15 and 16 show the performance of clustering algorithms after Weka's ICA filter is applied

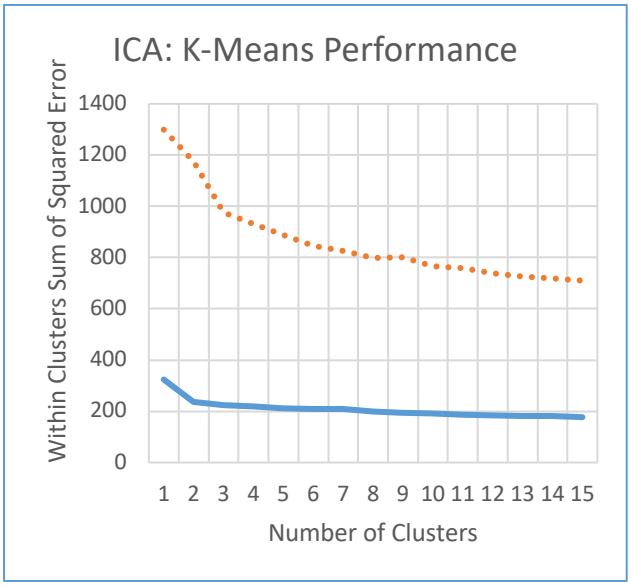


*Figure 15:K-Means Performance on Breast Cancer Dataset after ICA*

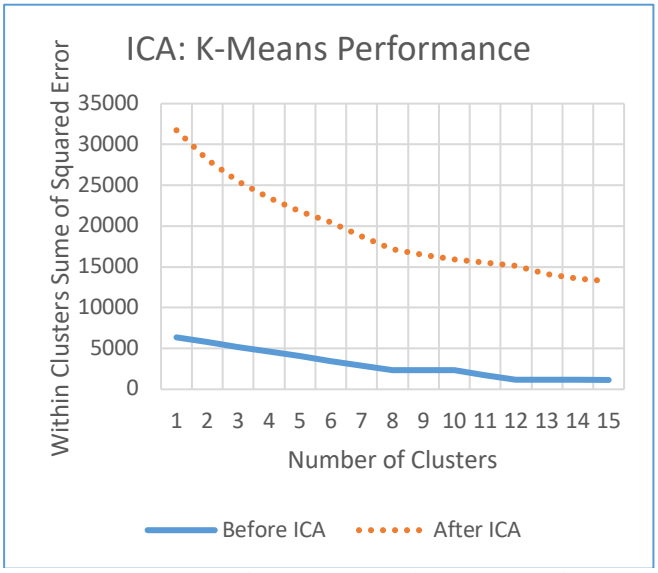


*Figure 16: K-Means Performance on Optdigits Dataset after ICA*

ICA clearly outperforms PCA. This is clear when observing the Breast Cancer dataset. The squared error is always remains significantly lower than before ICA is applied unlike when PCA is applied and the error converges as the number of clusters increase. This means that the attributes in the Breast Cancer Dataset help a clustering algorithm perform better when their independence is maximized as opposed to their reconstruction. The Optdigits dataset does not suffer this problem.

In order to analyze the effects of ICA on the new attributes, the kurtoses of each attribute was measured for both datasets. The results are shown below.

