Data Mining with Spectral Clustering and Comparative Clustering Analysis: An Exploration with R

OLIVIA SAMPLES
College of Computing & Technology, Lipscomb University, Nashville, TN, osamples@mail.lipscomb.edu
NATHAN MORGAN,
College of Business, Lipscomb University, Nashville, TN, nmorgan@mail.lipscomb.edu

ABSTRACT

The team was approached by Big Data Science Incorporated (BDSI) to determine how well spectral clustering compares to other well-known methods of clustering. The comparative algorithms chosen were K-Means, K-Medoids, and Spectral Clustering to determine whether or not a randomized dataset could be successfully clustered. After the code in R was run on our chosen dataset of Wine, the accuracy function and p-value showed quantifiably that the well-known method of K-Medoids and Spectral Clustering did a better job of correctly clustering the data than K-Means. The team understands that the work above could come up with different results for different datasets. The most important thing to understand is that clustering work should never be done with just one algorithm all the time. Any data scientist must try a few methods to make sure they come up with the best clusters for the data they have.

1 INTRODUCTION

In April 2020, Olivia Samples and Nathan Morgan ("the team") were approached by Big Data Science Incorporated (BDSI) to determine how well spectral clustering compares to other well-known methods of clustering. The goal was to come up with a quantifiable measurement in a test environment as opposed to the production environment. Using a popular dataset, BDSI wanted to know if the use of spectral clustering would allow the team to find new discoveries or a more accurate manner of predicting the correct clusters.

The comparative algorithms chosen were K-Means, K-Medoids, and Spectral Clustering to determine whether or not the dataset could be successfully clustered. The null hypothesis formed was that our dataset could not be successfully clustered, and the alternative hypothesis stated that our dataset could be successfully clustered. After building models on our dataset, these different algorithm performances were compared using their results from silhouette scores, respective Confusion Matrices, and prediction accuracy on test data.

A large amount of related work has been devoted to K-Means and K-Medoids cluster analysis, specifically with our chosen dataset of wine. Krista Zalik successfully clustered the wine dataset using an efficient form of the K-Means clustering algorithm. The purpose of the analysis was to cluster a dataset without initially assigning the number of clusters by involving a cost-function.[10] Francesco Gullo more accurately clusters the wine dataset using an effective k-medoids method called UK-medoids. In practice, his method outperformed others in accuracy.[5] This paper will compare the performance of these well-analyzed methods to that of spectral clustering.

2 METHOD

2.1 Dataset Description

Through a recommendation by Dr. Tim Wallace from Lipscomb University, the team began their search for a dataset online with the University of California, Irvine. The university has a large repository of datasets that are free to be used by the public. After doing rounds of initial testing with a few different datasets, the team landed on the wine dataset as the best option to pursue.[9] The wine dataset compares 178 different wines from the same region in Italy that were developed from three distinct cultivars. There exist 13 constituents as a result of a chemical analysis of these wines which are the 13 dimensions used in this analysis. Each chemical analysis record is labeled in respect to its cultivar, resulting in the "Class" attribute referred to in this paper as the true label. Since the dataset already had true labels, had more than two dimensions, and there are existing tests on this data for comparison, the team thought this dataset would be the most suitable for testing the clustering methods.

2.2 Data Exploration

In order to determine how each feature interacted with each other, the exploratory data analysis focused on attribute correlation and normalization of the data. The correlation matrix shown in Figure 1 displays the majority of the features having a weak to moderate correlation to one another. Additionally, results from singular value decomposition confirmed that each attribute contributes to the data variance, so we did not choose to remove any features for modeling purposes.

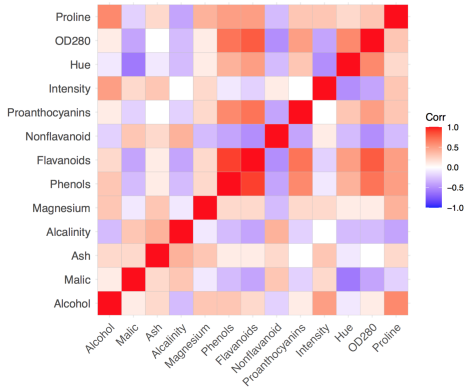

Figure 1: Feature Correlation

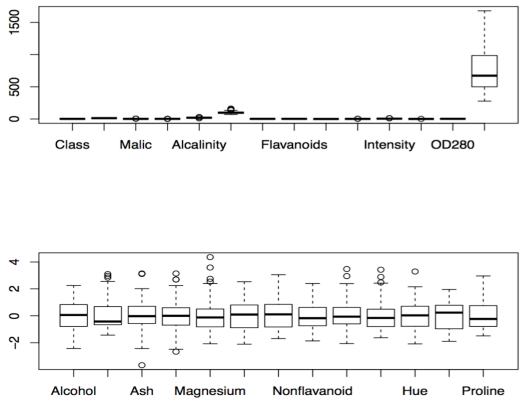

Figure 2: Original Boxplot Vs. Normalized Boxplot

However, we did choose to normalize our data based on the boxplot representation of the features presented in Figure 2. Here one can see that the original scale for the data attributes were skewed as "Proline" was substantially larger than the rest. The second box plot represents the normalized data that we used to perform the following methods.

