Wenjun Wu
April 2, 2017
CS 4641
Assignment 3

Analysis of Unsupervised Learning

Introduction

This report explores different unsupervised learning techniques. Unsupervised learning is an algorithm used to draw inferences from datasets without classification labels. The most common unsupervised learning technique is clustering algorithms and two clustering algorithms will be analyzed in this report: k-means (KM) clustering and Expectation Maximization (EM). KM clustering partitions data points into k clusters based on the centroid of the cluster. Expectation maximization (EM) is similar to KM clustering, but EM differs in that it determines cluster assignment based on prior probability distributions.

Besides, four dimensionality reduction (DR) algorithms are also investigated in this project: Principle component analysis (PCA), Independent component analysis (ICA), Randomized Projections (RP) and Feature selection based on Information Gain (IG).

Experiments with techniques discussed above will be performed on two datasets, which were also used in assignment 1. The result will be discussed and the following topic will be explored in this report: 1) the efficacy and characteristics of clustering algorithms, 2) the influence of DR on clustering algorithms, 3) the influence of DR on Neural Nets, which is a supervised learning algorithm.

Datasets Description

Wilt Dataset

This dataset was extracted from Quickbird imagery that detects diseased trees. The task is to detect diseased and non-diseased trees from those multispectral image attributes. This dataset in very interesting. Successful detection of the conditions of trees and surrounding areas through remote sensors could significantly reduce the cost to maintain forests and prevent spreading of disease and harmful insects. It is also interesting because all the attributes are continuous were derived from sensors. Since the data is derived from real-world images, it should contain certain level of noise.

Car Evaluation Dataset

This dataset was derived from a simple hierarchical decision model for decision making. The model evaluates cars according to a concept structure considering 6 attributes: the overall price, maintenance price, number of doors, capacity, the size of luggage boot and estimated safety of the car. In contrast to Wilt dataset, this dataset is smaller (1728 instances) and is consisted of categorical attributes with nominal values. For ease of data preprocessing, the nominal values are converted to numerical values.

	Wilt	Car Evaluation
Number of Attributes	6	6
Number of Instances	4889	1728
Number of classes	2	4
Attribute Types	Numerical, continuous	Categorical, nominal

Table 1. *Information of Datasets*

Comparison between experiments performed on Car Evaluation dataset and Wilt dataset can demonstrate how algorithms behave differently on different types of dataset (Table 1). Also, the difference in size can also demonstrate how size can influence the performance of algorithms.

Methods

Clustering Analysis

KM and EM were applied to the two datasets described above. For both clustering algorithms, different number of clusters were tested to determine the optimal number of clusters for each dataset. For Wilt dataset, the number of clusters tested are 2, 4 and 6 and for Car Evaluation dataset, the number of clusters tested ranges from 4-20.

Besides, I also examined different selection of distance function for KM: Euclidean distance (ED) and Manhattan distance (MD). Euclidean distance is the shortest distance between observation and cluster centroid in the hyperplane. Manhattan distance is absolute differences between coordinates of observation and cluster centroid. For both EM and KM, initial seed = {100, 200}, was also examined.

Dimensionality Reduction (DR)

PCA, ICA, RP and IG feature selection are used to reduce the dimensionality of the datasets. For PCA, different total variances covered were tested: 0.75, 0.95 and 1. For RP and IG, different number of attributes retained were tested (Table 2).

	PCA	ICA	RP	IG
Parameter varied	Total variance	NA	Number of attributes attained	Number of attributes attained
Parameter value	0.75, 0.95, 1	NA	Wilt: {1,2,4,8} Car Evaluation: {1,3,5,9}	Wilt: {1,2,4} Car Evaluation: {1,3,5}

Table 2. *Parameters tested in DR algorithms*

Clustering Analysis After DR

EM and KM were applied to dimensionality reduced datasets with different number of clusters (Table 3) to determine the influence of DR on clustering analysis.

	Car Evaluation Dataset	Wilt Dataset
Number of Clusters	4, 8, 10, 12, 15, 20	2, 4, 6

Table 3. *Number of clusters tested in clustering analysis after DR*

Supervised Learning After DR

In order to determine the influence of DR on supervised learning, NN is applied to dimensionality reduced datasets. Cross validation result will be compared to the result of clustering analysis. Finally, EM and KM are used as DR algorithms and NN was rerun on the newly projected data by EM and KM. The parameters used in the last steps will be the best performing parameters found in previous steps.

