Unsupervised Learning and Dimensionality Reduction

Alek Francescangeli

November 6, 2022

1 Introduction

In this paper, we will be comparing two datasets using six algorithms: k-means, expectation maximization, independent component analysis, random projection, and information gain. We will then discuss the results and problems that occurred during experimentation.

2 The Datasets

For these experiments we will be using two datasets, one that describes the quality of red wine and one that is used to determine whether credit card fraud has occured.

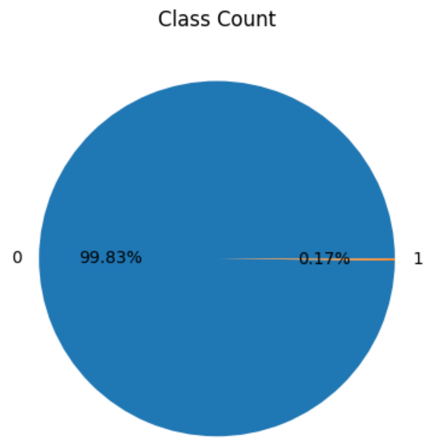
2.1 Red Wine Quality Dataset

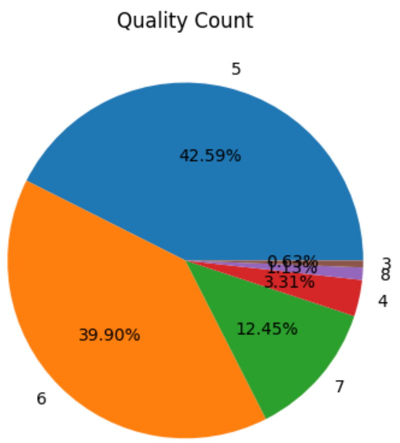
This dataset has approximately 1,600 instances in it with 11 inputs that are used to classify what quality the wine will be. This dataset is interesting because there is a degree of subjectivity to it. While one particular wine could have identical inputs the person who rated the wine could rate it differently based on their own personal biases. This dataset is unbalanced with over approximately 82% of the wine being rated either a 5 or a 6.

2.2 Credit Card Fraud Dataset

This dataset has approximately 280,000 instances in it with 28 inputs that are used to classify whether the specified transaction is fraudulent. This dataset is much more unbalanced with 99.83% of instances having a class of ”0” which indicates a non-fraudulent transactions, with a class of ”1” indicating a fraudulent transaction.

Dataset Distributions:





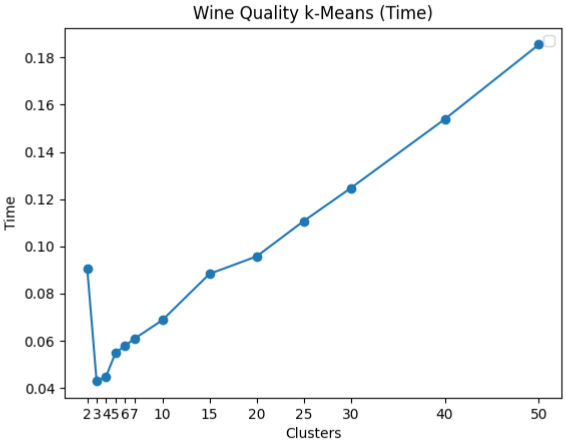
3 Clustering Algorithms

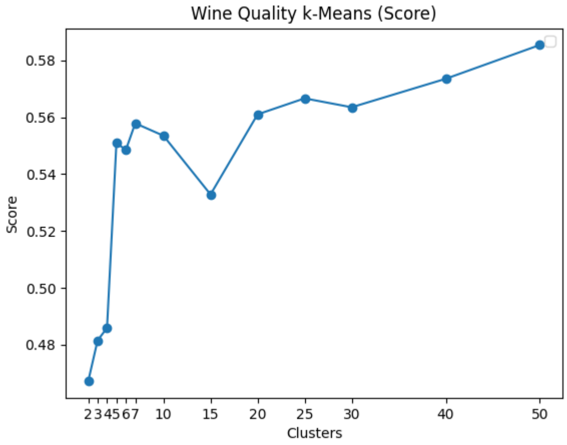
There are two clustering algorithms we will use in this section: k-Means and Expectation Maximization. K-means works by separating the dataset into random k-clusters, centroids are then calculated for each clusters. These centroids are then used by the K-means algorithm in whcih it prefeorms iterative calculations in order to optimize the positions of each of the centroids. [1]

