IEMS 308 HW1 Clustering

Taige Hong

Jan 2019

**Executive Summary**

The Center for Medicare & Medicaid Services (CMS) is a federal organization that administrates the Medicare program, which provide insurance to American citizens. With its public data set, it should be able to identify which of the future cases may contain abnormal values in reported charge or the Medicare payment.

Clustering is a process that group data points based on similarities of the features. In the case of CMS, clustering is used to find the most similar case and thus finding out if payment is overpriced. By using the data of number of beneficiaries receiving the service and the submitted charge from the provider, CMS can utilize clustering to gain insight into the pattern of payment within similar cases.

After developing model, 5 clusters are formed, and the amount paid by Medicare for each cluster is analyzed. The model is capable of helping CMS capture fraud data with the formed clusters and the characteristics of the payment column.

**Problem Statement**

CMS’s data contains information on the number of distinct beneficiaries receiving service on a daily basis, the average charge provider submitted for the service, and amount actually paid after deductibles for the line services. We want to use clustering to match new entries to existing clusters to see if the payment is overvalued, according to the cluster it belongs to.

**Assumptions**

All entries are final-action, which means all claim adjustments have been resolved.

All existing amount paid to healthcare provider is reasonable. If there is inaccurate payment it may cause the clustering to be biased and therefore falsely classify new entries.

**Methodology**

The first step is to find out which columns are relevant to solving the problem. We pulled the relevant columns in and calculated the pairwise correlation. To make the clustering more accurate, almost-perfect correlated columns are condensed and outlier rows are omitted.

Since the size of the data is so huge, mini-batch k-means algorithm is used instead of k-means to raise speed of execution at the cost of certain accuracy. After running scenarios with number of clusters ranging from 2 to 10, the point where best signaled the stabilization of the decreasing speed of the objective function would be chosen as the best number of clusters.

**Analysis**

In the first step, reasonable features to include are:

* PLACE\_OF\_SERVICE - Identifies whether the place of service submitted on the claims is a facility (value of ‘F’) or non-facility (value of ‘O’). Omitted in further inspections due to unexpected dominant effect on the clusters. (See Figure 1)
* HCPCS\_DRUG\_INDICATOR - Identifies whether the HCPCS code for the specific service furnished by the provider is a HCPCS listed on the Medicare Part B Drug Average Sales Price (ASP) File. Omitted due to unbalanced response ratio.
* 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. Omitted due to highly correlated with last column.
* 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. Omitted due to highly correlated with last column.
* AVERAGE\_MEDICARE\_STANDARDIZED\_AMT - Average amount that Medicare paid after beneficiary deductible and coinsurance amounts have been deducted for the line item service and after standardization of the Medicare payment has been applied.

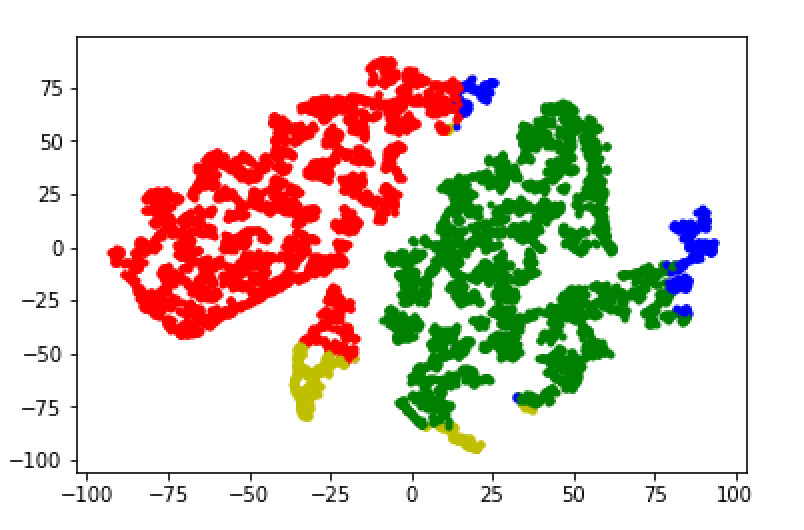


Figure 1: Including the dummy variable will make it a dominant factor in clustering

The features included in the clustering model includes: BENE\_UNIQUE\_CNT, BENE\_DAY\_SRVC\_CNT, AVERAGE\_SUBMITTED\_CHRG\_AMT, AVERAGE\_MEDICARE\_STANDARDIZED\_AMT. For each feature, outliers are excluded prior to the development of clustering model. After standardizing each column, mini-batch k-means algorithm is applied to the data with number of clusters ranging from 2 to 10, and the following relationship between objective value and number of clusters is recorded (Figure 2).