Results

	Decision Tree		Adaboost		SVM		Neural Network		kNN	
	Accuracy	Time (s)	Accuracy	Time (s)	Accuracy	Time (s)	Accuracy	Time (s)	Accuracy	Time (s)
Wilt	98.90%	0.01	98.50%	0.1	91.90%	0.23	98.62%	5.2	93.70%	0.01
Car Evaluation	91.60%	0.01	95.40%	0.21	92.60%	0.02	97.30%	8.6	90.30%	0.01

Table 4. *Summary of Performance of Supervised Algorithms*

Clustering Analysis

1. Car Evaluation Dataset

The performance of clustering analysis of EM and KM on Car Evaluation dataset with different parameters are shown in Figure 2-4. As can be seen from Figure 2, 4 clusters have the best performance on the Car Evaluation dataset with both EM and KM regardless of number of seeds and distance function. Also, the results indicate that both EM and KM performs poorly comparing with supervised algorithms (Table 4) and has error rate as high as 60%, but EM performs better than KM. The reason that both clustering algorithms have poor performance is the nature of the dataset, which can be seen from Figure 1. When plotting one attribute versus the other, there is no

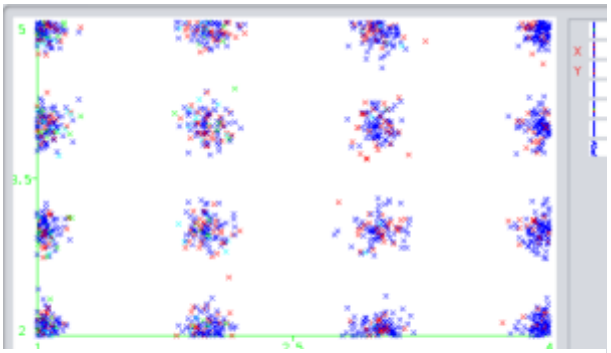


Figure 1 Class distribution of Car Evaluation dataset with horizontal axis as attribute 1 and vertical axis as attribute 2. Plotting different attributes yielded similar graph.

distinction between different classes. Also, due to the nature of nominal values, data are in clusters already. However, observations with same classification does not share similar feature values. Result from assignment 1 also supports this reasoning (Table 4): kNN has the lowest performance among all supervised learning. The reason that KM performs worse than EM is that KM algorithm predicts the classification using distance proximity, but distance proximity is not a meaningful criterion for this dataset as shown by Figure 1. EM performs better than KM because feature values have less significance in EM and similarities between feature values, which are misleading in this dataset are often omitted.

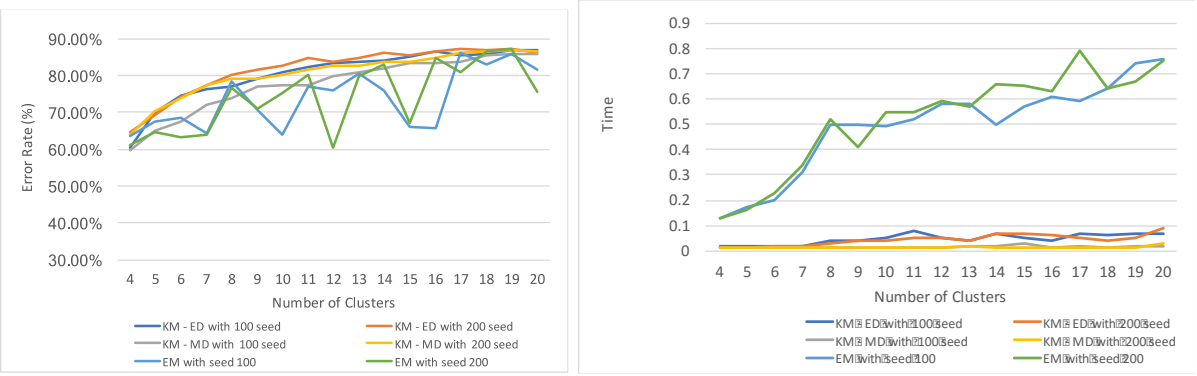


Figure 2 Clustering Error Rate (Left) and Clustering Time (Right) on Car Evaluation dataset with K-Means (KM) and Expectation Maximization (EM) Clustering analysis. Manhattan Distance (MD) and Euclidean Distance (ED) were tested for KM. EM and KM were tested with different number of seeds = {100, 200}.

Another important observation is that as the number of clusters increases, the sum of squared error for KM generally decreases, but the error rate of KM increases (Figure 2, 3). This is expected because as the number of clusters increases, the cluster becomes smaller, but more clusters still would not improve the model due to the nature of the dataset. Also, increasing the number of seed does not have a clear influence on the performance of clustering algorithms for this dataset (Figure 3). Moreover, the clustering time of both algorithms are less than a second and thus considered unimportant when comparing the two algorithms although EM requires longer time than KM (Figure 2 Right).

2. Wilt Dataset

The performance of clustering analysis of EM and KM on Car Evaluation dataset with different parameters are shown in Figure 5-7. EM performs much better than KM. Increasing the number of seed does not increase the performance of either EM or KM. One thing to note is that Wilt dataset only has small portion of positive classes (Figure 4 Right), which could be the reason that EM clustering works dramatically better than KM since EM takes prior probability into account. Although the classification result of EM is still not comparable with supervised algorithms (Table 4), component analysis could increase the performance of EM to be comparable with supervised learning since only two attribute can separate the data effectively (Figure 4 Right).