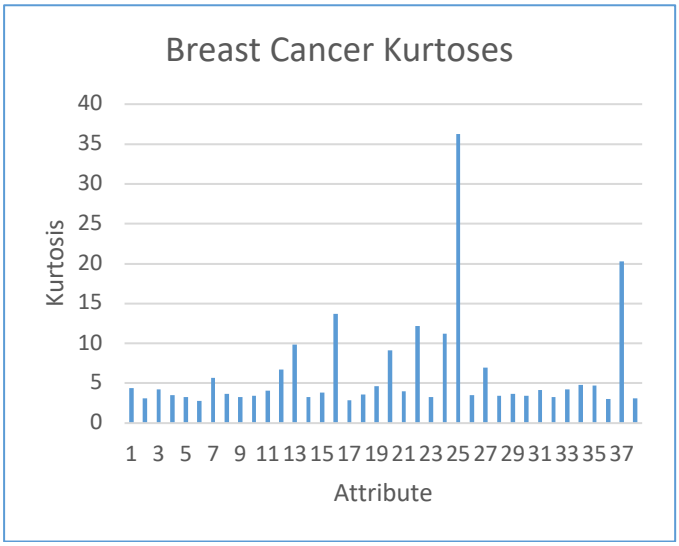


*Figure 17: Breast Cancer Kurtoses for all Attributes*

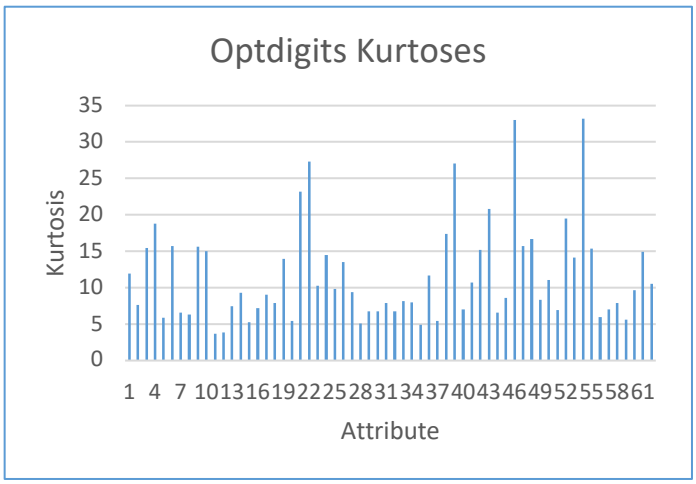


*Figure 18: Optdigits Kurtoses for all Attributes*

The ideal kurtosis value of a gaussian distribution is 3. The attributes for both datasets have kurtoses values that are significantly higher which indicates a high level of mutual independence in both datasets. This is evidence for the high performance of clustering algorithms after ICA is applied to both datasets. The Optdigits datasets have significantly higher variance and have values that are much greater than 3 which indicates that their attributes have a higher level of mutual independence than that of the breast cancer dataset.

*C. Randomized Projections (RP)*

Randomized Projections is a method that chooses a randomized matrix based on some target number of dimensions. This target is a hyperparameter so it must be tuned. In order to find the optimal value of this hyperparameter for both datasets, the number of dimensions was varied and the class to cluster error is measured for a set number of clusters. For this analysis, the number of clusters is equal to the number of classes in each dataset. Figure 19 shows the class to cluster error of each dataset after Randomized Projections is applied.

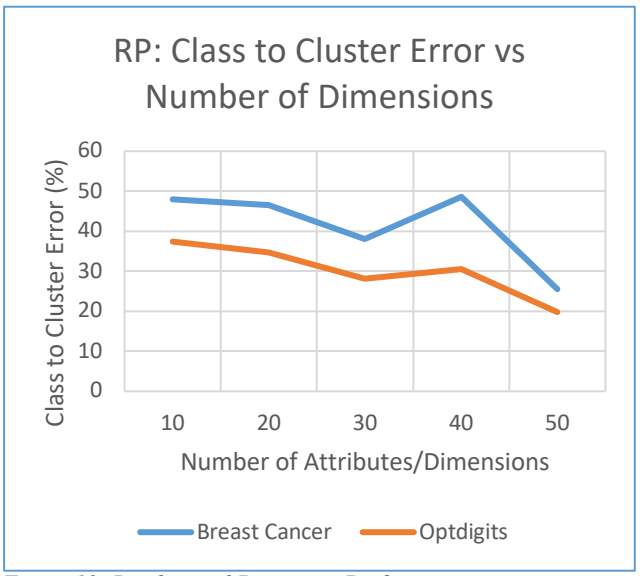


*Figure 19: Randomized Projection Performance*

Randomized Projection provides some interesting results. The optimal number of dimensions is found to be 50 for both dimensions. The class to cluster error curve seems to decrease as the number of dimensions increases. Jumps in the error can be attributed to the stochastic nature of the feature transformation algorithm, but it appears to decrease. One can surmise that as the number of dimensions increases, the clustering algorithms will eventually overfit. Especially since the Breast Cancer dataset is increasing greatly in the number of features considered as the target dimension increases. It is surprising that the clustering error decreases as the dimension increases so much, but if clustering can isolate classes better with the increasing number of dimensions, then the supervised learning algorithms will most likely perform better with an increased number of dimensions. The Optdigits also displays a similar trend, though the jumps in error are not as sharp as the other dataset.

