House of Representatives Congressional Voting Record

Kevin Kuo

Computer Science Department

Towson University

Towson, Maryland

kkuo1@students.towson.edu

*Abstract*— This report analyzes a data set of 16 critical votes based on the Congressional Quarterly Almanac made by all 435 Representatives from the House of Representatives. We will analyze this dataset to determine any association rules or trends based on how each representative voted.

Keywords— politics, voting, House of Representatives, association

# Introduction

This report studies the voting record of all 435 Representatives from the House of Representatives in the 2nd session of the 98th Congress.

## Dataset

This dataset, “1984 United States Congressional Voting Records Database”, was provided by Jeff Schlimmer on April 27, 1987. There are 17 attributes:

1. Class Name: 2 (democrat, republican)

2. handicapped-infants

3. water-project-cost-sharing

4. adoption-of-the-budget-resolution

5. physician-fee-freeze

6. el-salvador-aid

7. religious-groups-in-schools

8. anti-satellite-test-ban

9. aid-to-nicaraguan-contras

10. mx-missile

11. immigration

12. synfuels-corporation-cutback

13. education-spending

14. superfund-right-to-sue

15. crime

16. duty-free-exports

17. export-administration-act-south-africa

Each of these attributes with the exception of attribute #1 is categorized into either yay, nay, or no response.

## Objective

We will analyze this dataset to determine any frequent itemsets and mine for any association rules. This allow us to analyze voting trends.

# Methodology

This section will explain the two association methods used on the dataset and the approach that we have taken to mine the dataset and any pre-processing that was done on the dataset.

## Pre-processing

Minor pre-processing was needed on this data set. The format from the raw data file did not include header labels which makes the data not human readable. In order to make it human readable, it was converted to an Excel spreadsheet format using the ‘xlsx’. Also, each column is a separate vote but without a way to differential values (y, n, or ?) from those in another column we were not able to apply the association algorithms appropriately. I prefixed each value in the column with the column integer label.

## Apriori

K-means is a partitioning approaching to clustering. This data set is divided into subsets in a repetitive process and the seed points are computed to determine the centers of cluster. We will divide the data set into 2 and 3 clusters to see the algorithm’s clustering trends.

## Eclat

DBSCAN is a density based clustering analysis approach where two parameters are taken into account: maximum radius of neighborhood and minimum number of points in that group. This method discovers clusters in arbitrary shape in a spatial area with noise.

­

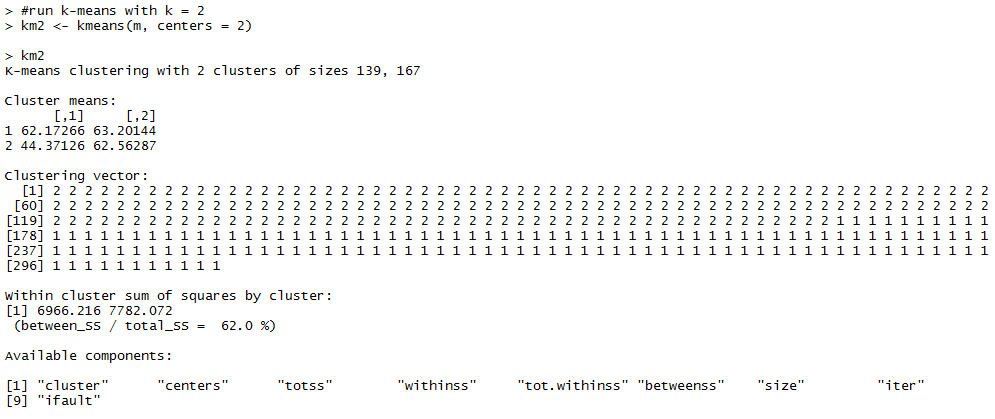
# Results

The dataset was not a good dataset in which to perform cluster analysis to begin with. Most clustering algorithms seemed to cluster based on age of patient. This is consistent with the knowledge that cancer generally afflicts the older population more than the younger population.

