Cluster Analysis for Reducing City Crime Rates

Adel Ali Alkhaibari, Long Island University, Brooklyn, NY, *Student Member, IEEE,*
Ping-Tsai Chung, Long Island University, Brooklyn, NY, *Senior Member, IEEE*

Abstract— **Data analysis plays an indispensable role in the knowledge discovery process of extracting of interesting patterns or knowledge for understanding various phenomena or wide applications. Visual Data Mining is further presenting implicit but useful knowledge from large data sets using visualization techniques, to create visual images which aid in the understanding of complex, often massive representations of data. As the amount of data managed in a database increases, the need to simplify the vast amount of data also increases. Cluster analysis is the process of classifying a large group of data items into smaller groups that share the same or similar properties. In this paper, different Clustering algorithms such as K-Means clustering, agglomerative clustering were studied and applied to the Stop, Question and Frisk Report Database, City of New York, Police Department, NYPD, for analyzing the location of the crime and stopped people using the reason of stopped in order to reduce city crime rates. Our analytic and visual results revealed that the best clustering algorithm is K-Means algorithm, and its good features ensuring that the models are helpful.**

Keywords-component; clustering analysis, k-means clustering, agglomerative clustering, internal validation, stop and frisk.

I. INTRODUCTION

Data analysis plays an indispensable role in the knowledge discovery process of extracting of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge for understanding various phenomena or wide applications adapted retail, telecommunications, banking, fraud analysis, bio-data mining, stock market analysis, text mining, web mining, …,etc. Visual Data Mining is further presenting implicit but useful knowledge from large data sets using visualization techniques, to create visual images which aid in the understanding of complex, often massive representations of data. Through Visualization techniques, we could gain insight into an information space by mapping data onto graphical primitives; provide qualitative overview of large data sets; search for patterns, trends, structure, irregularities, relationships among data; help to find interesting regions and suitable parameters for further quantitative analysis [4].

As the amount of data managed in a database increases, the need to simplify the vast amount of data also increases. Cluster analysis is the process of classifying a large group of data items into smaller groups that share the same or similar properties [1][2][3][4][5][6].

K-Means clustering algorithm is an iterative clustering algorithm. It works as follows: Initially, we pick K random points as cluster centers. Alternately, we assign data points to closest cluster center, and then change the cluster center to the average of its assigned points; the iterations will be stopped when no points' assignments change. Note that Properties of K-Means algorithm guaranteed to converge in a finite number of iterations. The running time per iteration can be estimated as follows, it takes O(KN) time to assign data points to closest cluster center; it takes O(N) to change the cluster center to the average of its assigned points. In addition, the cluster converges. The critical factor of K-Means clustering is to specify the number of clusters and it is hard to determine what the best number of clusters is.

Agglomerative clustering is one category of hierarchical clustering algorithms; the other category of hierarchical clustering algorithms is called divisive clustering. Agglomerative clustering works as follows, it first merges very similar instances; and then it incrementally builds larger clusters out of smaller clusters. That is, an agglomerative clustering algorithm will initially maintain a set of clusters (i.e., each instance in its own cluster). Then repeatedly, it picks the two closest clusters, merge them into a new cluster. The algorithm will be stopped when there's only one cluster left. Note that it produces not one clustering, but a family of clustering represented by a dendrogram. There are a variety of Agglomerative clustering algorithms depending on the measurement of "closest" for clusters with multiple elements. There are three options: the measurement defined on the closest pair (Single-link clustering); the measurement defined on the farthest pair (Complete-link clustering); the measurement defined on the average of all pairs (Group-average clustering). Note that different measurement choices will result in different clustering behaviors.