*D. Information Gain*

Information Gain is a feature selection algorithm. Feature selection is a subset of feature transformation, where the transformation is simply returning a subset of the given features. It uses entropy and mutual information to isolate the best features. In order to determine the optimal number of features to keep, this parameter was tuned by removing certain features from both datasets and comparing the performance of K-Means on both modified datasets.

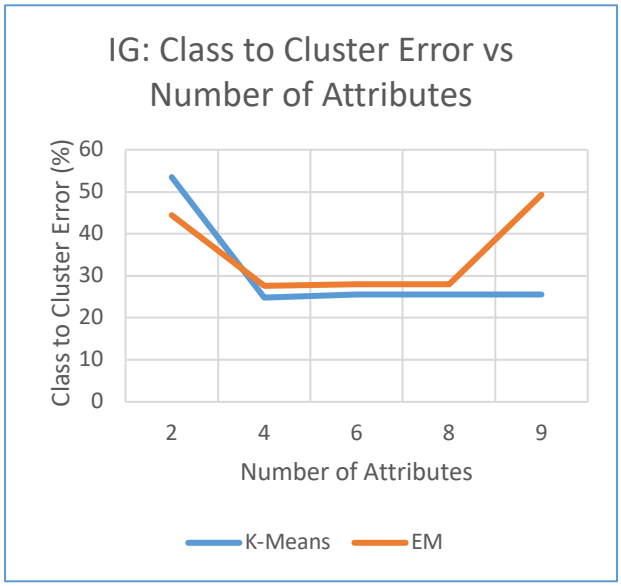


*Figure 20: Clustering Performance using Information Gain Feature Selection on Breast Cancer Dataset*

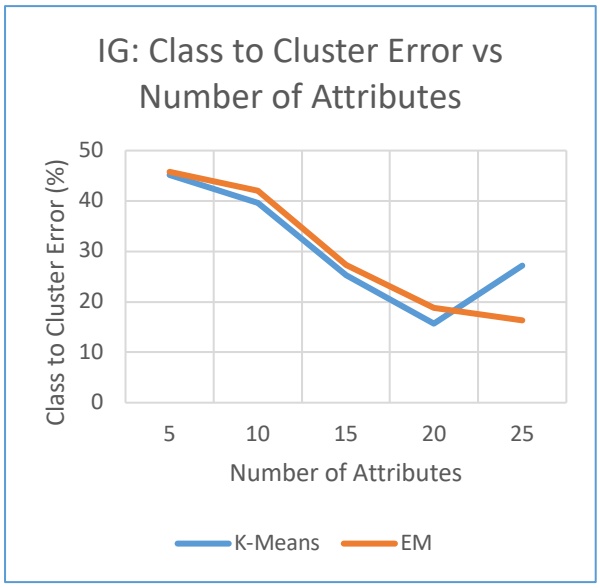


*Figure 21: Clustering Performance using Information Gain Feature Selection on Optdigits Dataset*

Both class to cluster error curves indicate signs of overfitting when too many attributes are added which indicates the presence of attributes that are irrelevant or extraneous. In fact EM performs poorly when all the attributes of the Breast Cancer dataset are considered. The Information Gain selection algorithm helps visualize the attributes that are most relevant and give the most information by pruning all of the attributes that provide the least information.

## III. Revisitng Neural Networks

Now it is time to analyze the effects of dimensionality reduction and clustering on supervised learning problems. In order to observe the effects of dimensionality reduction, Weka's MultilayerPerceptron Network was trained and run on both datasets on the new sets of features. Table 1 shows the performance of the neural net learner after all the algorithms are applied.

The neural net was applied on the Optdigits dataset with optimal parameters found in previous analysis. The learning rate is set to 0.4 and the neural net is trained for 600 training iterations. The dataset was split into a 70% training set and a 30% testing set.

For RP and Information Gain, the number of attributes hyperparameter was tuned and the optimal parameter that generated the least error was used. The values of the parameters are as follows: 50 attributes for RP and 20 attributes for InfoGain.

