Unsupervised Learning and Dimensionality Reduction

CS 7641 Machine Learning
Sang Yun Park (spark359@gatech.edu)

Section 1: Introduction

Abstract

This assignment includes analysis on two clustering algorithms and four dimensionality reduction algorithms on two datasets.

1) Breast Cancer: 286 instances and 9 attributes
2) Optical Recognition of Handwritten Digits: 5620 instances and 64 attributes

The clustering algorithms include k-Means Clustering and Expectation Maximization, and the dimensionality reduction algorithms include Principal Component Analysis (PCA), Independent Principal Component Analysis (ICA), Randomized Projection (RP), and Information Gain (IG) as feature selection.

Section 2: Clustering

K-Means Clustering

K-Means Clustering is a method that separates the data into k clusters. It randomly picks *k* number of center points, and each center claims its closest points with minimal sum of squared errors. It repeats the step of picking a new point and recomputing a center point until it converges.

On two datasets, K-Means algorithm is applied by increasing the number of *k*. Figure 1 shows the result of the test. As the number of cluster increases, a decrease in sum of squared error (SSE) is observed. Euclidean distance is used when calculating SSE. An optimal number of cluster can be observed using the elbow method. It observes a point where a rate of reduction in SSE is getting slower. On two datasets, we can clearly observe 'elbow' points as marked on Figure 1.

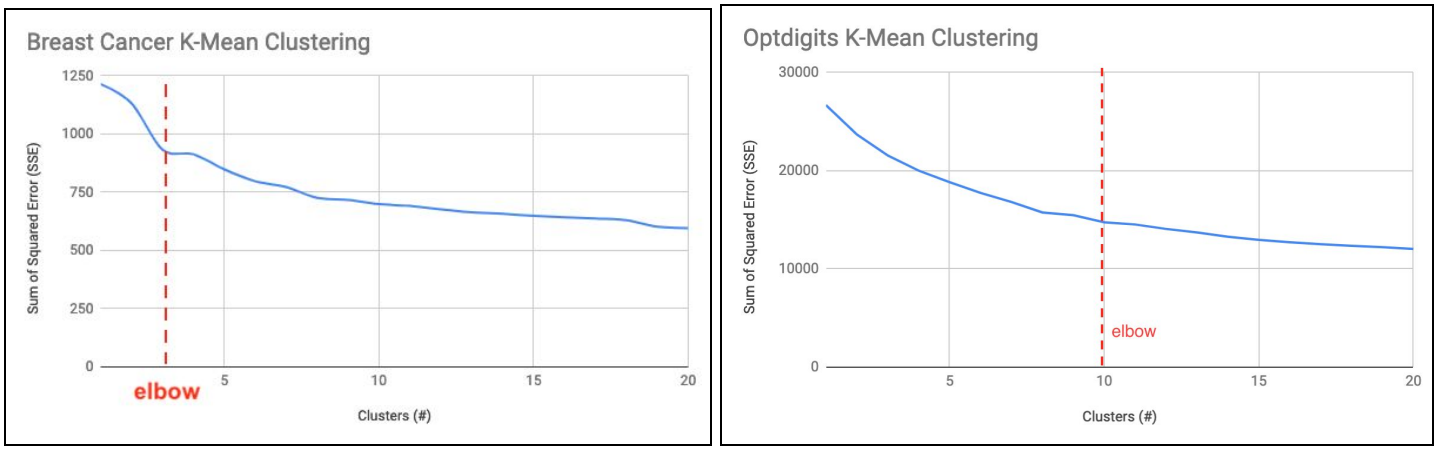

Figure 1. K-Means Clustering on Breast Cancer Data and Optdigits Dataset

For an ambiguous result, we can calculate k value using the average Silhouette method. The method measures the quality of a clustering by determining how close each dataset lies in its belonging cluster. Higher the average silhouette value, better a cluster. I also measure class to cluster error. It is defined as the number of instances with incorrectly defined cluster. Figure 2 shows the result of the evaluation. The error rate decreases until the total number of classes in each dataset and starts to increase afterward.