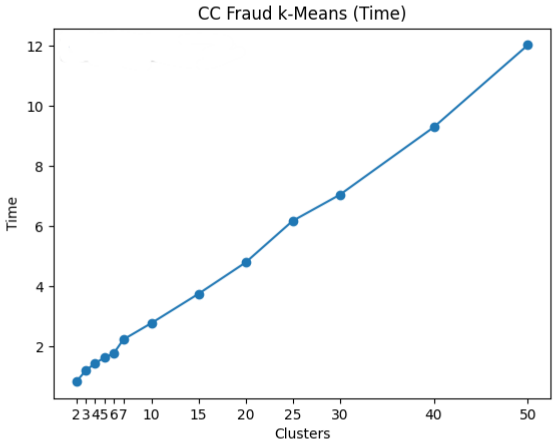
Expectation Maximization works by finding the k-distribution in data so that the Log Likelihood of the data distributions are maximized. The algorithm will cycle between two modes. The first mode will atempt to esitmate the missing or latent variables, with the second step optimizing the parameters of the models.[2]

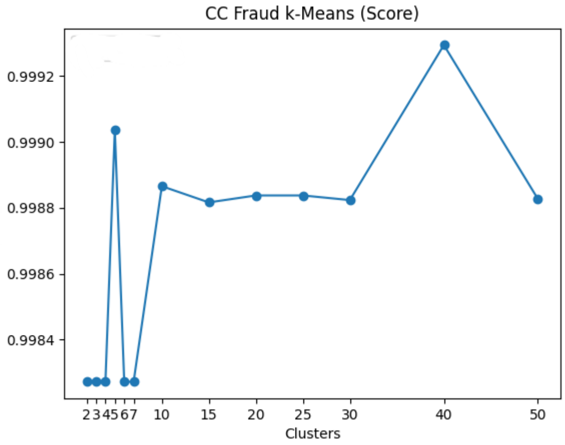
For both datasets we can see that increasing the number of clusters increases the accuracy score. It also greatly increases the time, while time is not a major issue for the small wine quality dataset, time proved to be a major issue for the larger credit card fraud dataset, with calculations taking many magnitudes longer to finish. Overfitting is also a large concern, I believe the graphs show that there may have been some issues in both k-means and EM. Some tweaks were made (such as running with smaller samples of the larger dataset and adjusting the initial value/seeds) but the behavior was still often erratic on that dataset. The wine quality dataset on the other hand acted in a much more expected manner, with more clusters correlating to a better accuracy score. Overall from these results, I determined that for both datatsets the optimal number of clusters was approximately 30, both in a sense of time and accuracy results.

