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

OLIVIA SAMPLES AND NATHAN MORGAN

Outline

1. Introduction/Background Information
2. Methodology
3. Results
4. Potential Challenges
5. Conclusion

1. Introduction/Background Information

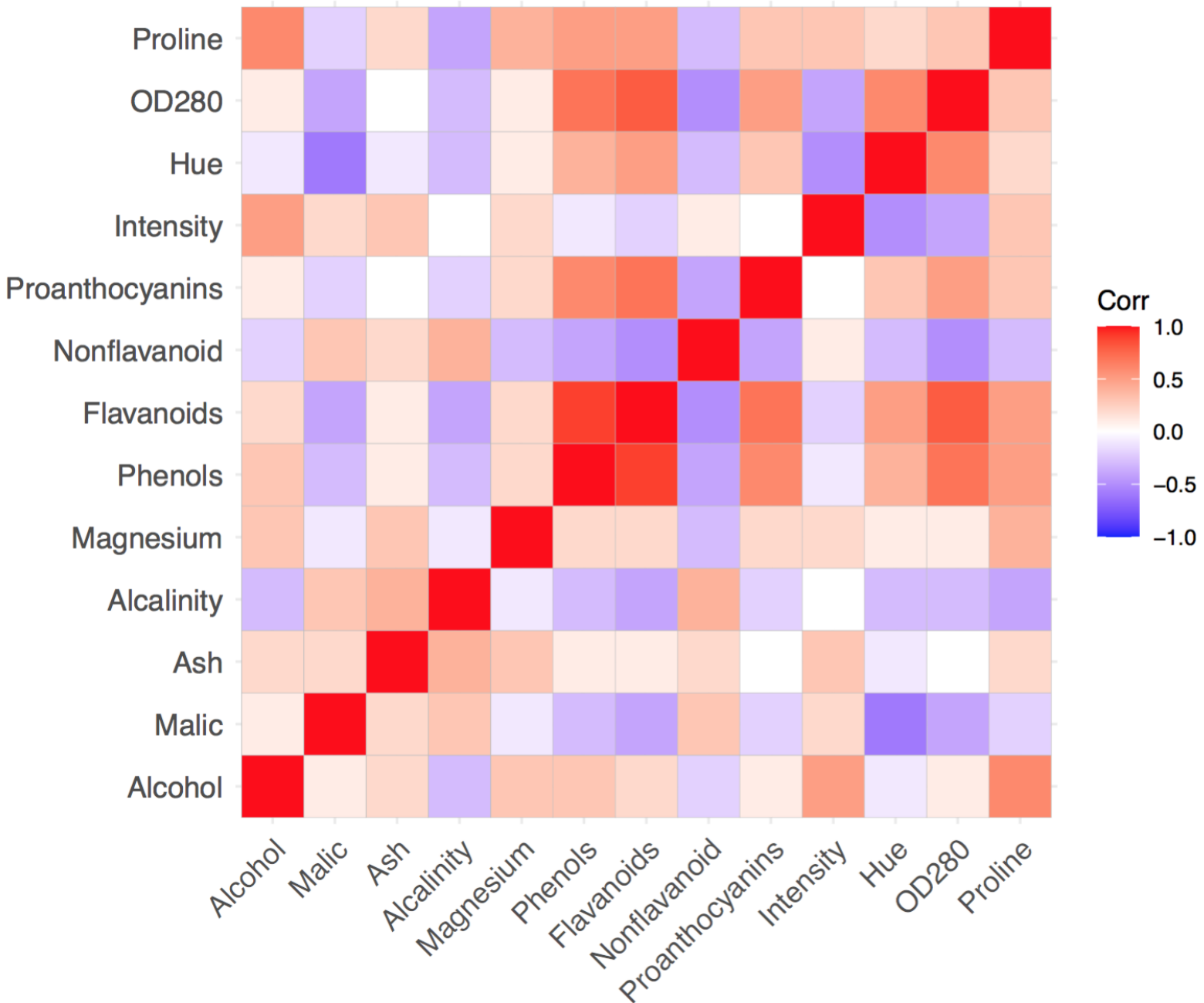
- Approached by Big Data Science Incorporated (BDSI) to compare spectral clustering to other well-known clustering methods
- Goal: find a way to quantifiably measure which methods are best
- Clustering Methods Chosen: K-Means and K-Medoids
  - Previous work had been done on the chosen data set with these methods

2. Methodology

- Chosen Dataset: The Wine Dataset
  - Found within the University of California, Irvine's data repository
  - Compares 178 different wines from the same region in Italy that were developed from three distinct cultivars
  - Has 13 dimensions derived from a chemical analysis of each wine
  - Keys to choosing this dataset:
    - Already contains true labels
    - Existing tests had been completed on this data