In this paper, different clustering algorithms such as K-Means clustering, agglomerative clustering were studied and applied to the 2015 data records for the Stop, Question and Frisk Report Database, City of New York, Police Department, NYPD [7], for analyzing the location of the crime and stopped people using the reason of stopped, in order to reduce city crime rates. A police officer plays important role

in every city at all time, and every time a police department has information about crime, especially in large cities. According to "NEW YORK CIVIL LIBERTIES website, NYPD stop people all the time, and when they do so there are two ways to write a report. One of the ways is to report on stop-and-frisk data. This project on big data is going to deal with stop and frisk data. This program has applied to New York City. The main goal of this program is to reduce crimes rates in New York City. The program included New York City five boroughs, Manhattan, Bronx, Brooklyn, Queens, and Staten Island. Each borough is coextensive with a county of New York State.

The location of making graffiti crimes has been clustered for 24 hours, and the location have been clustered for the same crime. The time is set between 8.00 pm to 8.00 A.M. There is precious information and learning “hidden” in databases; and without automatic methods for extracting this information, it is virtually unattainable to mine for them. Throughout the years, many algorithms were created to extract what is called nuggets of knowledge from large sets of data. There are many different methodologies to approach this problem like classification, clustering, etc. Our cluster analysis was conducted for the location of the crime and stopped people using the reason of stopped. The cluster analysis performed using different clustering algorithms such as K-Means clustering, agglomerative clustering (Single-link clustering, Complete-link clustering, and Group-average clustering). The cluster analysis provided results about reasons for a stop; each method included making Graffiti on it. Furthermore, the following data have been clustered: the location for the women who were arrested; the location for teenagers less than 17 years old who were arrested for carrying contraband; the location of people less than 17 years old who were carrying pistols. A comparative report were summarized how to determine the proper number of clusters for each method. An internal validation measures is used to evaluate how well the results of a cluster analysis fit the data without reference to external information.

Note that the visual results were generated by the programs written by R-programming, which is open-source and free to use. It provides state-of-the-art algorithm, with more than 7000 extension packages on the Comprehensive R Archive Network, CRAN 2015. It creates beautiful visualization, as seen in the New York Times and The Economist [8]. Note that the data clustered based on the density regions (i.e., the Density-based spatial clustering of applications with noise (DBSCAN) model), clusters defined based on areas, which has higher density. The cluster in this model consists of objects connected by density and all objects within these objects range. Two key things built this model: the border points and the core points. The border points located on margin of the cluster, whereas the core points located on its inner region. The bad thing in DBSCAN model is the border is not perfectly organized, so some points not belonging to any clusters, which called outliers.

Our analytic and visual results revealed that the best clustering algorithm is K-Means algorithm, and its good features ensuring that the models are helpful. This paper is organized as follows. In Section II, problem description, and clustering analyses are provided. In Section III, Deep clustering analyses are provided for the location for the women who were arrested; the location for teenagers less than 17 years old who were arrested for carrying contraband; the location of people less than 17 years old who were carrying pistols. In Section IV, we conducted Internal Validation. Finally, a conclusion of this paper is provided in Section V.

II. PROBLEM DESCRIPTION AND CLUSTERING ANALYSES

In Section II, problem description, and clustering analyses are provided.

A. Problem Description

How to analyze the location of a crime and stopped people? Our goal is to analyze the location of a crime and stopped people by using different clustering algorithms such as *K*-Means clustering, agglomerative clustering were studied and applied to the 2015 data records for the Stop, Question and Frisk Report Database, City of New York, Police Department, NYPD [7], for analyzing the location of the crime and stopped people using the reason of stopped, in order to reduce city crime rates.

B. Clustering Analyses

First, we cluster the location of making graffiti crime for 24 hours and after that we cluster the location for the same crime but the time will be between 8.00 pm to 8.00 am. We obtained the xcoord and ycoord for the making graffiti crimes, which will be clustered to find out the most frequent locations for this crime. The plot for the location of making graffiti crimes is shown in Figure 1, we could see that our dataset not well separated or defined which making it hard to cluster.