3.1 k-Means Visualizations

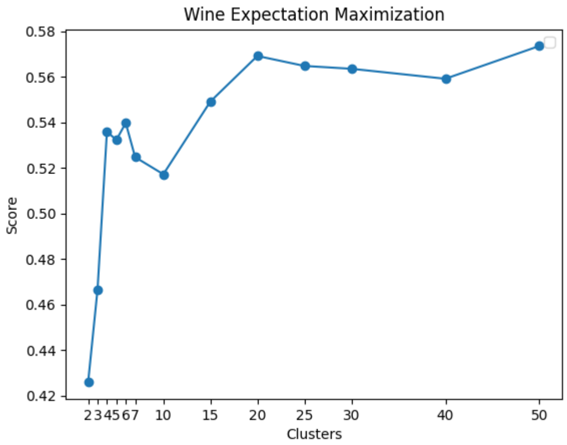


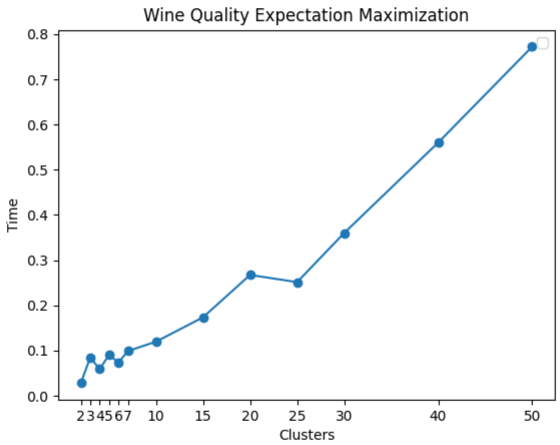


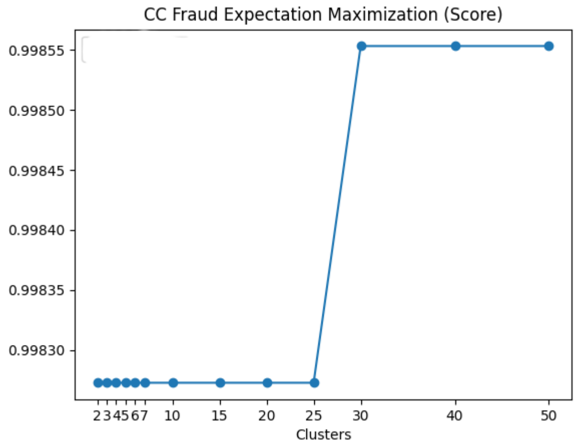


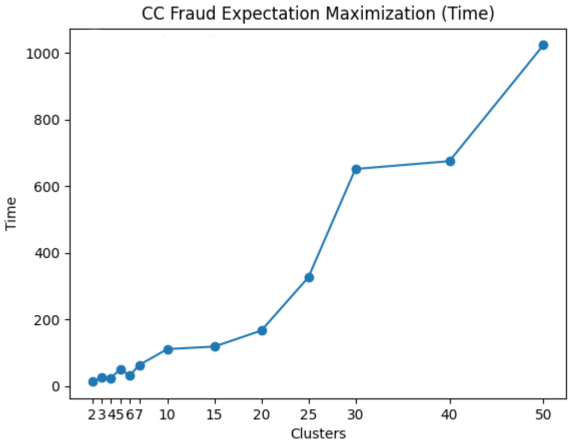


3.2 Expectation Maximazation(EM) Visualizations





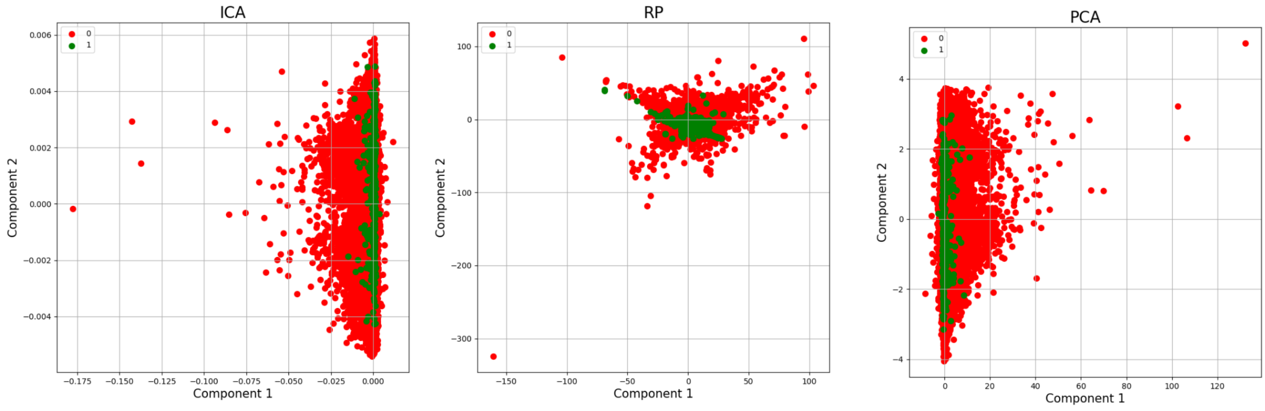




4 Dimensionality Reduction Algorithms

For the following experiments we will be using dimensionality reduction algorithms on the previous EM and K-means method in order to see changes and improvements that can be made. These algorithms act as a pre-processing step in order to improve supervised learning algorithms such as k-means and expectation maximization that were discussed earlier in this paper. We will be using values of 1-11 to test which components perform better.