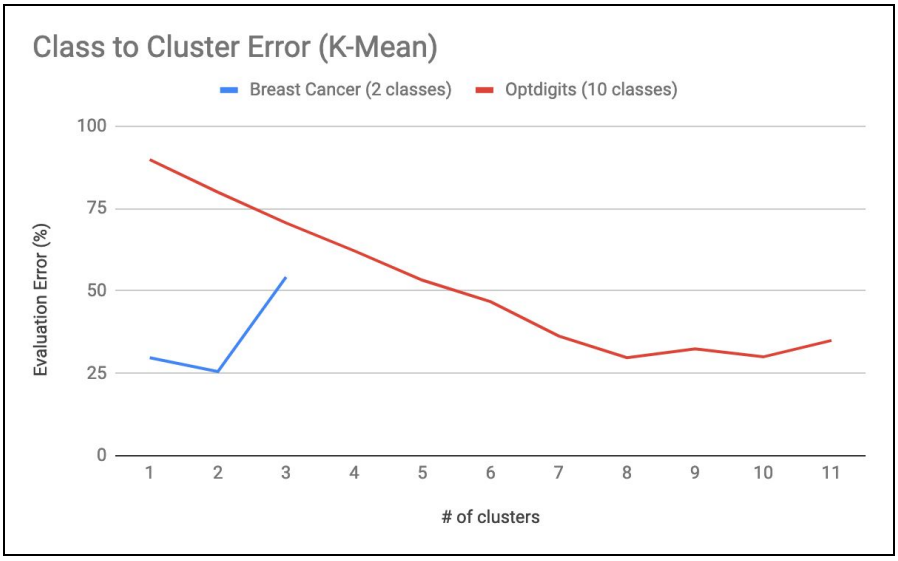

Figure 2. Class to Cluster Error for K-Mean Clustering

Expectation Maximization

This clustering algorithm is a softer clustering method compared to k-means clustering. In k-means clustering, if a data point is close to two different clusters, it is only bounded to the closest cluster and has no effect on another cluster regardless of distance to the second closest cluster. However in Expectation Maximization (EM) method, each data point has probability based on its likelihood of getting into one cluster. The algorithm alternate between probabilities of expectation and maximization. In the experiment, I alter the number of clusters and observe the changes in likelihood.

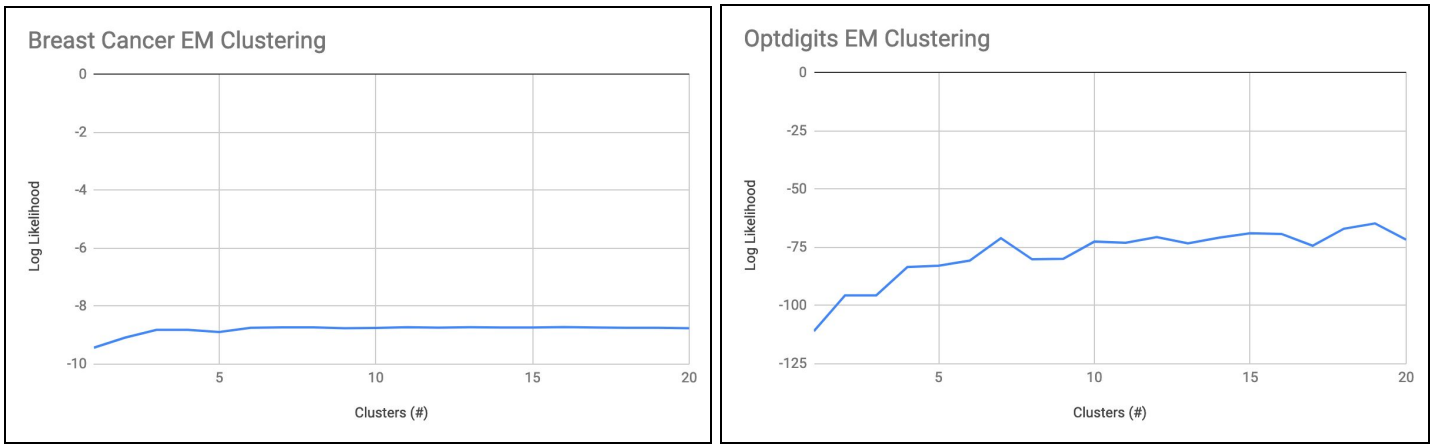

Figure 3. Expectation Maximization Clustering on Each Dataset

On the breast cancer dataset, it is observed that the log likelihood converges in early phase. One explanation for such result can be a vulnerability of Expectation Maximization to stucking at local maxima. In contrast, log likelihood increase proportional to the number of clusters in the optical digits dataset. Figure 4 shows the visual representation of clusterings with optimal k value found above. As noted, data points of the breast cancer dataset (left) are more clustered compared to the optical digits dataset (right). This implies that the data points in the optical digit dataset are more influenced by Expectation Maximization algorithm.