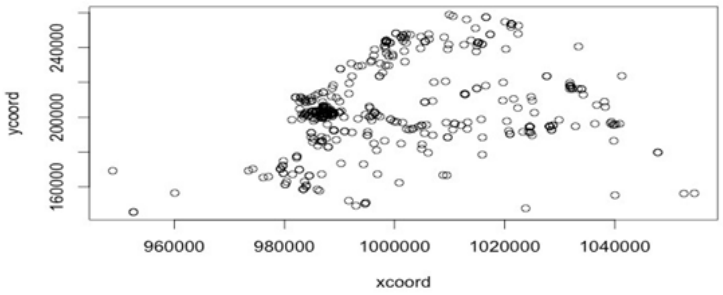

Figure 1. Plot (Making GRAFFITI crimes)

C. Using K-Means clustering (Making Graffiti crimes)

Now let us apply *K*-Means clustering algorithm. Before doing that, we must determine what the best number of clusters is. We used the Sum of squared error to determine the optimal number of clusters. The best number that we found is 4 clusters.

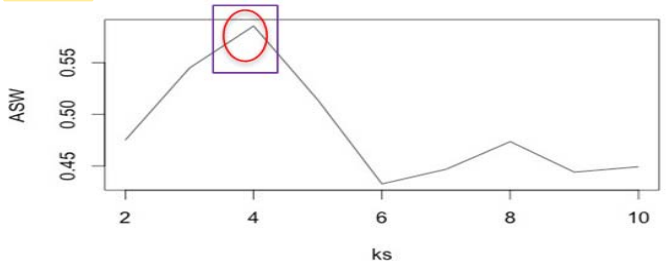

Figure 2. Criminal Possession of Controlled Substance

We can see and notice the suitable number of cluster is 4. *K*-means algorithm clustered and divided the location in plot graph with 4 colors (i.e., areas). Each area indicates to different cluster. We have 5 sides but the best clusters number that we got is 4. If the number of making graffiti crimes in Staten Island is larger than 3 we could get 5 clusters instead of 4.

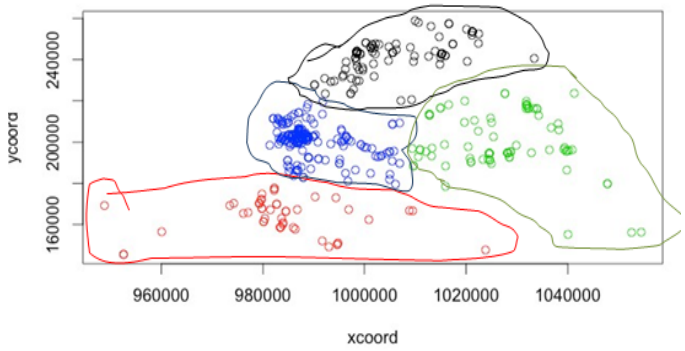

Figure 3. *K*-Means Clustering (Making Graffiti crimes)

As we can see below Figure 4, the map contains all location for making graffiti crime. We circled different clusters with different colors (i.e., areas).

Figure 4. A map contains all locations for making graffiti crimes shown in Figure 3.

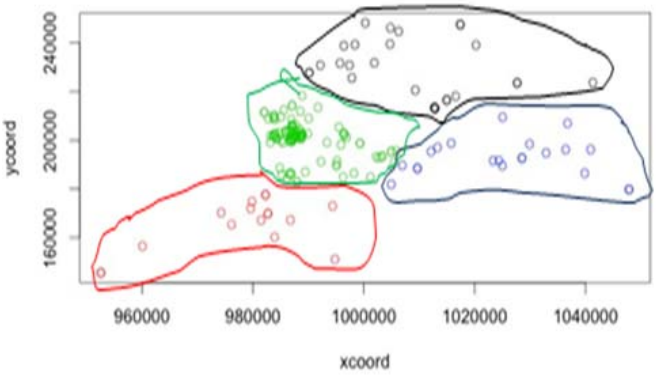

Figure 5. *K*-Means (Making Graffiti crimes 8:00 PM ~ 8:00 AM)

