Haberman’s Survival Data

Kevin Kuo

Computer Science Department

Towson University

Towson, Maryland

kkuo1@students.towson.edu

*Abstract*— This report analyzes a data set of patients who had undergone surgery for breast cancer between 1958 and 1970 at the University of Chicago Billing’s Hospital. We will analyze if there are similarities among patient age, year of operation, number of axillary nodes.

Keywords—breast cancer, University of Chicago Billings Hospital, axillary nodes, cluster

# Introduction

Breast cancer is a cancer that develops within the breast tissue. Breast cancer is a treatable type of cancer with high survival rates in developed countries if diagnosed early enough. We will perform cluster analysis on Haberman’s dataset to find similarities between the data by grouping similar data into clusters.

## Dataset

This dataset, “Haberman’s Survival Data”, was provided by Tjen-Sien Lim on March 4, 1999. There are 4 total attributes: age of patient at time of operation (numerical), patient’s year of operation (numerical), number of positive axillary nodes selected (numerical), and survival status of patient (class attribute). The last attribute is “1” if the patient survived 5 years or longer and “2” if the patient died within 5 years.

## Objective

We will analyze if there is a clustering pattern or relationship among patient age, year of operation, and number of axillary nodes. If there is a correlation among these variables, they may impact the patient’s post operation survival status.

# Methodology

This section will explain three different clustering algorithms used on this dataset and the approach that we have taken to mine the dataset and any pre-processing that was done on the dataset.

## Pre-processing

Pre-processing was not needed on this data set. We are comparing age of patient against operation year. This data set also did not have missing data values.



Above is the raw plot of the dataset.

## K-means

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.

### K-medoids

K-medoids addresses the outlier influence that plagues the K-means method. Instead of using the mean, k-medoids uses the median as the most centrally located object in a cluster. We will divide the data set into 2 and 3 clusters to see the algorithm’s clustering trends.

## Linkage

Linkage is a hierarchical clustering approach that is based upon a distance matrix. The downsides to this approach are its inability to undo or go back and

### Single Linkage

Single linkage is based on the smallest distance between an element in one cluster and an element in the other.

### Complete Linkage

Complete linkage is based on the largest distance between an element in one cluster and an element in the other.

## DBSCAN

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.

­

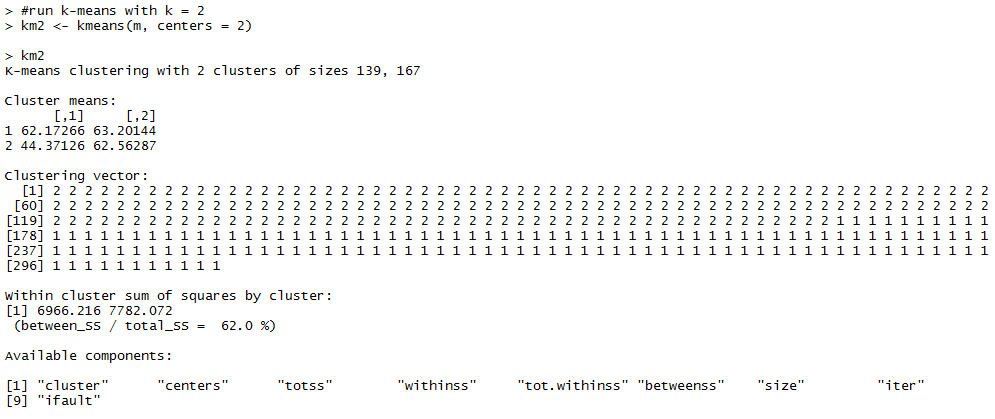
# Results

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.