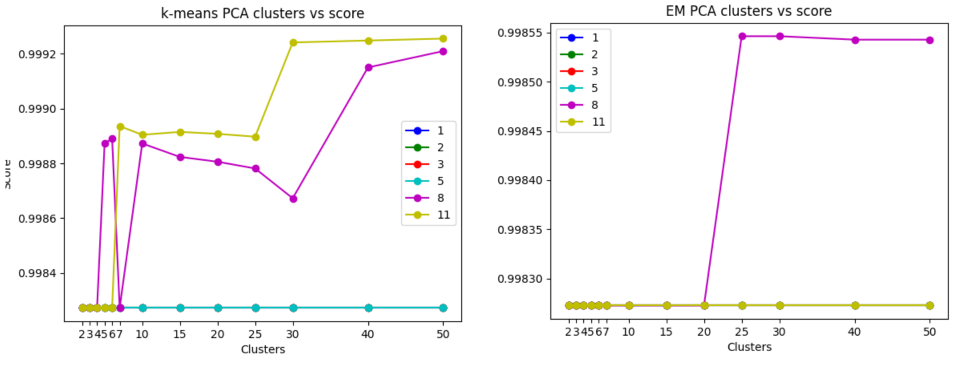
Dimensional Visualizations for Credit Card Fraud dataset:



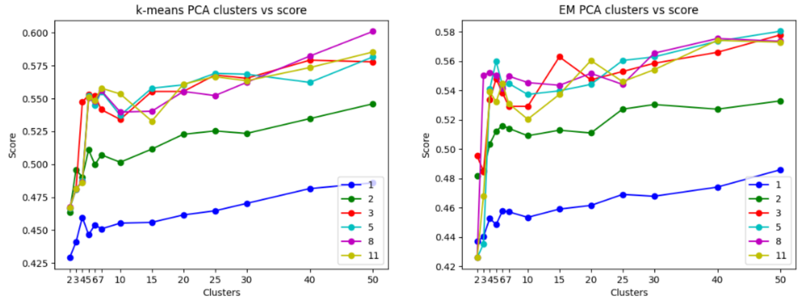
4.1 Principal Component Analysis(PCA)

Principal Component Analysis is a technique that linearly transforms the data into a coordinate system, so that the data can be described with fewer dimension than in the initial data set. For the credit card dataset, PCA was not very effective with only two component settings (1 and 11) yielding different results. It seems to have yielded results for the wine quality dataset with clear results for different components.

Figures for using PCA with k-means/EM on credit card dataset:



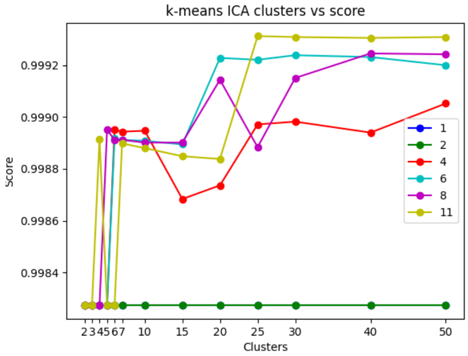
Figures for using PCA with k-means/EM on wine quality dataset:

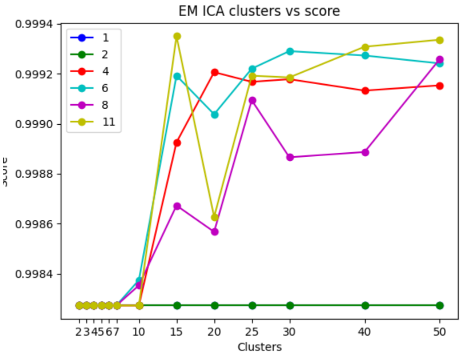


4.2 Independent Component Analysis(ICA)

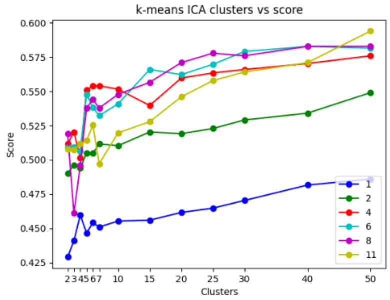
Independent Component Analysis works by having the variables by the output of multiple unobserved sources in the data. [3] This method provides more clear cut improvements based on the component count.