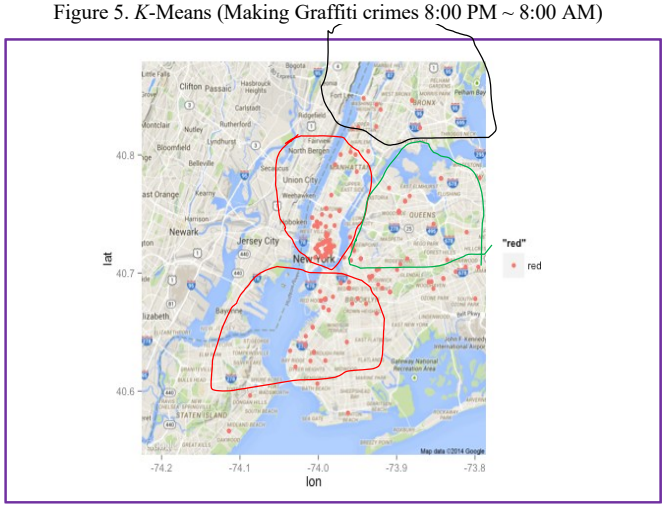

Figure 6. A map contains all locations for making graffiti crimes shown in Figure 5.

K-Means clustering (Making Graffiti 8.00 pm to 8.00 am): The next step we clustered the Making Graffiti crimes based on time. We made the time to start from 08.00 pm to 08.00 a.m. the total of making graffiti crimes based on this time is 171, whereas the total number in all dataset is 394 times. We can say most making graffiti crimes happened in the daytime. If we closer look to Figure 4 and Figure 6, we can notice that both data sets have clusters that are in the same location. Some of our clustering similar to the plot we have been discussed above where in Staten Island seems having the same clustering plot while in Manhattan and Bronx have overlap objects which doesn't match the previous plot above as same as in Queens and Brooklyn. The map contains all location of making graffiti crimes. We circled different clusters with different colors (i.e., areas).

D. Using Agglomerative clustering (Making Graffiti)

Next, we clustered the making graffiti crime location using the Group-average clustering for agglomerative clustering. We also applied the Complete-link clustering and Single-link clustering but we found that the Group-average clustering gives the best result among three methods.

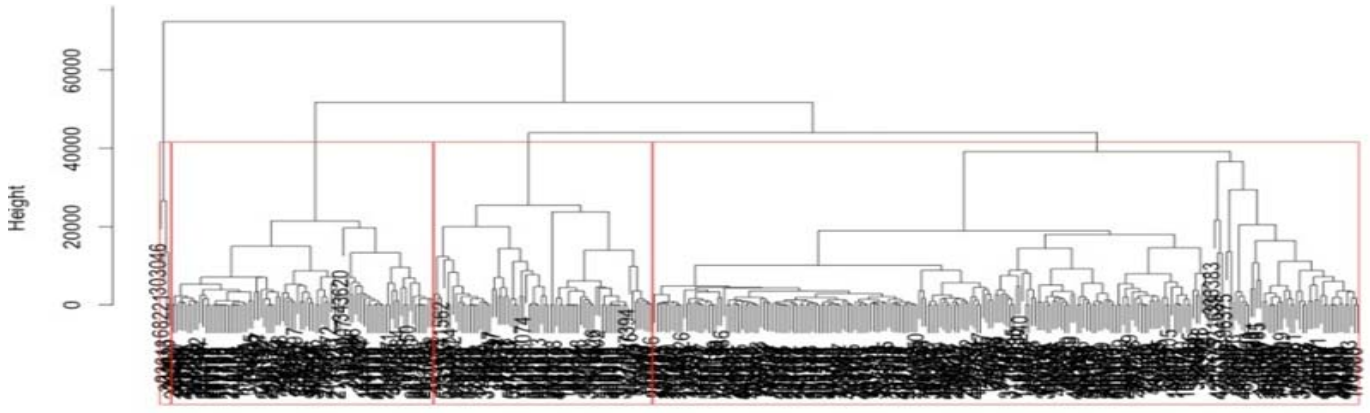

Figure 7. A dendrogram of Agglomerative clustering (Making Graffiti crimes)

Note that after plotting the clusters, we got the following results shown in Figure 7, As we can see, the boxes were not separated very well. All 4 cluster groups extended in the right side. It is clear that Agglomerative clustering clustered the objects in different way than K-Means clustering.