Finally, we separated our dataset for training and testing purposes. In order to have normal representation, the data was randomly separated at a *70/30* split. Seventy percent of the normalized, randomly separated data is the dataset used to train each clustering model, leaving the other 30% for testing.

3 RESULTS

3.1 K-Means Algorithm

The first algorithm we use to train our model is K-Means. K-Means is an iterative algorithm that finds the points closest to the centroid and the clusters furthest away from one another by minimizing the mean squared error. We applied this to our training dataset using the function kmeans() in R for both $k = 2,3$. The resulting clusters vs. true label confusion matrix for 3 clusters produced a p-value of *0.994* which means we will not reject our null hypothesis for K-Means. The probability of accurately predicting the class, given it is Class 2 is *0%*. Similarly, giving it is Class

3, the probability of predicting accurately is *0%.* Confirming the inability to cluster the dataset using K-Means, the accuracy using the model to predict on the test data was *55.56%.* Hence, it is likely that we will obtain these results at random.

However, it is interesting to note that the resulting plot has three distinct clusters with clear centroids. While the confusion matrix was inaccurate, the algorithm correctly identified all Class 3 wines but misidentified most Class 1 wines for Class 2 wines and vice versa. In other words, the data was mostly separated into the correct number of classes, just incorrectly labeled. The silhouette coefficients for each cluster, which measure similarity between intra-cluster points, were moderate (*0.37, 0.36, 0.18*). The closer the coefficients are to *1*, the more similar the intra-cluster points. Conversely, the closer the coefficients are to *-1*, the more dissimilar the intra-cluster points.[2] For our trained algorithm, the intra-cluster points are not strongly dissimilar which might explain the three distinct clusters in Figure 3.

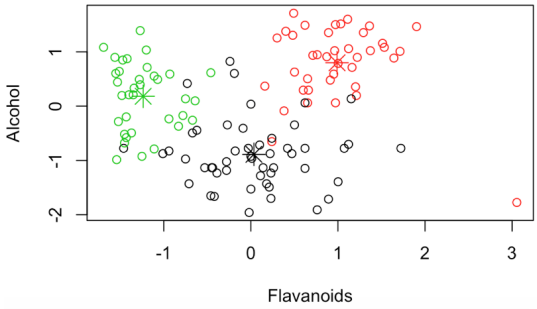

Figure 3: K-Means Cluster Relation

When running the model for two clusters, the Bayesian probability for accurately predicting class given it is Class 2 went from *0%* to *50%* as it was the only Class with accurate predictions. However, the silhouette coefficients decreased in size (*0.22, 0.27*) and the Bayesian probabilities decreased for Class 1 and Class 3. Overall, kmeans() performed worse for *k=2*.

3.2 K-Medoids Algorithm

K-Medoids is similar to K-Means in that it partitions data into a set, k, number of clusters based on minimizing distance to a centroid, but it is better at working with noise and outliers. K-Medoids uses an actual point in the cluster for the centroid whereas K-Means uses the mean of the cluster.[7] Beginning with two clusters, the algorithm has a good accuracy rate on the test data with *51.85%.* For the trained data in the confusion matrix, *88.09%* of Class 1 was accurate and *22.44%* of Class 2 was accurate.

When considering K-Medoids for 3 clusters, the accuracy for the training data is *89.52%* and *87.04%* for the test data. The p-value is the smallest value that R returns, $< 2.2\ e^{-16}$, meaning that we can reject the null hypothesis at a significant value of 0.05. Because the silhouette coefficients (*0.36, 0.16, 0.31*) are moderately close to 1, they perform well enough for our model. The original data set had 42 Class 1 wines, 49 Class 2 wines, and 35 Class 3 wines. As shown in Figure 4, the probability the model predicts the cluster correctly given the Class 1 cluster is *71.43%*, given the Class 2 cluster is *93.88%*, and given the Class 3 cluster is *100%.*

	Reference: 1	Reference: 2	Reference: 3
Prediction: 1	30	2	0
Prediction: 2	7	46	0
Prediction: 3	0	4	35

Figure 4: K-Medoids Correlation Matrix

3.3 Spectral Clustering Algorithm

Spectral clustering is a technique that applies well known clustering techniques, such as k-means, onto the eigenvectors of the Laplacian matrix of the dataset. As this is the technique specifically requested by BDSI for analyzing, we saved it for last; however, it seemed to have performed the similar to K-Means on our wine data set.