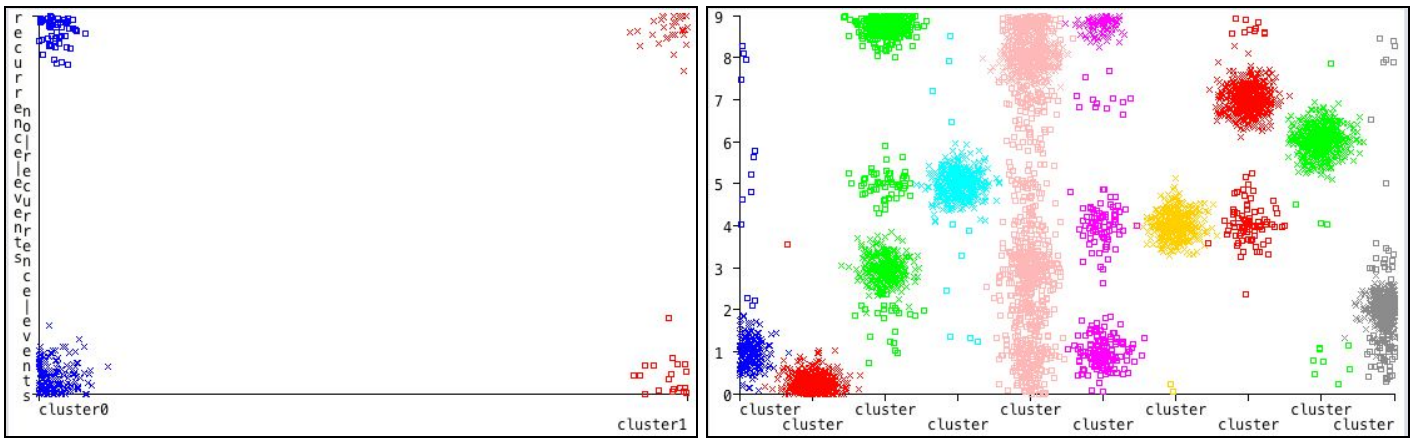

Figure 4. K-Mean Clustering Visual Representation on Both Datasets

As shown in Figure 5, class to cluster error looks similar to k-mean clustering result as expected. Breast cancer is binary data, and therefore having more than 2 classes results in increase of error rate. Also, in the optical digits dataset, the error rate starts to increase when the number of clusters exceeds the total number of classes.