E. K-Means clustering (Reasons for Stop)

In this part, we are going to cluster stopped people using reasons for stop. Since we have a large data set, we sampled your data to produce a new subset that contains 800 Objects. After that, we determined the best clusters number, which are 10 clusters. We clustered stopped people using reasons for stop. Since this subset is binary, we got lift plot that describes how well a model ranks samples for one cluster. Fig 9 shows us the clusters from 1 to 5, whereas fig 10 shows us clusters from 6 to 10. The red line shows us the life when its value equals 1. Cluster 1 has two reasons of stop with value 40, which are they furtive movements and suspect acting as a lookout. The second cluster has suspicious bulge as the highest of reason of stop with value 5. The third cluster has the reason wearing clothes commonly used in a crime, as the highest one with lift value equals 30. The fourth one has actions indicative of a drug transaction, as the highest with lift value is 5 and furtive movement's reason is the highest reason in the fifth cluster with lift value equals 10. Cluster 6 we got two reasons with same lift value. The two reasons are actions of engaging in a violent crime and casing a victim or location with lift value equals 10. The carrying suspicious object reason has the maximum lift value in cluster 7 with value equal 2. Cluster 8 has the reason suspect acting as a lookout with highest lift value 8. Cluster 9 has three reasons that have the same lift value that equals 10. The histogram in Figure 8 and Figure 9 showed the relationship between different reasons. For example, the relationship between carrying suspicious object and other reasons is not strong since its value is not high.

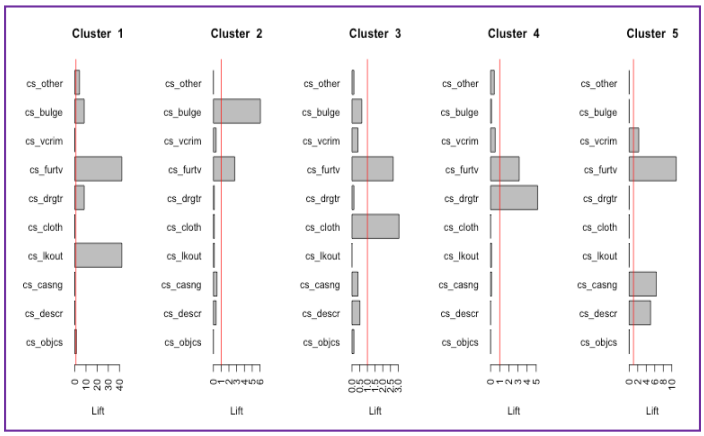

Figure 8. K-Means clustering (Reasons for Stop).

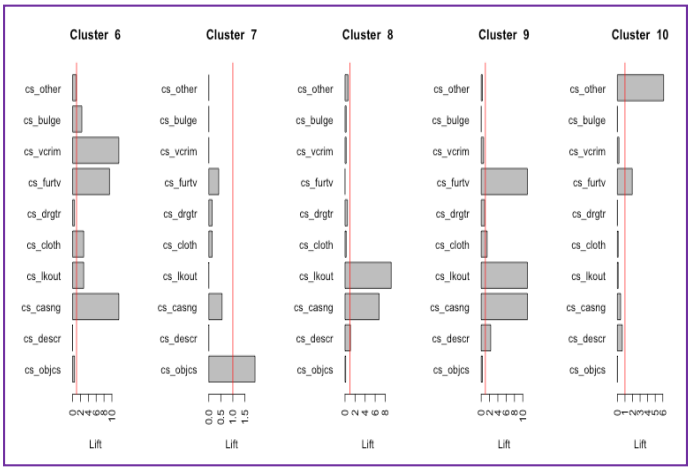

Figure 9. K-Means clustering (Reasons for Stop) (continued).

