MSiA 431 Big Data Assignment 3

Julia Greenberger
Bryce Codell

A. Approach

1. Data Standardization

The first step in performing the k-means analysis was to normalize the data to ensure incommensurate units would not skew the cluster means. MapReduce was used to calculate the column minimums and maximums. Using these inputs the columns were scaled to a value from 0 to 1. The dimension of the data was then reduced by selecting six numeric columns of interest in MapReduce.

2. K-means

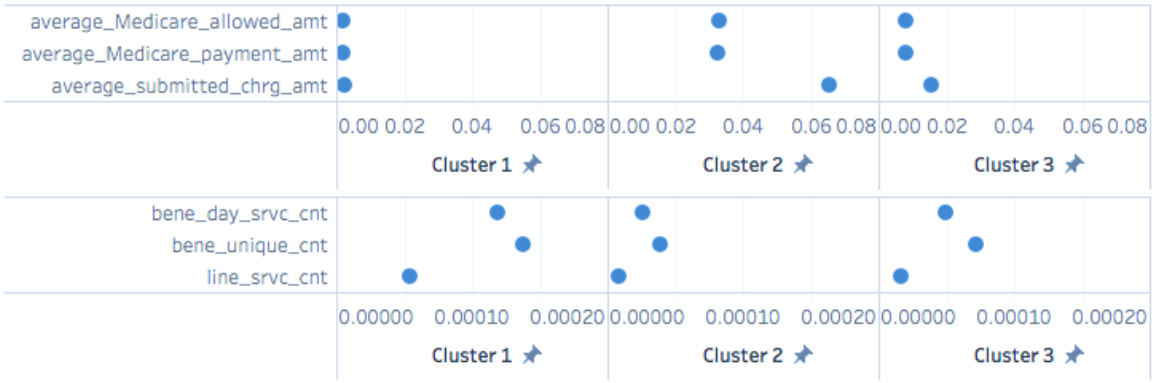
A Python wrapper script was written to call the Kmeans MapReduce job iteratively. Before performing the clustering, the cluster centers were initialized by randomly selecting three existing points in the data. The Python script then ran the Kmeans MapReduce job, which reads in these cluster centers in the `setup`. After an iteration completes, `getmerge` was used to concatenate the output and overwrite the cluster centroid input file.

B. Results

The data was clustered using the following six numeric features. These features provide insight into the number of services and dollar amount of Medicare benefits given by providers.

- ***line_srvc_cnt*** - Number of services provided.
- ***bene_unique_cnt*** - Number of distinct Medicare beneficiaries receiving the service.
- ***bene_day_srvc_cnt*** - Number of distinct Medicare beneficiary/per day services.
- ***average_Medicare_allowed_amt*** - Average of the Medicare allowed amount for the service.
- ***average_submitted_chrg_amt*** - Average of the charges that the provider submitted for the service.
- ***average_Medicare_payment_amt*** - Average amount that Medicare paid after deductible and coinsurance amounts have been deducted for the line item service.

K-means was run for 15 iterations to group the data into three clusters.



Cluster 1 contains providers with high numbers of services and beneficiaries but low charges and deductibles. Providers with low numbers of services and high charges comprise the second cluster. Cluster 3 contains providers with both low quantity of service and low charges. These clusters show distinction among providers who give more costly services to fewer people.