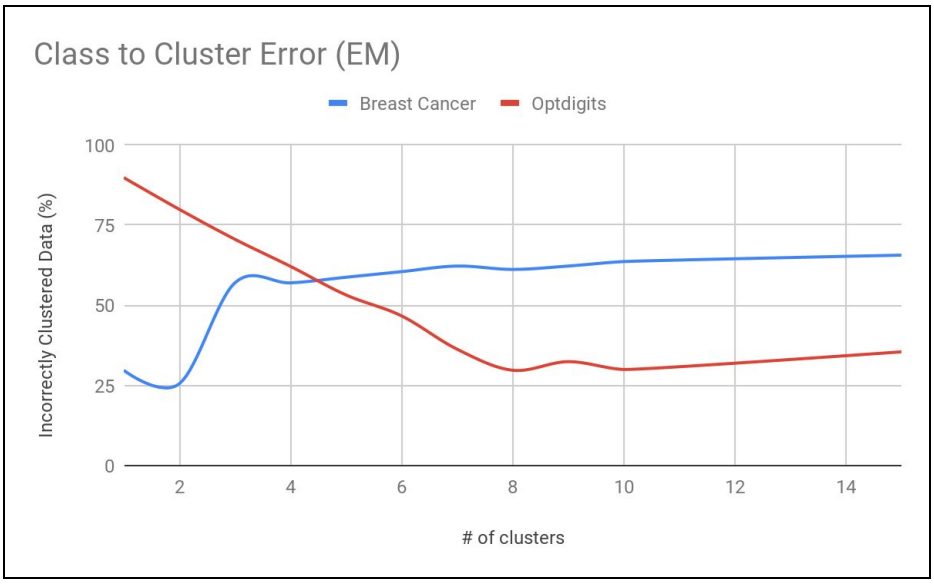

Figure 5. Class to Cluster Error for EM Clustering

Section 3: Dimensionality Reduction

PCA

By using Principal Components Analysis (PCA), we can transform the features of dataset into dimensions by having feature elimination and feature extraction. It finds orthogonal eigenvectors that maximizes the variance of data. I use Weka's library to implement PCA with default parameters. The results are as follows:

Dataset	Default Number of Attributes	Number of Features after PCA
Breast Cancer	9	28
Optical Digits	64	42

Table 1. Dimensionality Change After PCA

The dimensionality for the breast cancer dataset is increased while reduced for the optical digits data. As noted from Figure 6, sum of squared errors on both datasets are significantly lowered after applying PCA. This implies that there exists more advantages for the breast cancer dataset when having more attributes. In contrast, the number of attributes decreased for the optical digits dataset. This feature extraction gives simplicity while maintaining interpretability of variables.

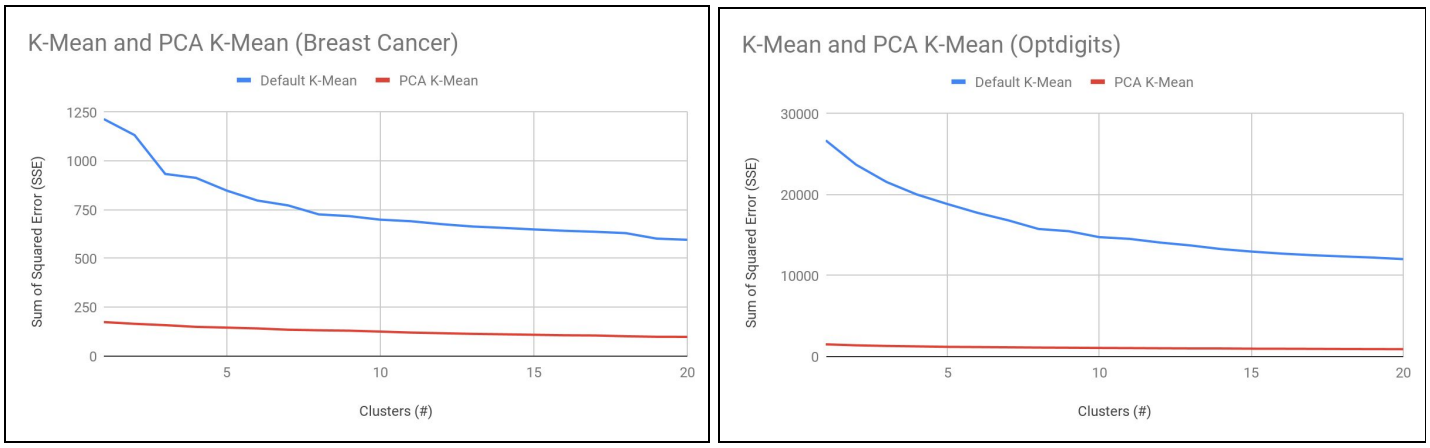

Figure 6. Performance Comparison of K-Mean Clustering after PCA Applied

ICA

Independent Component Analysis (ICA) is a similar feature transformation method as PCA. PCA's goal is to maximize the variance of data. In contrast, ICA tries to maximize the independence of the feature. I use Weka's IndependentComponents filter with default parameters. The results are as follows:

Dataset	Default Number of Attributes	Number of Attributes after PCA applied
Breast Cancer	9	38

Optical Digits	64	62

Table 2. Dimensionality Change After ICA

As shown in Table 2, dimensionality increased for the breast cancer data and slightly reduced for the optical digits data. It is also noted that sum of squared error in both datasets look very similar to PCA results. The performance of clustering algorithm increased when applied ICA on both datasets. The squared error is significantly lowered, and log likelihood increased as well. Also, Figure 7 shows that the convergence of the squared error occurs in relatively small number of clusters on both cases. This implies that transformed data points are less spread out, therefore creates lower sum of squared errors and higher log likelihood as shown in Figure 7 and 8.