Similar pattern observed on the Car Evaluation dataset can be seen here: as the number of clusters increases, the performance of both algorithms decreases, but sum of squared error and log of likelihood increases. Regarding time spent by clustering algorithms, as the number of clusters increases, clustering time also increases (Table 5). The clustering time of EM is much longer than KM and the increase rate is much higher (Table 5).

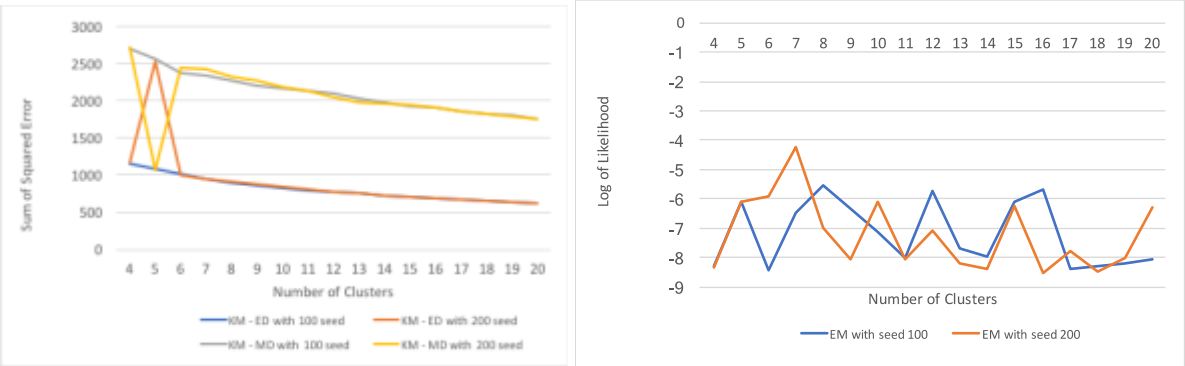


Figure 3 Sum of Squared Error and Log of likelihood on Car Evaluation dataset with Left: K-Means (KM) and Right: Expectation Maximization (EM) Clustering analysis respectively. Manhattan Distance (MD) and Euclidean Distance (ED) were tested for KM. EM and KM were tested with different number of seeds = {100, 200}.

Time spent by Clustering Algorithms (second)	Number of Clusters		
	2	4	6
KM - ED with 100 seed	0.02	0.05	0.08
KM - ED with 200 seed	0.02	0.05	0.07
KM - MD with 100 seed	0.04	0.12	0.07
KM - MD with 200 seed	0.05	0.06	0.13
EM with seed 100	0.17	1.24	2.25
EM with seed 200	0.17	1.17	2.22

Table 5. *Clustering time in seconds of K-Means (KM) and Expectation Maximization (EM) Clustering analysis on Car Evaluation dataset. Manhattan Distance (MD) and Euclidean Distance (ED) were tested for KM. EM and KM were tested with different number of seeds = {100, 200}.*

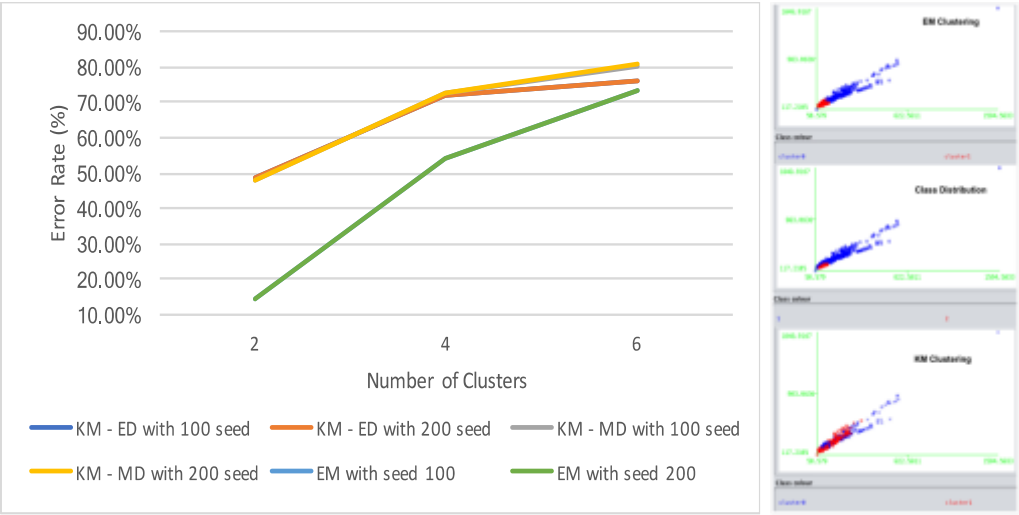


Figure 4 Clustering Result on Wilt dataset with K-Means (KM) and Expectation Maximization (EM) Clustering analysis. Left: Clustering Error Rate versus number of clusters; Right: EM and KM cluster visualization of attribute 3 projection onto attribute 2. Manhattan Distance (MD) and Euclidean Distance (ED) were tested for KM. EM and KM were tested with different number of seeds = {100, 200}.