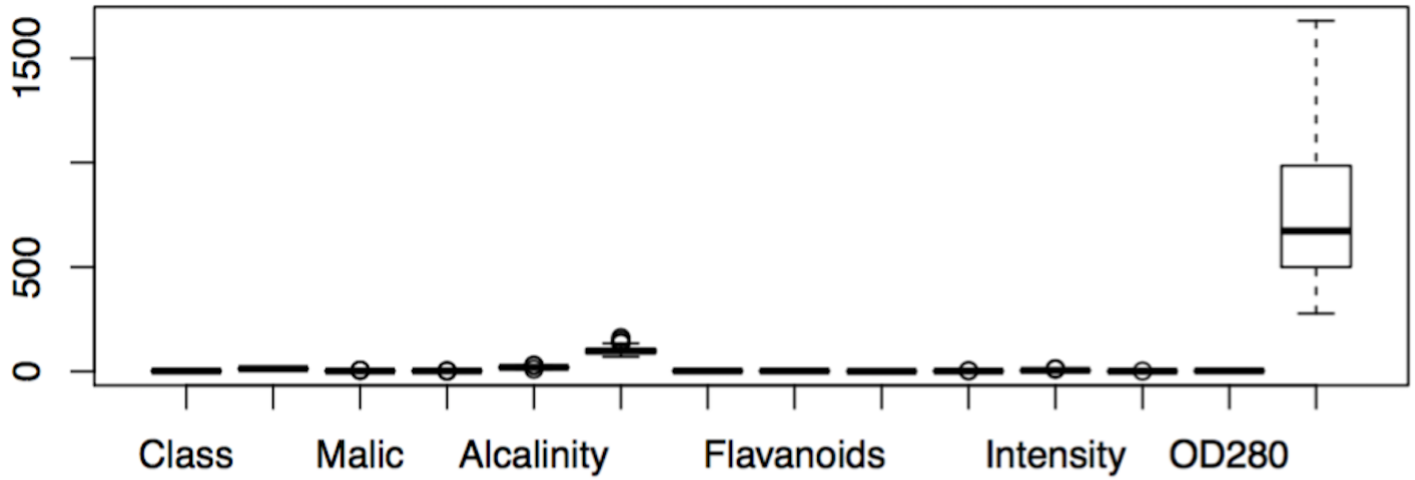
Correlation Graph

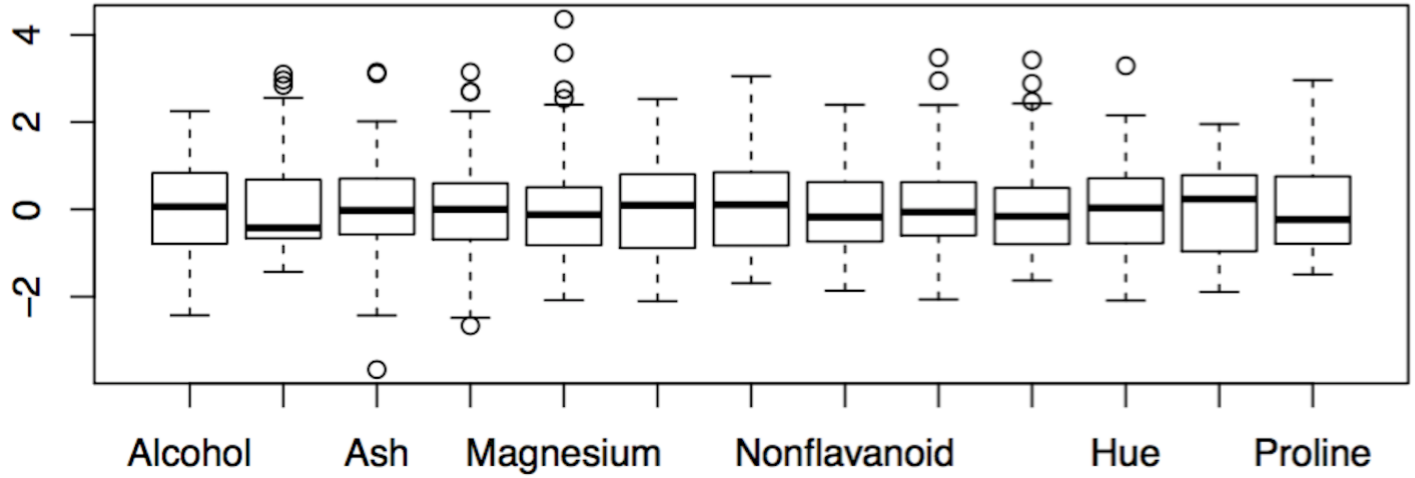
- Data Exploration
  - Before digging into the clustering, a few functions were carried out on the data to find correlations and normalize it
  - This first graph shows the correlation of the data using the ggcorrplot function