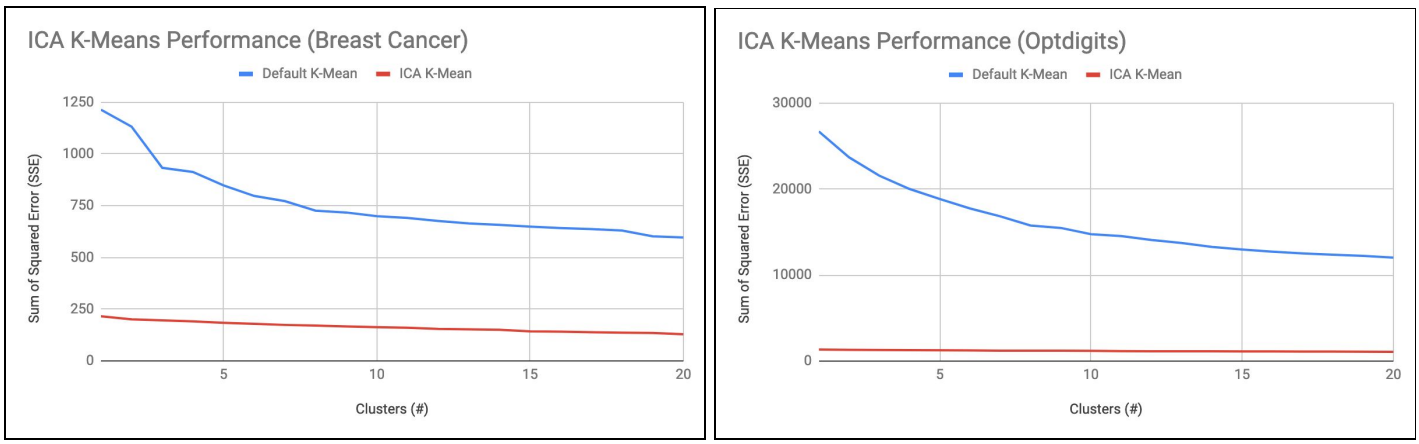

Figure 7. Performance Comparison of K-Means Clustering after ICA (Both Datasets)

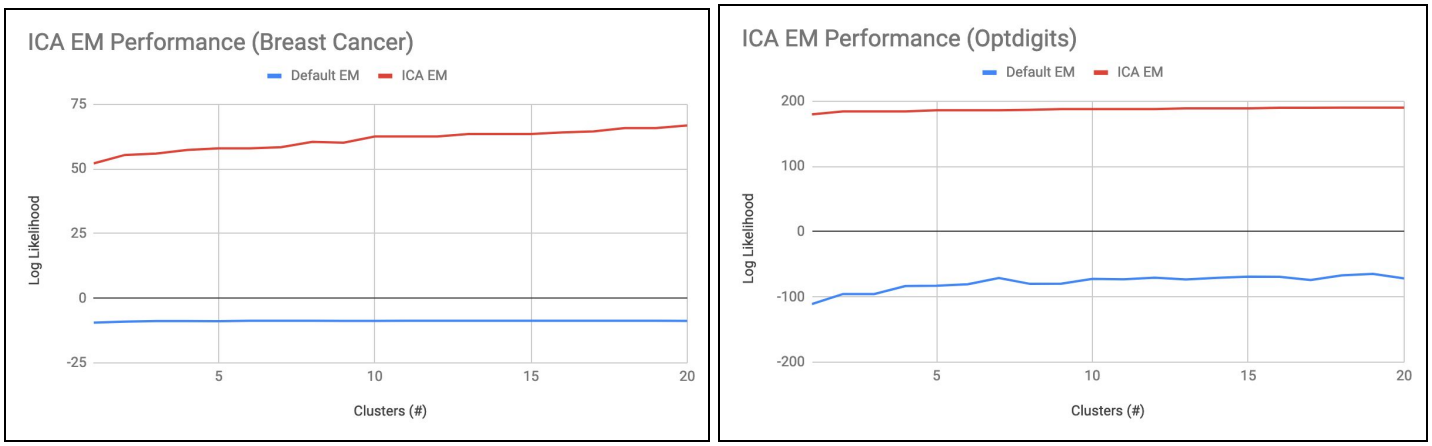

Figure 8. Performance Comparison of EM Clustering after ICA (Both Datasets)

Randomized Projection

Randomized Projection is a method that generates a randomized matrix based on a target number of dimension, and it transforms the feature with created matrix. In this experiment, we find an optimal number of attributes by measuring the class to cluster error in different number of dimensions. The number of cluster (k) used for this experiment is same the optimal k values found above: 2 for the breast cancer and 10 for the optical digits dataset. Weka's RandomProjection is used with changes in the number of attributes.