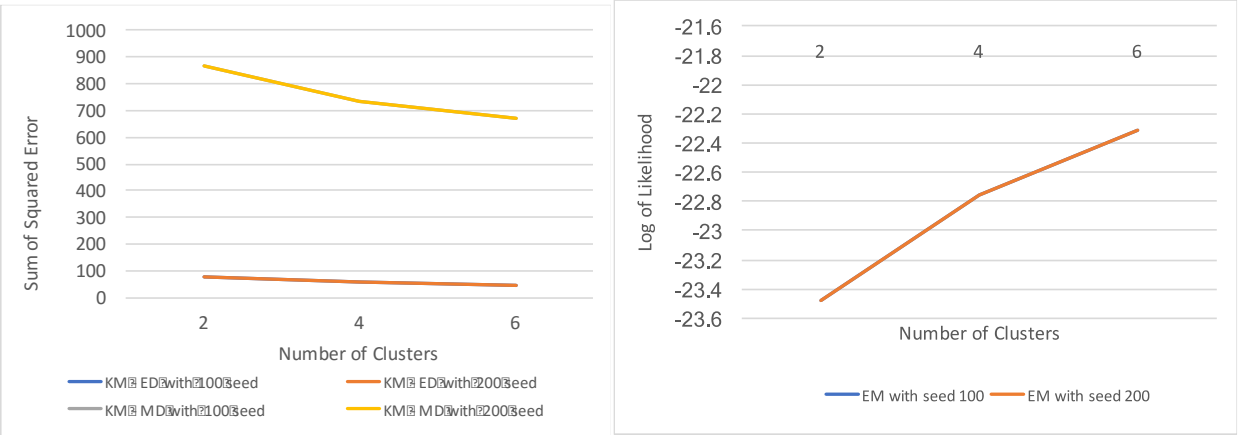


Figure 5 Sum of Squared Error and Log of likelihood on Wilt dataset with Left: K-Means (KM) and Right: Expectation Maximization (EM) Clustering analysis respectively. Manhattan Distance (MD) and Euclidean Distance (ED) were tested for KM. EM and KM were tested with different number of seeds = {100, 200}.

Component Analysis

1. PCA

PCA uses orthogonal transformation which maximizes the total variance, to convert the attributes into linearly independent variables. In general, PCA is inappropriate for discrete data type, which is the case for Car Evaluation dataset since a practically important violation of the normality assumption underlying the PCA occurs when the data are discrete. This is confirmed by results shown in Figure 6 to 7, which shows that all attributes have similar variances and projection onto new space does not change the distribution of data (Figure 1 and Figure 6). Also, the correlation matrix indicates that attributes are linearly independent and each of them has the same eigenvalue of 1 (Figure 7).

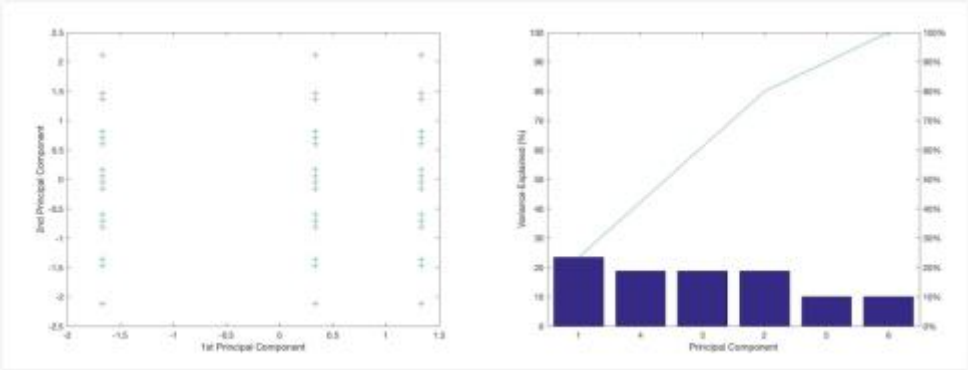


Figure 6 Projection of top 2 Principal Components (Left) and the Percent Variability Explained by each Principal Component (Right) of the Car Evaluation Dataset.