Normalization w/ Boxplots

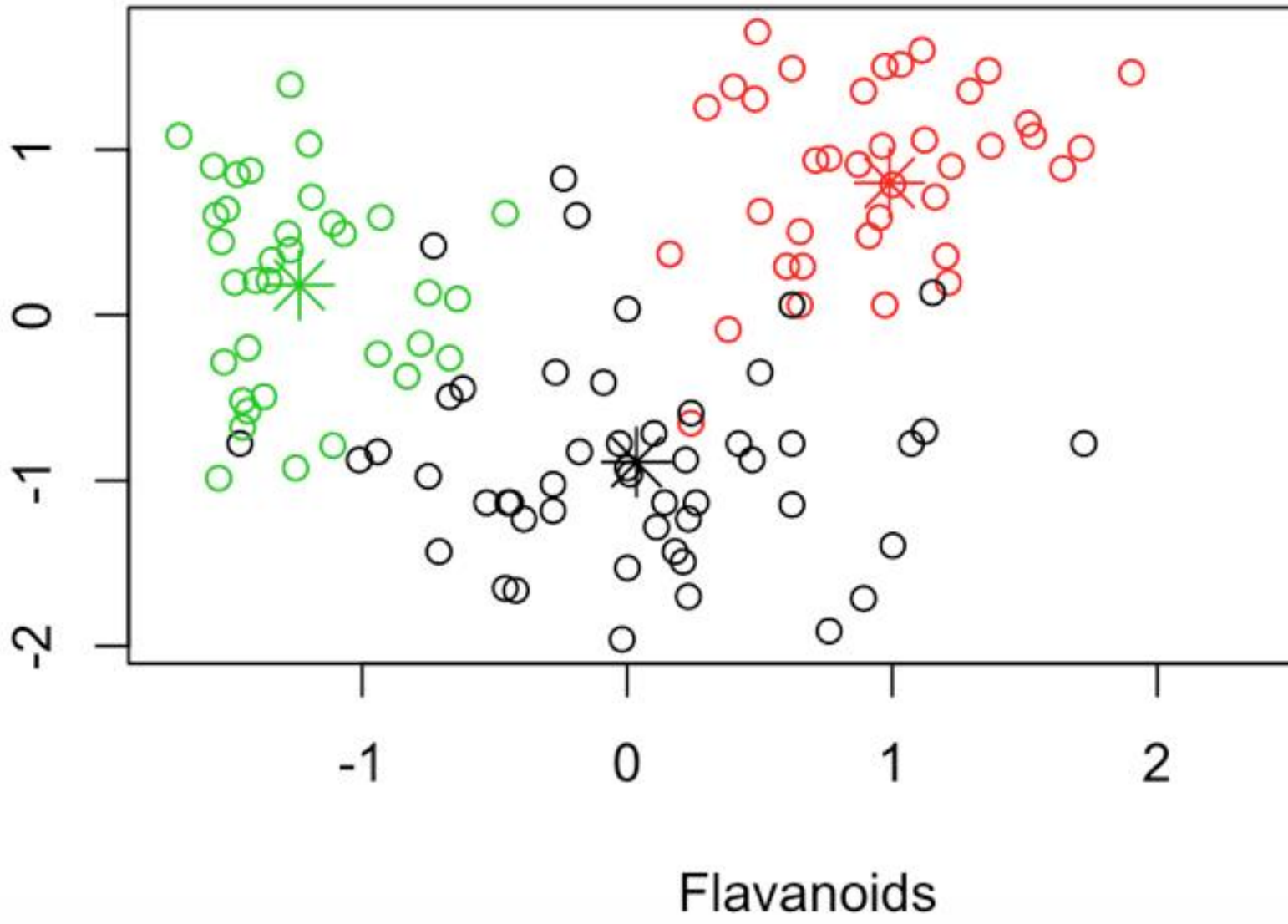
- Data Exploration
  - These two graphs show the normalization of the data through the use of boxplots
  - Finally, the data was split into training and testing groups before clustering





3. Results

- K-Means Algorithm
  - P-value: 0.994
  - Accuracy: 55.56%
  - Well-defined clusters that were incorrectly labeled
  - Given 3 well-defined classes, clustering with k=2 produced worse results



3. Results

- K-Medoids Algorithm
  - P-value: < 2.2e^-16
  - Accuracy: 87.04%
  - Graph shows strong similarities in predictions

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

3. Results

- Spectral Clustering Algorithm
  - P-value: *0.01188*
  - Train Accuracy: 52.42%
  - Test Accuracy: 40.74%
  - Silhouette Coefficients: (*0.38, 0.16, 0.37*)

Table of Comparison of the 3 Clustering Methods

Top: K-Means

Middle: K-Medoids

Bottom: Spectral Clustering

	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

4. Potential Challenges

- Traditionally, K-Means and K-Medoids work better with larger sample sizes and numbers of dimensions:
- The Wine Dataset has 178 instances and 13 dimensions, which is low for use of these methods
- The Travel Reviews Dataset (one we considered) has 980 instances, and, in preliminary testing, produced better results for K-Means and K-Medoids than Spectral Clustering

5. Conclusion

- With the Wine Dataset, K-Medoids clustering proved to more accurately cluster the data
- We were able to reject the null hypothesis with K-Medoids and Spectral Clustering
- Given the challenge of a relatively small dataset, the other methods may have proved more accurate had the population increased
- The key to any clustering analysis is to use multiple methods to find which will provide the best results

Thank you!
Any questions?