*Table 1: Neural Network Performance after Dimensionality Reduction Algorithms Applied*

| Algorithm | Accuracy (Optdigits) |
|---|---|
| None | 97.92% |
| PCA | 96.8% |
| ICA | 97.56% |
| RP | 97.45% |
| Information Gain | 92.64% |

The original dataset performs the best with the neural network. ICA and RP are not too far behind. What this shows is that the by transforming the features of the Optdigits dataset, the burden of the curse of dimensionality is reduced with minimal loss to performance. Furthermore, the attributes of the dataset help a supervised learner perform better when the independence between each feature is maximized as opposed to maximizing reconstruction. This is apparent as ICA performs significantly better than PCA.

Information Gain performs the worst out of all the datasets which reveals the difference between relevant and useful features. Although many features were thrown away because they gave very little information, they are still useful as having them helped the learner classify instances better.

The next step is to consider clustering's role in supervised learning. Clustering algorithms were run on the dataset and each cluster was considered as a new feature.

First, the original features were removed from consideration and only the clustered features were used to train the learner. The following table shows the results of the learner given up to 5 clustering features.

*Table 2: Accuracy of a Neural Network Learner Applied only to Cluster Attributes*

| Cluster Attributes | Accuracy | Training Time (s) |
|---|---|---|
| 1 | 73.31% | 17.3 |
| 2 | 86.95% | 35.43 |
| 3 | 87.18% | 57.53 |
| 4 | 93.89% | 89.72 |
| 5 | 94.31% | 137.88 |

As expected the training time of the learner increases as the number of attributes considered increases since the neural network has more attributes to tune its weights for. The accuracy increases as the number of features increases. This is intuitive as the more information the neural network has, the better it can learn. Also having only up to 5 features will not cause it to overfit. At some point adding more cluster attributes will cause the accuracy to decrease since the learner will overfit to the attributes given.

Next, each clustering feature was added to the dataset and the neural net was run against all the features. K-Means was applied to the Optdigits dataset starting at 10 clusters and increasing in order to add five features total to the dataset. The result of each addition is listed below in Table 3.

*Table 3: Accuracy of a Neural Network Learner After Clustering Attributes are added*

| Cluster Attributes Added | Accuracy |
|---|---|
| 1 | 97.74% |
| 2 | 98.16% |
| 3 | 97.69% |
| 4 | 98.51% |
| 5 | 98.39% |

These results show that adding more cluster attributes can cause the learner to overfit and the performance begins to decrease. This can also be seen as an effect of the curse of dimensionality as adding more dimension tells the learner a little less about the class of each instance, meaning that given the same number of instances, the learner performs worse with more attributes and needs more samples in order to make up for the performance loss. Still, adding more clustering attributes helps the learner perform better than the original dataset as seen in Table 1 which shows that providing some clustering attributes adds useful features that can help a supervised learner classify instances of datasets better. This is one possible use for the clustering algorithms analyzed previously. It also demonstrates unsupervised learning's purpose as a method of data description, since adding features to describe a dataset is a way of describing the actual data in the dataset.

## IV. Conclusion

The intention of this analysis was to compare clustering algorithms to supervised learning algorithms and to determine the effects of dimensionality reduction algorithms on both clustering algorithms and supervised learners. In this analysis, we have seen that clustering algorithms can be compared to supervised learners, but they perform poorly as supervised learning has the advantage of taking learner bias into account during training. Wrapping would likely be a better comparison to supervised learning, but it would take significantly more time to run wrapping on datasets. We have also observed dimensionality reduction and how it attempts to reduce the number of features while also transforming the original features into more relevant and useful features using linear transformations and have run supervised learners on a reduced dataset to see the impact of these feature selection and feature transformation algorithms.

## References


[1] Mitchell, T. M. (1997). Machine Learning (1st ed., Graw-Hill Series in Computer Science). McGraw-Hill education

[2] "Intro to Machine Learning Course | Udacity." Course | Udacity. Udacity, n.d. Web November 2017

[3] T. Kodinariya and P. Makwana, "Review on determining number of Cluster in K-Means Clustering", *International Journal of Advance*, 2013

[4] T. K. Moon, "The Expectation Maximization Algorithm", *IEEE Signal Processing Magazine*, 1996

[5] F. Dallaert, "Expectation Maximization Algorithm", *College of Computing Technical Report*, 2002

[6] J. Shlens, "A Tutorial On Principal Component Analysis", *Princeton University*, 2003

[7] A. Hyvärinen and E. Oja, "Independent Component Analysis: Algorithms and Applications", *Neural Networks Research Center*, 2000