MSIA 431 Analytics for Big Data
Homework 3

Alejandro Avalos Mar
Ameer Khan

Introduction

Aiming to making the healthcare system more transparent, Centers for Medicare & Medicaid Services (CMS) has released a public dataset, which provides aggregate information on services provided by healthcare professionals to Medicare beneficiaries. This dataset includes demographic information of the healthcare providers, count of services provided, average charged amount by the professional and paid by Medicare, among others. k-Means clustering can be used to explore groups of professionals with similar characteristics, with the objective of finding trends and situations where some professionals are over or underperforming, or are practicing differently that other practitioners.

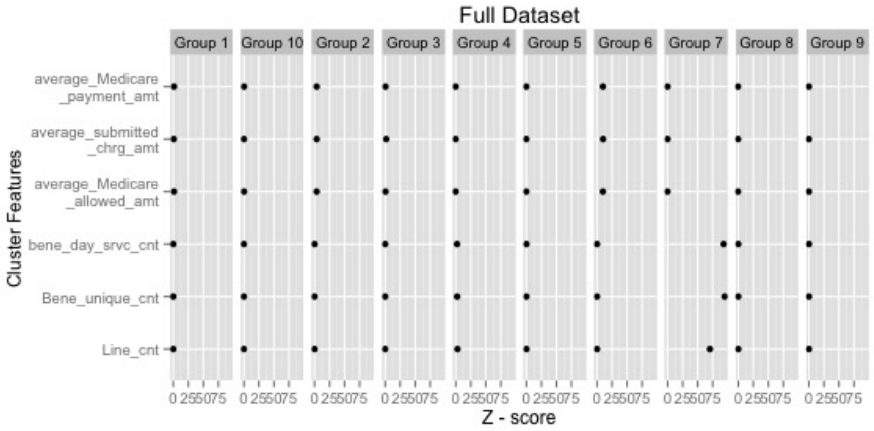
k-Means with Hadoop

The k-Means clustering for Medicare physician data was performed on a Hadoop cluster. A total of four separate jobs were used to perform the following operations to analyze the dataset: data subset creation, computing the mean and standard deviation of the subset, k-means clustering of the subset and assigning cluster labels to the data rows.

The data subset creation and cluster label assignment are map-only jobs, while the computation of mean and standard deviation uses a single reducer. The data subsets were created by filtering out the rows not relevant to the case study, but all columns were retained.

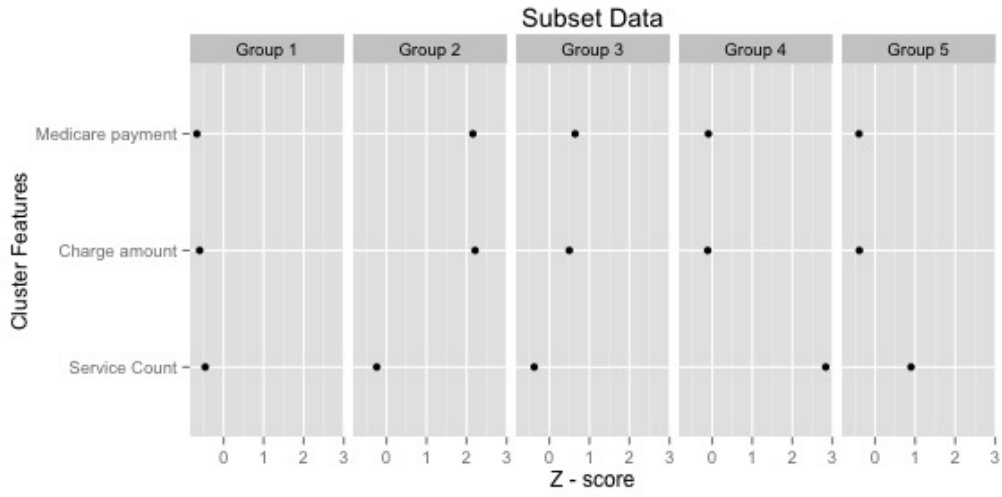
Column filtering was carried out in the k-Means routine, where only the desired columns are passed as input. The k-Means procedure also takes an initial selection of cluster centroids, the means and standard deviations of the features and the maximum number of iterations as input. The k-Means procedure iteratively assigns data points to clusters in the map step, computes partial means of each cluster for each mapper in the combine step, and recomputes the cluster centroids in the reduce step. This process repeats until convergence is achieved (the change in cluster centroids between successive iterations is not significant) or the maximum iteration limit has reached. Finally, a map-only cluster assignment job assigns the cluster numbers to data points and outputs the dataset.

The cluster centroids for k as 10 are visualized below for the full Medicare dataset. Clustering was performed on the z-scores of the data rows.