Correlation matrix

1	0	-0	-0	-0	0
0	1	-0	0	-0	0
-0	-0	1	-0	0	0
-0	0	-0	1	-0	-0
-0	-0	0	-0	1	-0
0	0	0	-0	-0	1

eigenvalue	proportion	cumulative	
1	0.16667	0.16667	-1V3+0.022V2+0.005V1-0.003V4-0.002V5...
1	0.16667	0.33333	1 V1-0.021V2-0.005V5+0.004V3+0.002V4...
1	0.16667	0.5	1 V2+0.023V3+0.021V1+0.003V4+0.002V5...
1	0.16667	0.66667	0.806V5-0.593V4+0.005V1+0.001V6+0 V2...
1	0.16667	0.83333	-0.805V4-0.593V5+0.008V6+0.004V3+0.004V2...
1	0.16667	1	1 V6+0.007V4+0.004V5-0V2-0V3...

Figure 7 Correlation Matrix and Part of Eigenvalues of the PCA processed Car Evaluation Dataset.

For Wilt dataset, PCA is appropriate and the projection of processed dataset is shown in Figure 8. As confirmed by previous observation (Figure 4 Right), for Wilt dataset, two attributes can effectively explain the dataset with 95% variances (Figure 8 Right, Figure 9). Also from the left, we can see that there is some separation between the data points. As mentioned before, PCA is expected to enhance the performance of both clustering algorithms.

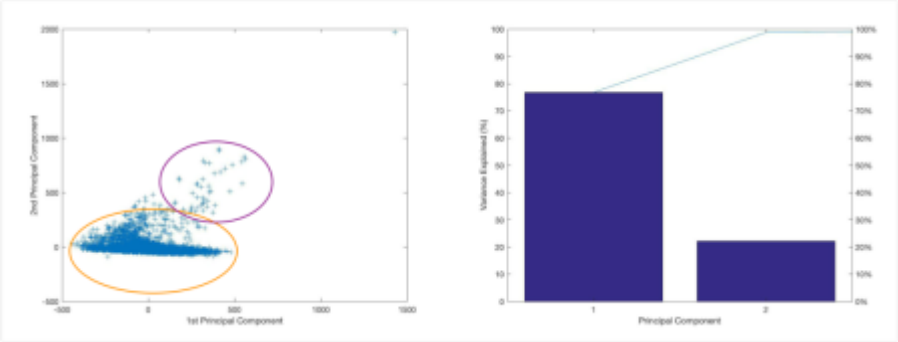


Figure 8 Projection of top 2 Principal Components (Left) and the Percent Variability Explained by each Principal Component (Right) of the Wilt Dataset.

Correlation matrix

1	-0.11	-0.1	-0.08	-0.07
-0.11	1	0.96	0.27	0.23
-0.1	0.96	1	0.16	0.23
-0.08	0.27	0.16	1	0.15
-0.07	0.23	0.23	0.15	1

eigenvalue	proportion	cumulative	
2.19177	0.43835	0.43835	-0.645V2-0.63V3-0.295V5-0.278V4+0.149V1
1.00436	0.20087	0.63922	-0.772V1+0.45 V4-0.281V3+0.271V5-0.221V2
0.92618	0.18524	0.82446	0.608V1+0.546V5+0.523V4-0.197V3-0.142V2
0.8489	0.16978	0.99424	0.736V5-0.663V4-0.107V1-0.08V2+0.006V3

Figure 9 Correlation Matrix and Part of Eigenvalues of the PCA processed Wilt Dataset.

2. ICA

ICA separates multivariate dataset into additive subcomponents, with the assumption that all attributes are non-Gaussian signals and statistically independent. The performance ICA was evaluated by kurtosis value of processed Car Evaluation and Wilt dataset. Kurtosis is a measure of the combined sizes of the two tails and the amount of probability in the tails. For a normal distribution, the kurtosis should equal to 3.

For Car Evaluation dataset, kurtosis values are [2, 2, 2, 2, 2, 2].

For Wilt dataset, kurtosis values are [20, 19, 17, 20, 20].

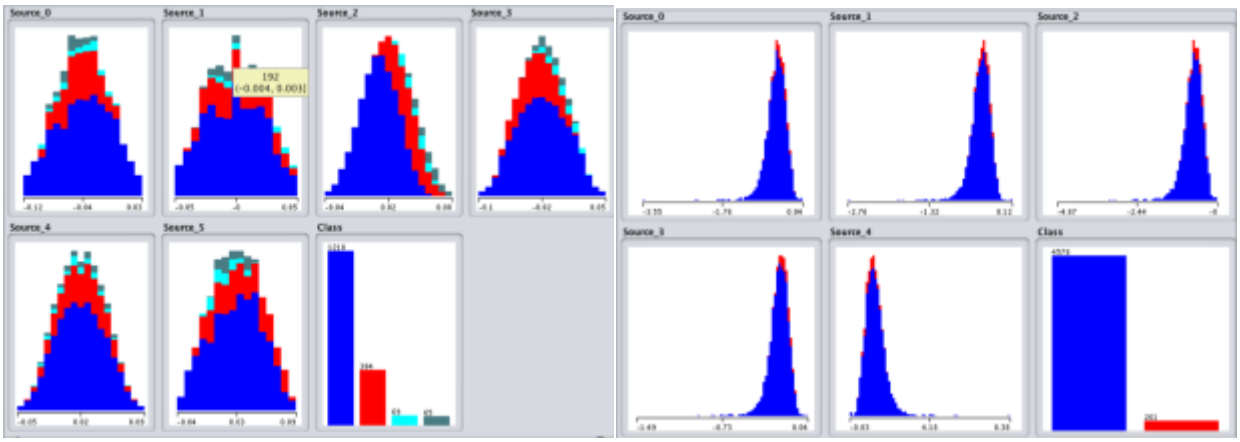


Figure 10 Distribution of sources in ICA processed Car Evaluation Dataset (Left) and Wilt Dataset (Right).