III. DEEP CLUSTERING ANALYSES

In Section III, Deep clustering analyses are provided for the location for the women who were arrested; the location for teenagers less than 17 years old who were arrested for carrying contraband; the location of people less than 17 years old who were carrying pistols.

A. What else can you use cluster analysis for in the data set?

We can use cluster analysis for the data set in more than one way, for example, we could cluster the location for the female who were arrested.

B. Carrying Contraband (K-Means)

Another example, we cluster the location for teenagers less than 17 years old who were arrested for carrying contraband as shown above. The best cluster number that we got is 4 after we used the average silhouette width. The total data points that we got are 389 points. We will cluster the data set using the K- Mean and Hierarchical clustering and after that, we are going to compare the location with people less than 17 years old who were carrying pistol. In the following Figure 10, we can see the locations where arrest has been made for people less 17 years old who were carrying contraband.

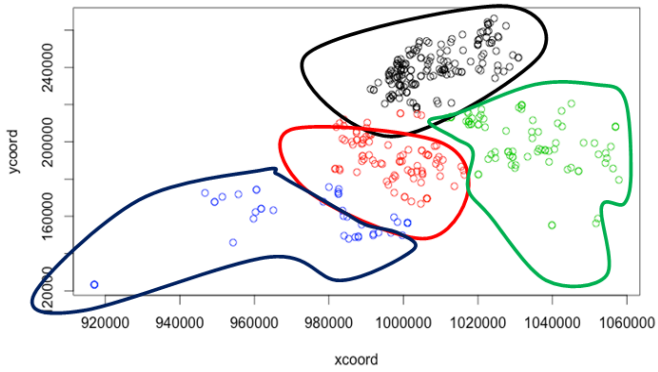

Figure 10. K-Means Clustering (Contraband Less 17 years old).

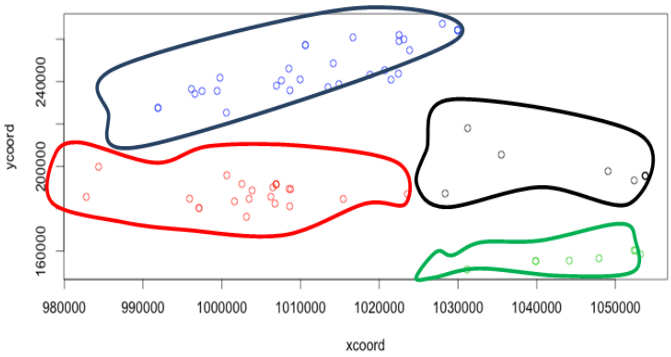

Figure 11. K-Means Clustering (Carrying Pistol Less 17 years old)

C. Carrying Pistol (K-Means)

We also clustered the location of people less than 17 years old who were carrying pistol. As shown below the number of people who were carrying pistol is 79. We found that the best clusters number is 4 (see Figure 11). We are going to make the K-Means clustering to compare it with the carrying contraband location. As we can see in the below map in Figure 12, the Staten Island city did not have any arrest for carrying pistol. If we compare it with carrying contraband, we can notice that there are some people who were stopped in Staten Island city.