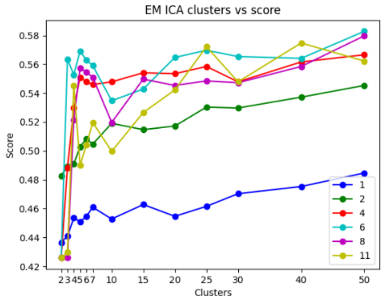
Figures for using ICA with k-means/EM on credit card dataset:





Figures for using ICA with k-means/EM on wine quality dataset:

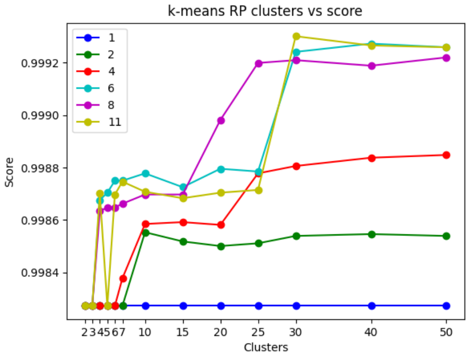


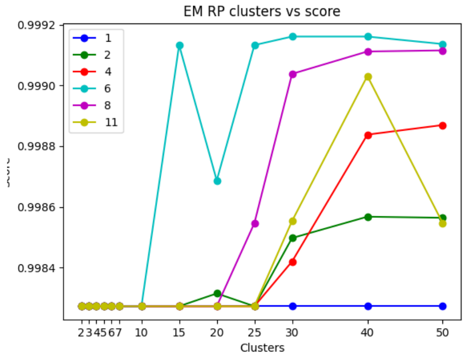


4.3 Random Projection (RP)

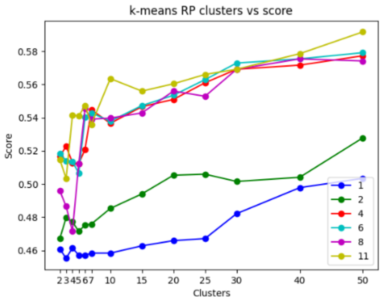
Random Projection is like PCA in that it maps data into a coordinate system with a key distinction being that being that the projections are independent from the data. This method also provides clear improvements based on the component count.

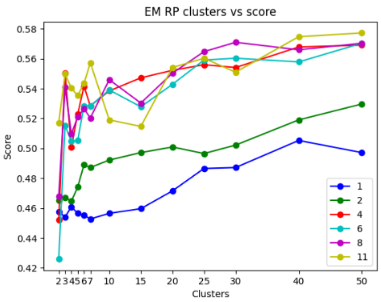
Figures for using RP with k-means/EM on credit card dataset:





Figures for using RP with k-means/EM on wine quality dataset:





4.4 Information Gain (IG)

Information Gain serves as a way to reduce entropy by splitting the datasets into groups that are best for classification. This is done by evaluating information gain for each variable and then selecting the variable that maximizes the information gain. [4]

5 Conclusion

Clustering algorithms can be helpful as a pre-processing technique. While no improvement for the large unbalanced credit card dataset was found, the wine quality dataset performed much better with the clustering algorithms. I believe this may be indicative of larger datasets having a higher computation cost to dimensionally reduce. With 11 components the wine dataset performed significantly better across all dimensionality reduction algorithms. A different component range may of improved the results of the credit card results but the computational power needed to run those tests may not be feasible.

References

[1] Understanding K-means Clustering in Machine Learning https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1

[2] A Gentle Introduction to Expectation-Maximization https://machinelearningmastery.com/expectation-maximization-em-algorithm/

[3] What is Independent Component Analysis: A Demo http://research.ics.aalto.fi/ica/icademo/

[4] Information Gain and Mutual Information for Machine Learning
https://machinelearningmastery.com/information-gain-and-mutual-information/