Comparing the ICA result of Car Evaluation dataset and Wilt dataset (Figure 10), we can see that the Wilt dataset has a non-Gaussian distribution while Car Evaluation dataset has a relatively Gaussian distribution. This is also shown by the kurtosis values above: Wilt dataset has kurtosis value much greater than 3 while Car Evaluation dataset has the same kurtosis values 2. Since ICA assumes that all attributes are non-Gaussian, ICA is expected to be effective on Wilt dataset and ineffective in Car Evaluation dataset.

3. Random Projection (RP)

RP, as referred by its name, is a method used to reduce the dimensionality of random variables into a suitable lower-dimensionality space which preserves much of the variation. To investigate the effect of RP, different number of attributes retained were tested (Table 2).

4. Information Gain (IG) feature selection

IG feature selection, unlike the other 3 techniques, is a supervised feature selection method. It ranks features based on the amount of information the attribute gives with respect to the classification target. To investigate the effect of IG feature selection, different number of attributes retained were tested (Table 2).

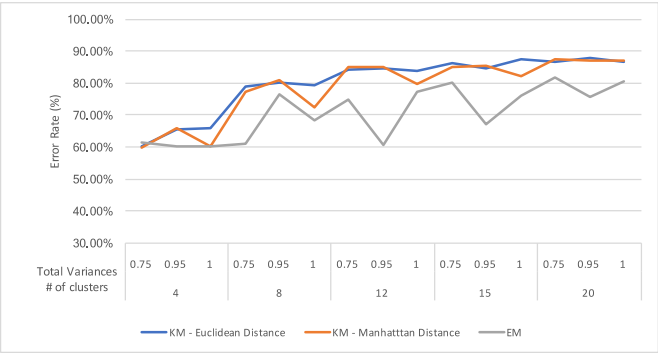
Clustering Analysis After DR

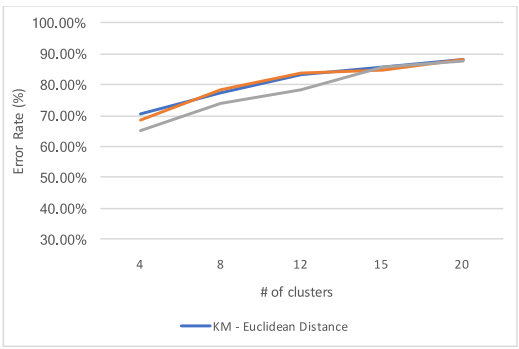
1. Car Evaluation Dataset

KM algorithms with Euclidean distance and Manhattan distance function is applied to dimensionality reduced Car Evaluation Dataset. The number of initial seed is set to be 100 since as discussed before, increasing the number initial seeds did not increase the performance. The result is summarized in Figure 11.

First, as can be seen from the graph, there is no major improvement in the performance of both EM and KM on this dataset. The best clustering accuracy of KM and EM is achieved after IG feature selection of 1 attribute attained, where all clustering algorithms have the same error rate of 52.49%. This is expected since IG feature selection is only DR algorithms investigated involves label information. This in certain level, corrects the bias toward certain attributes, which does not add information regarding the correct classification. However, as discussed before, due to the nature of this dataset, clustering algorithms still perform poorly comparing with supersized algorithms (Table 2).

Another important observation that we can see from Figure 11 is that for PCA, as the total variances increases from 0.75 to 1, the performance of KM in most cases, increases first and then decreases. However, the performance of EM does not have a clear pattern. For ICA, the performance of both KM and EM worsens comparing with non-DR reduced clustering result. This is expected from previous analysis of ICA: ICA assumes non-Gaussian data while this dataset exhibit relatively Gaussian nature. For RP, there is no clear relationship between the performance and number of attributes attained. Finally, for IG feature selection, there is a relationship between the number of attributes attained and performance of the clustering algorithms. As the number of attributes increases from 1 to 3 and then 5, the performance in general worsens. This is happening because instances of same class do not show any proximity in feature values, as shown by figure 1. Thus, more attributes will only bring more confusion.