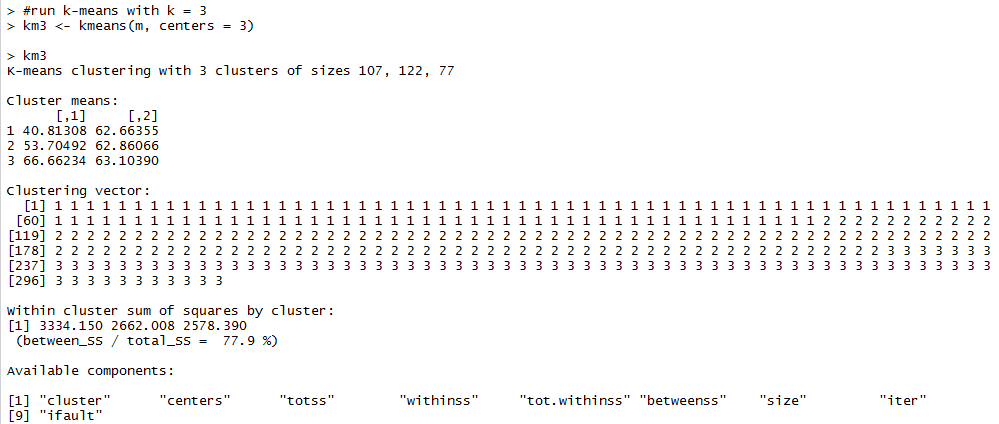
## K-means

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.





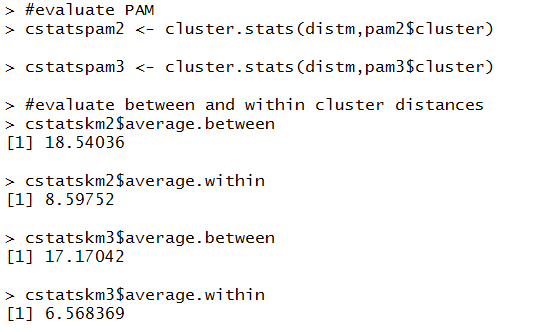




By clustering into 2 and 3 clusters, we can see that the k-means algorithm selected to cluster mostly on age of the patient. There is a minor trend in the centers of the k-means that the greater the age the greater the operation year.