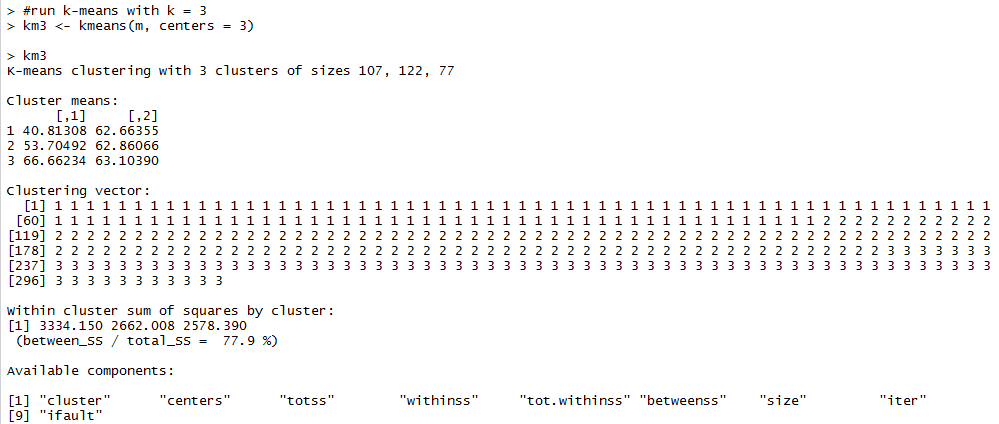
## K-means

K-means is may not be useful in this case because a disadvantage of K-means is that it not suitable in discovering clusters in non-convex shapes and sensitive to outliers. However, its graphical clustered data was more understandable than any other clustering algorithm used in this study.





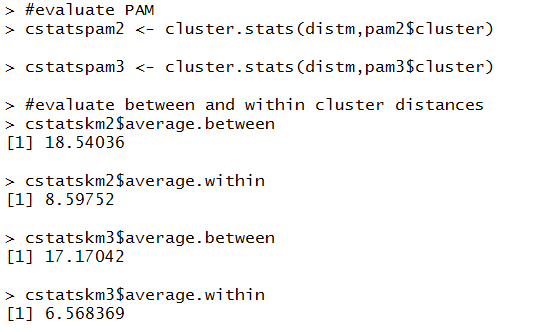




By clustering into 2 and 3 clusters, we can see that the k-means algorithm selected to cluster mostly on age of the patient. There is a minor trend in the centers of the k-means that the greater the age the greater the operation year.

Unfortunately the dataset is not larger which may include patients with a greater range of age and operation year. This is expected due to the possibility that this illness might affect older people more than younger adults.

### K-medoids



The k-medoids result did not vary greatly from the k-means result based on the respective centers.