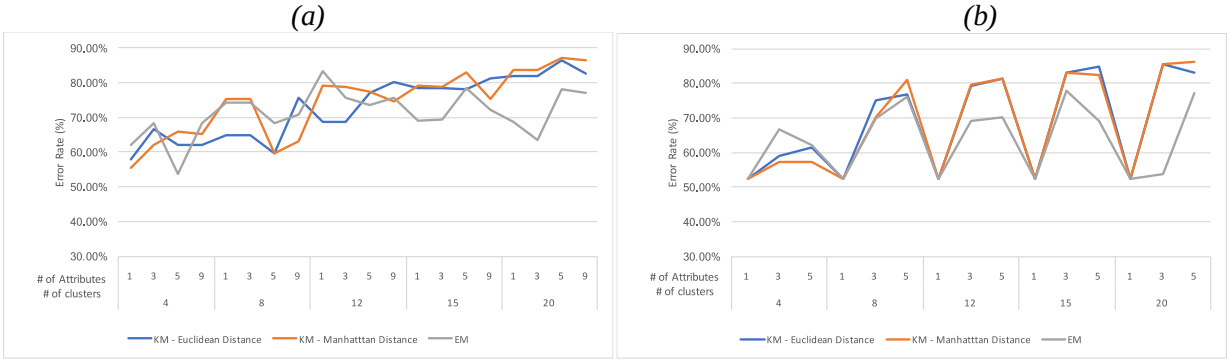


Figure 11 Result of Clustering Analysis after 4 Dimensionality Reduction Techniques on Car Evaluation Dataset. (a) PCA; (b) ICA; (c) RP; (d) IG feature selection.

2. Wilt Dataset

KM algorithms with Euclidean distance and Manhattan distance function is applied to dimensionality reduced Wilt Dataset. The number of initial seed is set to be 100 since as discussed before, increasing the number initial seeds did not increase the performance. The result is summarized in Figure 12.

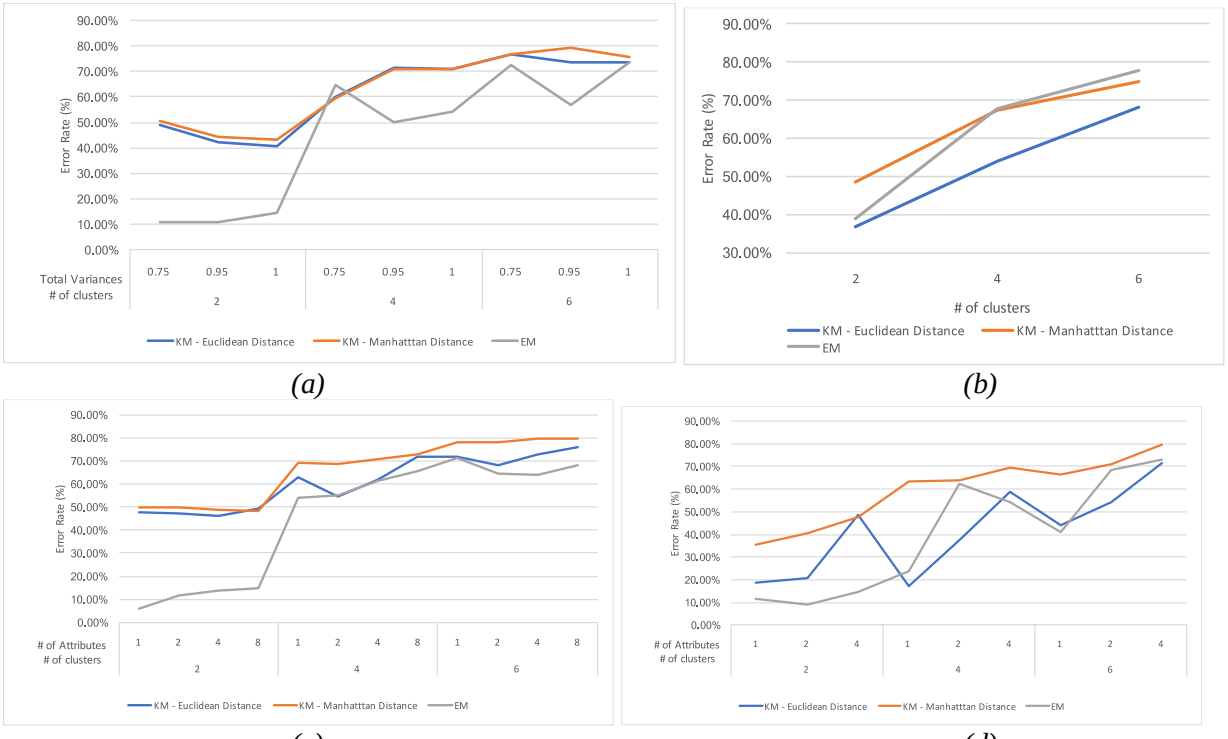


Figure 12 Result of Clustering Analysis after 4 Dimensionality Reduction Techniques on Wilt Dataset. (a) PCA; (b) ICA; (c) RP; (d) IG feature selection.