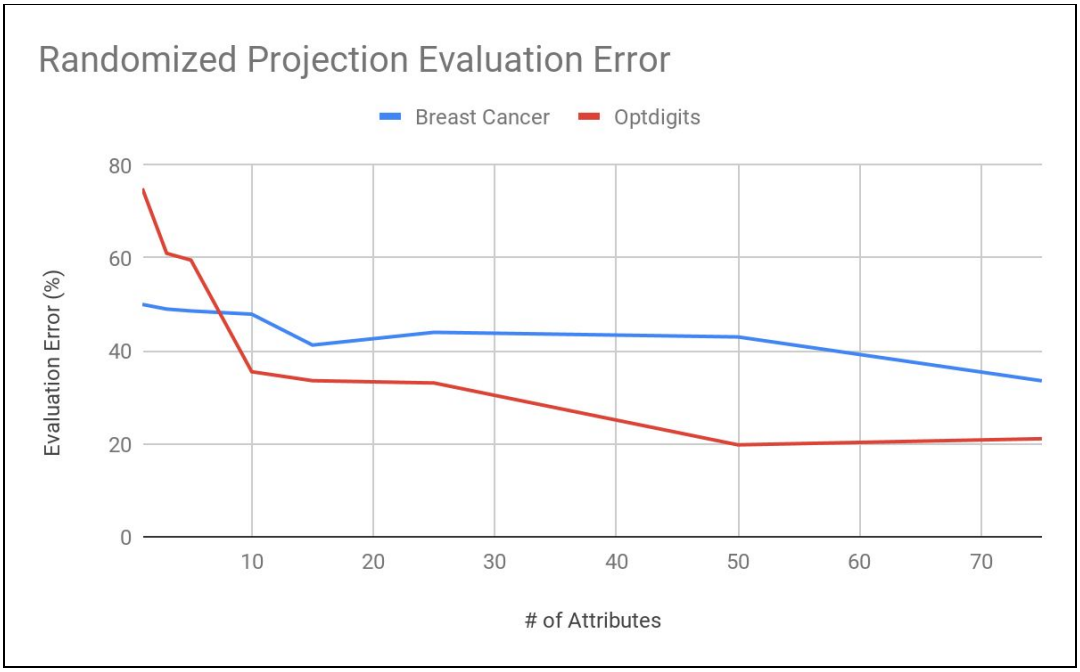

Figure 9. Randomized Project Evaluation Error on Two Clustering Algorithms (Both Datasets)

As shown on Figure 9, both datasets converges on the number of dimensions = 50. It looks like the performance increases as the number of dimensions increases. However, this should not be concluded, since the method is based on stochastic behavior. It is also noticeable that steeper increase in performance occurs in early phase of the experiment. Initial class to cluster error was **25%** for the breast cancer and **29%** for the optical digit with an optimal *k* as shown on Figure 2. Having compared to such result, performance increase by applying randomized projection is not significant.

Information Gain

Information Gain (IG) is a feature selection method that utilize specific number of attributes with given information scores. This dimension reduction algorithm uses entropy and mutual information to pick the best attributes. For each dataset, attribute is getting removed from the lowest rank of information. Weka's InfoGainAttributeEvaluator is used to rank attributes on each dataset.

On both dataset, the performance increases by removing some attributes with lower rank until some points as shown on Figure 9. However, the performance starts to fall when too many attributes are removed (marked with purple dashed square on FIgure 9). For the breast cancer data, there is slight performance increase, whereas noticeable increase is observed in the optical digit dataset (29% → 20% reduction in error rate). This is most likely due to the characteristic of the dataset. By removing some attributes that brings ambiguity on clustering, we can better cluster the dataset.