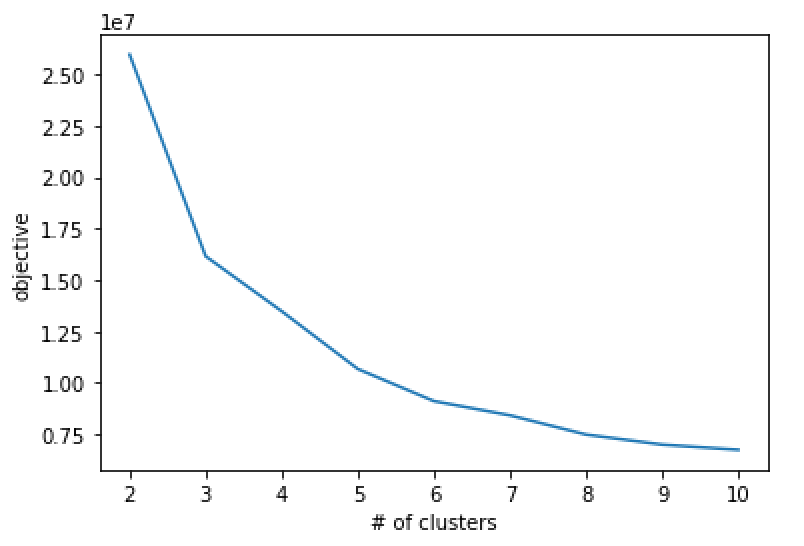


Figure 2: objective value vs number of clusters

Starting from the point when the number of clusters is 5, the shape of the curve starts to get smooth and horizontal, so 5 is chosen to be the number of clusters. The first 10,000 points are plotted on a 2-dimenstional space and the clustering effect is shown below (Figure 3):

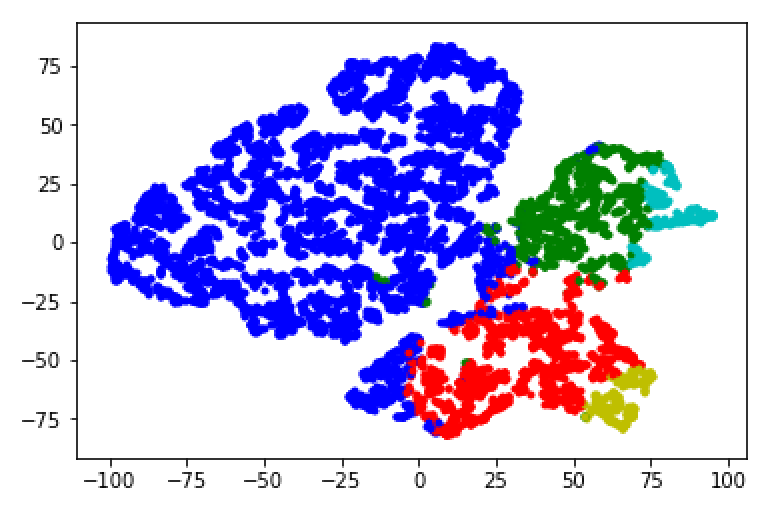


Figure 3: First 10,000 data entries divided into 5 clusters

The silhouette score is calculated to assess the quality of the clustering and calculating with all the data yields a memory error, so the first 10,000 rows are used in calculating the score. The outcome is ﻿0.4826, which is fairly good.

The next step is to analyze the pattern of amount paid by Medicare within each group, and the result is shown in the table below:

|  |  |  |
| --- | --- | --- |
|  | Mean | Std |
| Cluster 1 | ﻿49.7524 | ﻿39.2523 |
| Cluster 2 | 148.7966 | 61.4100 |
| Cluster 3 | 41.2544 | 31.6308 |
| Cluster 4 | ﻿490.5073 | ﻿263.0824 |
| Cluster 5 | ﻿51.4280 | ﻿52.2343 |

Table 1: Summary of AVERAGE\_MEDICARE\_STANDARDIZED\_AMT for each cluster

|  |  |  |  |
| --- | --- | --- | --- |
| Mean | Std | Min | Max |
| 0.7956 | 0.6986 | 0.0090 | 17.1751 |

Table 2: Distance to its centroid for all valid entries in the data

New data points are assigned to clusters based on whichever centroid it is most close to, and after assessment of the fitting (with respect to Table 2), a test will be performed on the payment column to see if it is overpriced.

**Conclusions**

It can be concluded that the clustering is successful, and the result can be used in future outlier prediction. A function is included for assessing future payments, which assesses whether the row is outlier (certain columns have unusual values) or the payment is overpriced (based on a 2 standard deviation standard).

**Next Steps**

The major next step is to find ways to include the one hot encoding for location and some other features in the model. For unknown reason, the model is performing rather poor when such dummies are included. Another path for further exploration is to use random forest to better account for the standard of clustering, which will make it easier to explain to actual users.