Case Study

We decided to explore Family Practice and Internal Medicine providers for Individual practitioners, as those two type of providers were among the highest services providers. To complete this analysis, a sub dataset was created performing a MapReduce job, filtering for these three characteristics. Additionally, outliers were eliminated from this dataset in the same MapReduce job, by removing data rows that had line service count more than 500 or average charged amount greater than $1,000. New means and standard deviations were obtained for this dataset for computing z-scores. Then, we standardized the data in the MapReduce job that assigned the practitioners to each cluster. The features used for clustering were the number of distinct Medicare beneficiary per day services, the average of the charges that the provider submitted for the service, and the average amount that Medicare paid after deductible and coinsurance amounts for the service provided to the beneficiary. The final cluster centroids obtained are illustrated below:

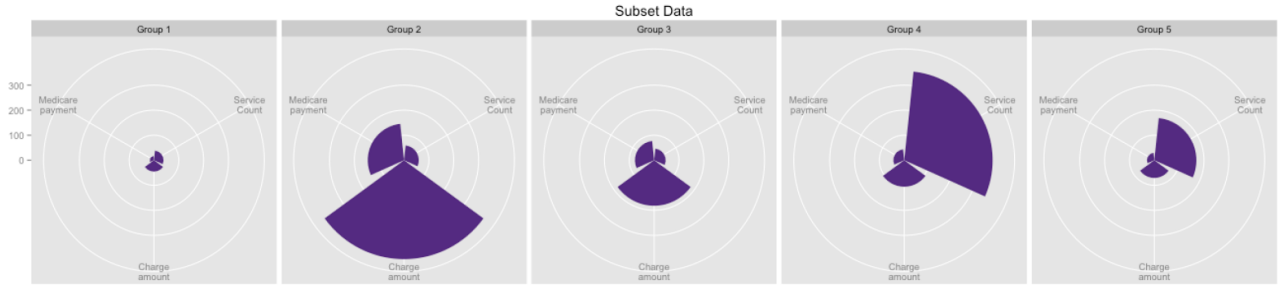


Note: The cluster features in the figure above are: Medicare payment for average_Medicare _payment_amt, Charged amount for average_submitted_chrg_amt and Service Count for bene_day_srvc_cnt.

Based on their cluster centroids, each cluster group can be described as:

1. **Group 1 -** Providers with the lowest amount of services provided, amount charged for the service, and amount that Medicare paid for the beneficiary on average
2. **Group 2 -** The highest amount charged on average for the services and amount paid by Medicare, although it had an extremely low amount of services provided.
3. **Group 3 -** Has low number of services provided, and medium amount of charges
4. **Group 4 -** Provides the most services, but charges very few compared to other groups
5. **Group 5 -** Similar to Group 4, but in a lesser scale

Additionally, one can explore the unstandardized the values for every feature with its respective centroids:



Group	Cluster Size	Beneficiary Day Service Count	Average Submitted Charged Amount ($)	Average Medicare Payment Amount ($)
1	910,536	38.65	45.22	17.66
2	187,454	59.24	396.34	145.86
3	525,094	47.60	179.95	77.04
4	143,355	357.31	108.38	44.90
5	265,925	171.85	67.30	28.31

From the table above, one can note that Group 1 is the cluster with the highest amount of observations. This is due to the right-skewness observed in the data. Furthermore, one can note that the average_Medicare_payment_amt is approximately 40% of the average_ submitted_chrg_amt for every group, which is relatively close to the actual coverage (48%) Medicare covers[1].

Additionally, cluster profiling was done to provide additional information of every cluster (see the Distribution of Place of Service Across Groups and Distribution of Provider type Within Group tables below):

[1] http://en.wikipedia.org/wiki/Medicare_%28United_States%29

Group	Distribution of Place of Service Across Groups	
	Facility	Non-Facility
1	13.93%	55.56%
2	21.80%	4.84%
3	47.86%	18.16%
4	06.98%	7.08%
5	09.41%	14.36%

Group	Distribution of Provider type Within Groups	
	Family Practice	Internal Medicine
1	53.82%	46.18%
2	25.74%	74.28%
3	39.62%	60.38%
4	39.87%	60.13%
5	45.27%	54.73%

In the Distribution of Provider type Within Group table above, one can notice the distribution of the Provider type across every cluster group (the row for every group equals to 1). All groups, but Group 1, are composed mainly by Internal Medicine Providers, which is dominated by Family Practice. This indicates that Family Practice doctors generally charge lower and have fewer daily visits compared to Internists. Additionally, we can see the distribution of the places of service across all of the clusters in the Distribution of Place of Service Across Groups. This table is read by column (the sum of the columns equals to 1). An interesting note is that from all the non-Facilities, 55.56% is concentrated in Group 1, and 47.86% of Facilities are concentrated in Group 3.

Conclusions and Next Steps

Groups 2 and 4 seem out of the ordinary; Group 2 seems that it's charging great amounts for few services, while Group 4 seems to be charging very few on average while providing above average amount of services. Group 2 seems to be performing high-cost services, which could

indicate a trend in the community, perhaps identifying a high-risk area, where this practitioners are providing their services.

Expanding on this work, one could continue profiling the groups found in this analysis: locations (State) and the credentials provider credentials. This could provide further insight into the reasons of why Group 2 provides few services but charges more, and the reasons Group 4 provides high amount of services but charges low compared to other groups. Additionally, looking into the geography could give us further insight on the discrepancy between the ~40% Medicare pays of the entire charge found in our clusters, and the average amount Medicare actually covers (48%), as Medicare's payment for a given service can change based on several factors, including geography.[2]

[2] https://questions.cms.gov/faq.php?id=5005&faqId=9928