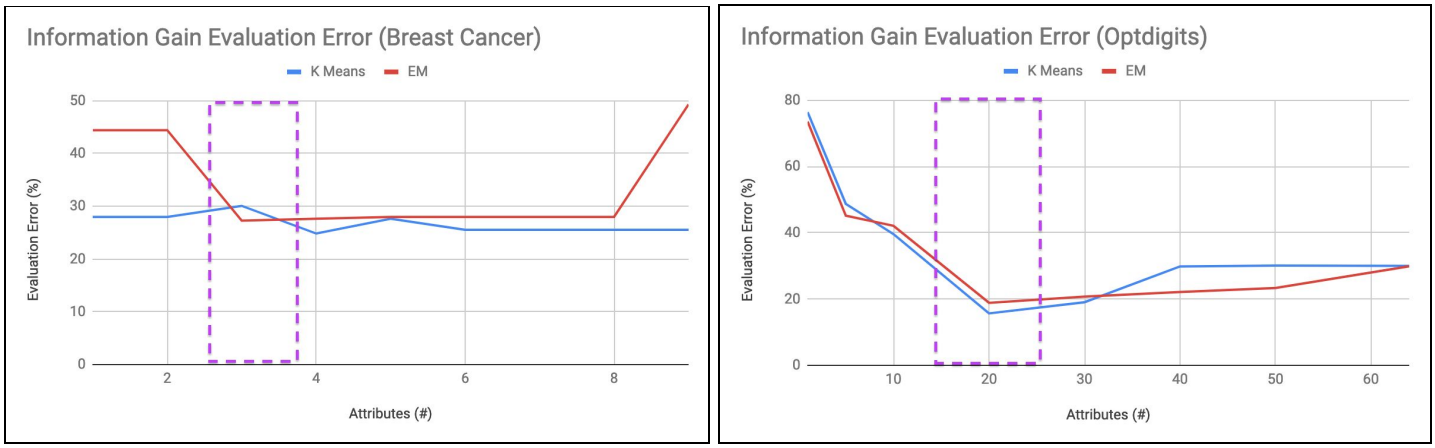

Figure 9. Information Gain Evaluation Error (Both Datasets)

Section 4: Neural Network

Neural Net with Dimensionality Reduction

We test neural network with the breast cancer dataset (9 attributes). Measured accuracies on datasets with dimension reduction applied are shown on Table 2. I apply an optimal hyperparameter for Randomized Projection and Information Gain that is derived from the previous section. Neural network is implemented using Weka's MultilayerPerceptron filter with default parameters.

Algorithm	K-Mean Accuracy (%)	EM Accuracy (%)
Original Dataset	66.7832	66.7832
PCA	66.0839	66.0839
ICA	65.7343	66.0839
Randomized Projection (50 attributes)	67.4825	67.4825
Information Gain (4 attributes)	71.3287	71.3287

Table 2. Neural Network with Dimensionality Reduction Applied

As shown on Table 2, there is slight decrease in performance in ICA dataset while PCA dataset is barely different from an original dataset. This represent that neural network performs better, for this dataset, with variance maximization than independence maximization on data feature. Also, it is noted that Information Gain dataset has increased the performance of 5%. Having removed 5 attributes out of 9 from the dataset, the neural network has better performance. We may conclude that the classification highly relies on the highest 4 attributes in the breast cancer dataset.

Neural Net with Clustering Feature

Finally, I apply clustering algorithms on the dataset with cluster is considered as a attribute. Weka's AddCluster function is used to add cluster. We observe that adding cluster does not increase the performance in both

clustering algorithms. This implies that adding clustering attribute can possibly decrease the performance by overfitting.

Cluster Attributes Added	K-Mean Accuracy (%)	EM Accuracy (%)
0	66.7832	66.7832
2	65.7343	66.7832
5	66.0839	69.5804
10	66.4336	66.0839
25	67.1329	66.7832

Table 3. Neural Network with Clustering Feature Included

Conclusion

In this assignment, we examined clustering algorithms with different dimensionality reduction methods. We observed pros and cons of each method by analyzing the performance difference. We also analyzed how supervised learner is affected by feature transformation. While other dimensionality reduction methods beside Information Gain has negligible effect in applied dataset, that does not necessarily mean you will observe the same in other dataset. The effect and outcome of each method can vary depending on characteristics of dataset.