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

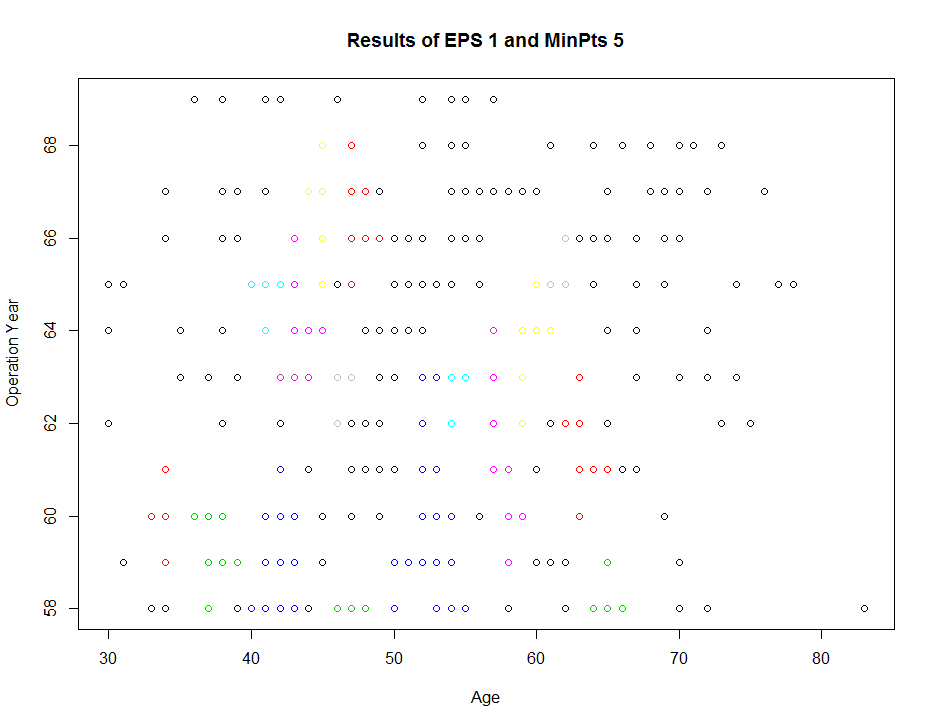
### Single Linkage

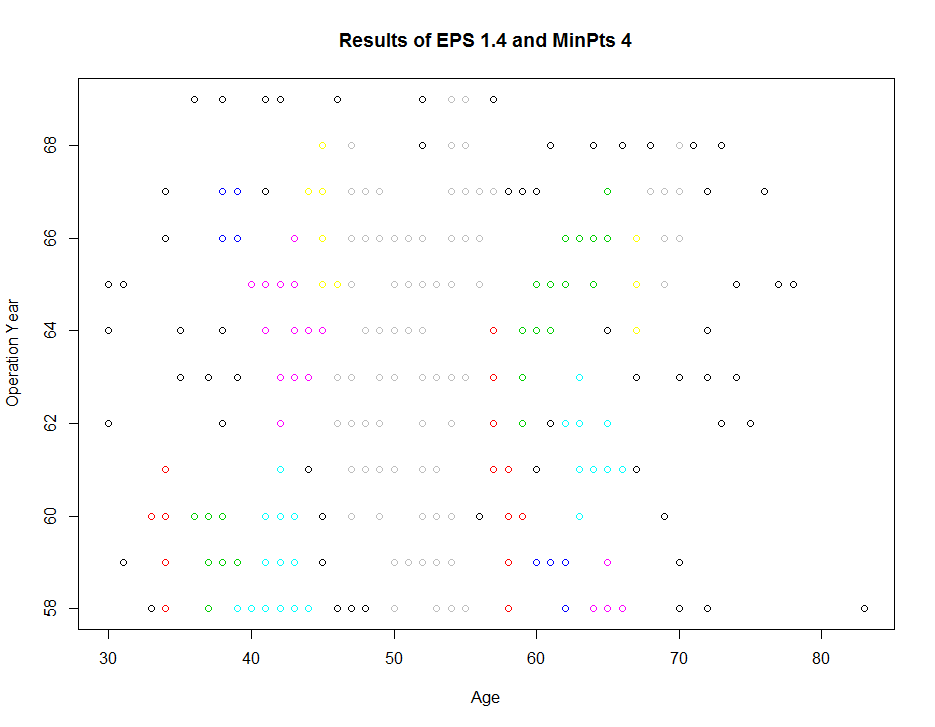


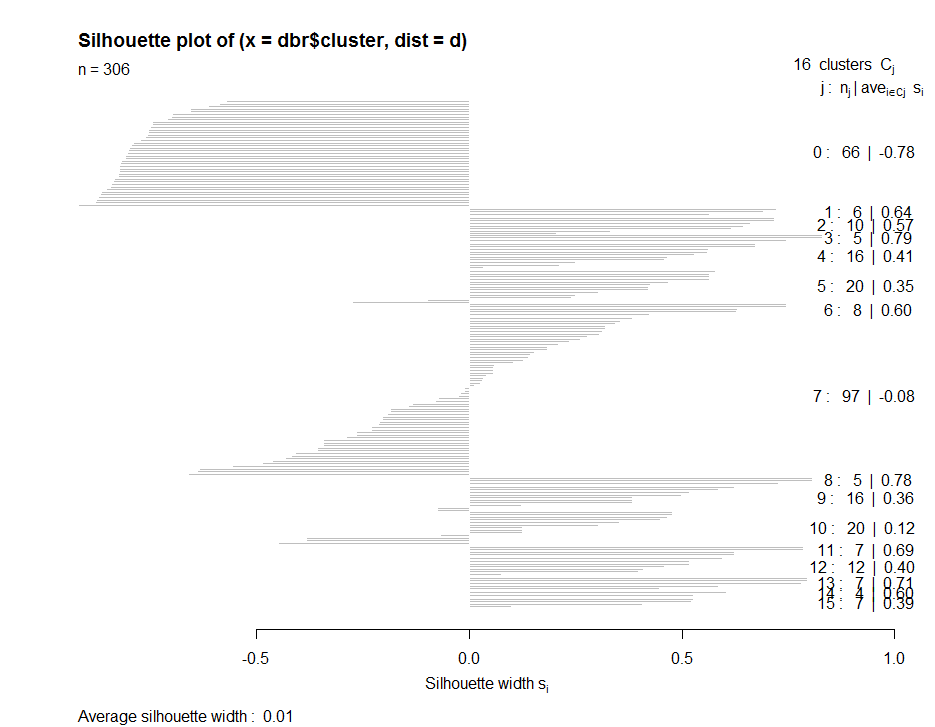
### Complete Linkage



## 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.