For *k=3*, our silhouette coefficients (*0.38, 0.16, 0.37*) are moderate which implies that the three clusters formed are somewhat similar. Also, the confusion matrix accuracy on the training dataset is *52.42%* with a p-value of *0.01188*. Here we can reject the null hypothesis as the p-value is less than the significant value. And so, spectral clustering can correctly cluster our dataset.

	Reference: 1	Reference: 2	Reference: 3
Prediction: 1	37	2	0
Prediction: 2	0	0	35
Prediction: 3	0	50	0

	Reference: 1	Reference: 2	Reference: 3
Prediction: 1	30	2	0
Prediction: 2	7	46	0
Prediction: 3	0	4	35

	Reference: 1	Reference: 2	Reference: 3
Prediction: 1	29	0	0
Prediction: 2	8	1	0
Prediction: 3	0	51	35

Figure 5: Correlation Matrices of K-Means (Top), K-Medoids (Middle), and Spectral Clustering (Bottom) on Wine Dataset

When comparing the confusion matrices from each algorithm (Figure 5), it is clear that Class 1 and 3 are more distinct than Classe 2 as they are predicted accurately most often. Overall, each cluster remains relatively distinct, but is sometimes mislabeled. It is clear that K-Medoids predicts the most accurately, but that the three different wine classes are strongly unique.

4. POTENTIAL CHALLENGES

As the dataset is small, only *178* instances, the algorithms are limited. There are only *~13* records for each feature. This sizing can be a drawback for iterative algorithms especially. A further analysis of spectral clustering compared to K-Means and K-Medoids can be performed on a larger dataset in order to receive clearer results.

5. CONCLUSION

The purpose of this paper was to compare two clustering algorithms to spectral clustering. Using the wine dataset, we were able to show that K-Medoids and Spectral Clustering can accurately cluster the data, rejecting the null hypothesis. However, K-Means was unable to provide significant results. These methods may be able to perform better using a larger dataset, as they are iterative in nature. K-Medoids also tends to work better with outliers. Future work may include a more thorough analysis on a larger well-known dataset with fewer outliers.

REFERENCES

[1] Aditya Garg. October 9, 2017. R Commands. RPubs by RStudio. Retrieved on April 29, 2020 from https://rpubs.com/ID_Tech/S1

[2] Alboukadel Kassambara. 2018. Cluster Validation Statistics: Must Know Methods. Retrieved on April 29, 2020 from https://www.datanovia.com/en/lessons/cluster-validation-statistics-must-know-methods/

[3] Edwin de Jonge and Mark van der Loo. 2013. An introduction to data cleaning with R. Retrieved April 29, 2020 from https://cran.r-project.org/doc/contrib/de_Jonge+van_der_Loo-Introduction_to_data_cleaning_with_R.pdf

[4] ggcorrplot: Visualization of a Correlation Matrix using ggplot2. STHDA. Retrieved on April 29, 2020 from http://www.sthda.com/english/wiki/ggcorrplot-visualization-of-a-correlation-matrix-using-ggplot2

[5] Gullo F., Ponti G., Tagarelli A. 2008. Clustering Uncertain Data Via K-Medoids. In: Greco S., Lukasiewicz T. (eds) Scalable Uncertainty Management. SUM 2008. Lecture Notes in Computer Science, vol 5291. Springer, Berlin, Heidelberg

[6] How to Compute Distances Between Centroids and Data Matrix (for kmeans algorithm). 2014. Retrieved April 29, 2020 from https://stackoverflow.com/questions/27082378/how-to-compute-distances-between-centroids-and-data-matrix-for-kmeans-algorithm/27088515

[7] Jin X., Han J. 2011. K-Medoids Clustering. In: Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA

[8] Wallace, T.L. 2020. Data Mining and Analytic Methods, Theory and Practice: R, Julia, and Python DRAFT WIP.

[9] Wine Dataset: http://archive.ics.uci.edu/ml/datasets/Wine/

[10] Zalik, K.R. 2008. An efficient k'-means clustering algorithm. Pattern Recognit. Lett., 29, 1385-1391.