The first thing to note is that except ICA (Figure 12b), three out of 4 DR algorithms enhanced the performance EM. What's more, as the number of attributes/variances increases, there is a general trend of worse performance. This is expected as discussed in previous sections due to the nature of this dataset. Moreover, the best performing number of clusters is 2, which makes sense since this is a binary classification problem.

Although ICA worsen the performance of EM, ICA dramatically improved the performance KM with Euclidean distance function. Besides, all DR algorithms favored Euclidean distance functions over Manhattan distance function. This makes sense since this dataset has low number of attributes. Euclidean distance works better in low dimensionality while Manhattan distance usually works better in high dimensionality.

Neural Network Learning After DR

I then explored the effect of DR algorithms on the performance of Neural Nets (NN) on Wilt dataset. Based on the result of assignment 1, I setup a multilayer perceptron NN with learning rate 0.2, momentum 0.3 and 10 hidden layers. NN is trained using dimensionality reduced dataset and the classification accuracy on testing dataset is reported in the following Table.

PCA			
Total Variances Explained	0.75	0.95	1
NN Accuracy	72.05%	94.61%	98.10%
ICA			
NN Cross Validation Accuracy		97.07%	
RP			
Number of Attributes Attained	2	4	6
NN Accuracy	94.61%	94.61%	94.98%
IG feature selection			
Number of Attributes Attained	2	4	6
NN Accuracy	94.61%	97.71%	97.91%

Table 6. *The result of NN Cross Validation Accuracy after Dimensionality Reduction on Wilt Dataset.*

As can be seen from the table, the highest accuracy is obtained by PCA reduced dataset with 100% total variances explained. Besides, for PCA, as the number of total variances explained increases, there is an increase of NN accuracy. For RP, the number of attributes attained does not have huge impact on the accuracy. For IG feature selection, as the number of attributes attained, the testing accuracy increases accordingly.

The highest accuracies gained from PCA, ICA and IG feature selection-reduced dataset are very high and very close to the accuracy found in assignment 1 (Table 4). This indicates that although ICA did not generate good results in clustering analysis, it still reconstructs the data well. For PCA, NN performs well with higher total variances explained makes sense since NN can approximate any nonlinear mapping through learning and thus there is no need to reconstruct dimension in advance, which can be achieved just as well by the input layer weights. Thus, more variances (more input information) means higher accuracy. Although DR is not needed for NN, the benefit of dimensionality reduction is that reduces the size of the training data and thus the size of the network. For example, for RP and IG feature selection, only 2 features are needed to achieve accuracy as high as 94.61%.

Finally, NN trained on dimensionality-reduced dataset has lower accuracy than the original dataset, on which the accuracy of NN is 98.62% (Table 4). This is since the discriminative information that distinguishes one class from another for some data points might be in low variance components. So, using DR algorithm can make performance worse.

Neural Network Learning After Clustering

The best performing parameters as shown in Table 7 is used for KM and EM in the experiments.

	KM	EM
seed	100	100
Distance function	Euclidean	NA
Number of clusters	2	2

Table 7. *The parameters used for KM and EM Clustering for Neural Network on Wilt Dataset*

	Total Variance	0.75	0.95	1
PCA	KM	57.74%	58.36%	59.41%
	EM	70.26%	92.65%	94.07%
	No clustering	72.05%	94.61%	98.10%
	# of Attributes	2	4	6
RP	KM	53.59%	56.42%	55.05%
	EM	93.61%	91.65%	90.93%
	No clustering	94.61%	94.61%	94.98%
	# of Attributes	1	2	4
IG	KM	60.97%	61.57%	66.60%
	EM	89.42%	90.20%	90.60%
	No clustering	94.61%	97.71%	97.91%
ICA	KM	68.22%		
	EM	72.63%		
	No clustering	97.07%		

Table 8. *The result of Neural Networks with clustering as dimension reduction.*

Clustering Information is added to the original data to test the effect of cluster analysis on NN performance. The result is summarized in Table 8.

As can be seen from the table that the classification accuracy of NN trained with clustering analysis is lower than NN trained with dimensionality reduced dataset (No Clustering). NN trained by EM clustering analysis on PCA, RP and IG-reduced dataset has reasonable performance. The pattern of increasing classification accuracy as the total variances of PCA increases can also be observed here. In general, the general trend of classification accuracy of NN trained with clustered data is the same as that of NN trained with dimensionality reduced dataset (No clustering) except for RP where the accuracy decreases as the number of attributes increases. This indicates change of overfitting or RP might not reconstruct the feature space well.

It is expected that EM has better performance than KM according to the clustering analysis result after dimensionality reduction. The probability model used by the EM clustering algorithm is more effective in recognizing the classification patterns, especially in this dataset which consists of only 5% of positive class.