## Linkage

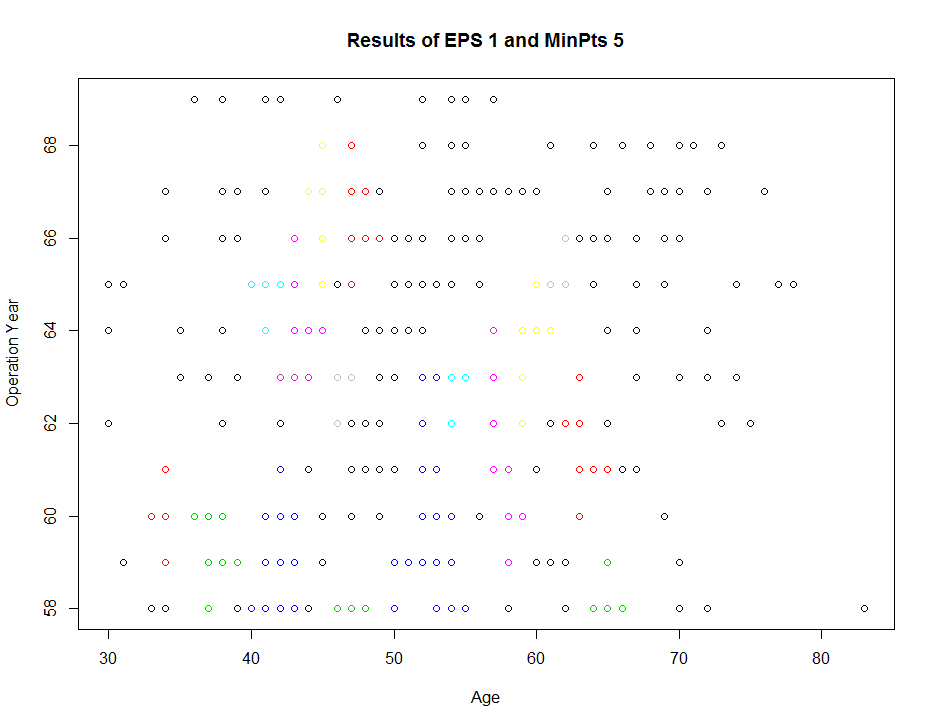
### Single Linkage

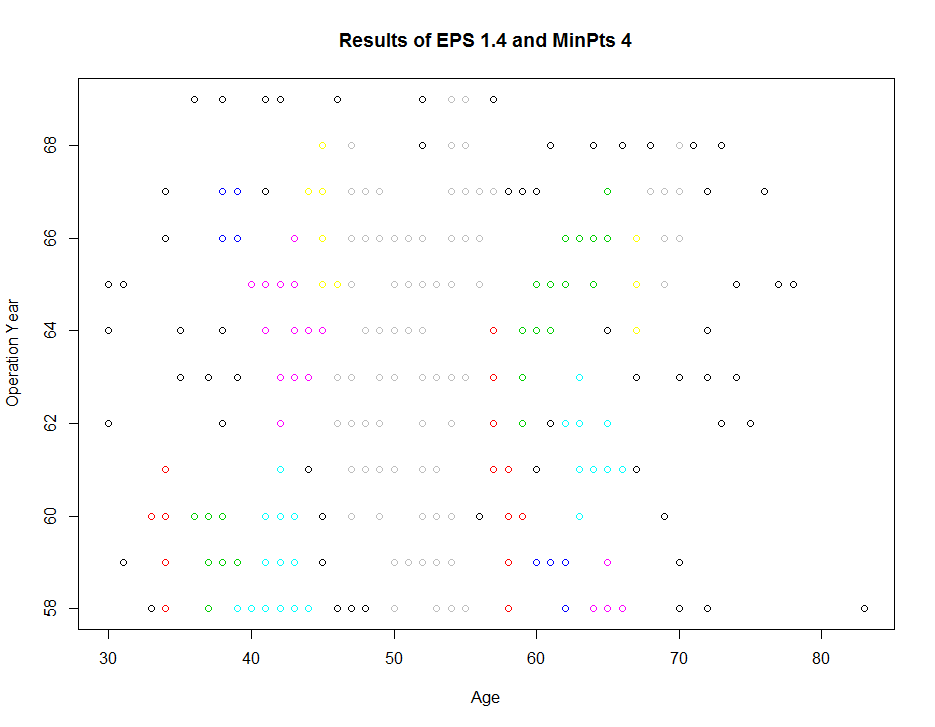


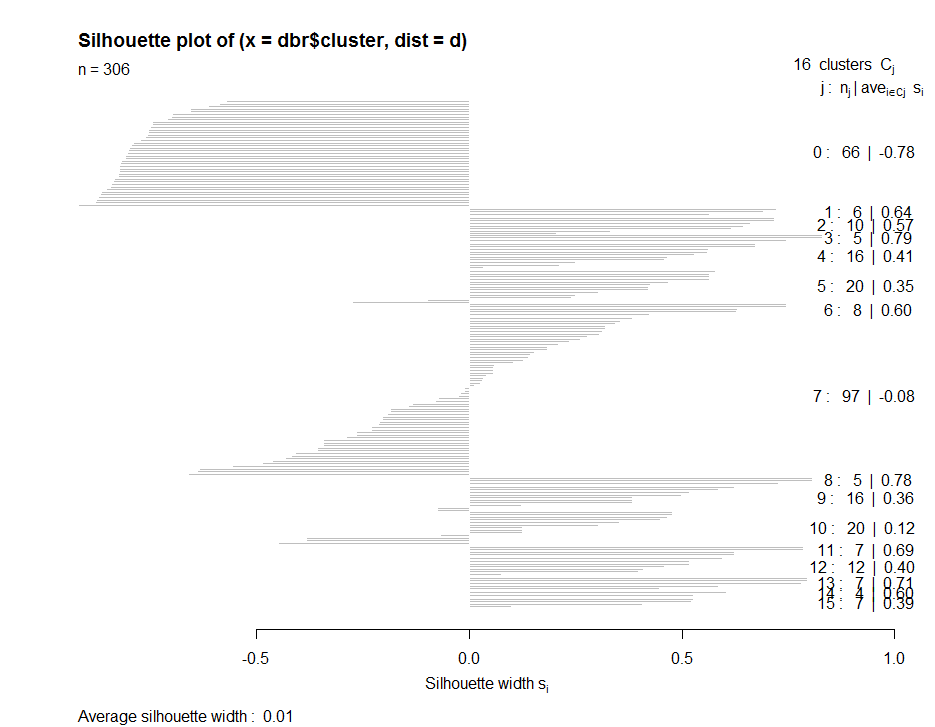
### Complete Linkage



## DBSCAN







DBSCAN proved to not cluster the data set well as it generated far too many clusters in what appears to be random spatial order.

# conclusion

All the clustering algorithms used in this analysis were ill suited for the data set provided. The k-means and k-medoid clustering algorithms produced the more appropriate clustering based on the data set.