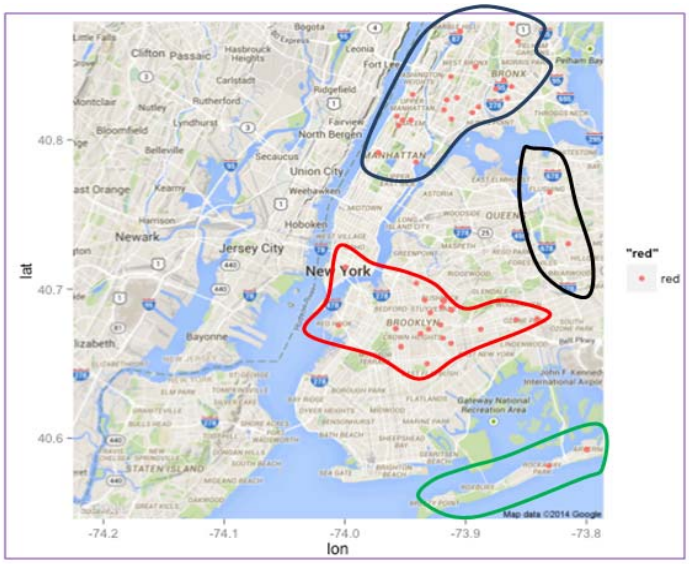

Figure 12. A map for Carrying Pistol (K-Means).

IV. INTERNAL VALIDATION

In Section IV, we conducted Internal Validation. Internal validation measures are used to evaluate how well the results of a cluster Analysis fit the data without reference to external information.

A. Internal Validation

The analytic results are shown as follows (as to how well they were clustered). We compared the clustering by first looking at the within. Clusters, which the low number considered better than high number. Since the K-Means clustering has the smaller number that means it is better than the Group-average agglomerative clustering. The average width indicates that how well the dataset clustered. If we get number that close to 1 means we have well clustered, whereas the negative numbers means we got some points that are closed to other group center than their group center. The K-Means value is 57%, while the Group-average agglomerative clustering is 51%. The best algorithm that we found is K-Means for this dataset.

Cluster#	K-Means clustering	Agglomerative clustering (Average)
Cluster 1	61%	47%
Cluster 2	41%	46%
Cluster 3	54%	65%
Cluster 4	64%	67%
Average	**57%**	**51%**

Table 1. A Comparative report for Internal Validation.

V. CONCLUSION

In this paper, we have applied different clustering techniques to get the best result that can help police officers to improve their work. Clustering is a popular. It is intended to identify several clusters discovered in databases using different measures of interestingness. First, we prepared the selected attributes to use them in creating the clusters. Second, we used some measures to determine the optimal number of clusters for each algorithm. Then, we created different clusters by using different methods. Finally, we used several visualization techniques to represent the clusters' profiles.

The best number of clusters is an important element that we should select carefully for clustering analysis. For clustering, two measures of cluster goodness or quality are used. One type of measure allows us to compare different sets of clusters is called an internal quality measure. We found that some of the models did not seem to have a lot of internal validity; however, they are reasonably accurate. Different Internal Validation techniques have been used to evaluate how well the results of a cluster analysis fit the data. We found that the best clustering algorithm is K-Means algorithm. In addition, we realized that good features play an important role in ensuring that the models are helpful.

References

[1] R. Xu, B. and D. Wunsch II, "Survey of Clustering Algorithms," IEEE Trans. On Neural Networks, VOL. 16, NO. 3, pp. 645–678, May 2005.

[2] M. Steinbach, G. Karypis, and V. Kumar, "A Comparison of Document Clustering Techniques," KDD Workshop on Text Mining, 2000.

[3] R. Ali, U. Ghani, and A. Saeed, "Data Clustering and Its Applications," Rudjer Boskovic Institute, 2001

[4] J. Han, M. Kamber and J. Pei and M. Kamber, Data Mining, Concepts and Technologies, 3rd Edition, The Morgan Kaufmann, , 2011.

[5] Sang C. Sug, Practical Applications of Data Mining, Jones & Bartlett, 2012.

[6] M. Kantardzic, Data Mining: Concepts, Models, Methods, and Algorithms, 2nd Edition, Wiley-IEEE Press, August 2011.

[7] The 2015 data records of the Stop, Question and Frisk Report Database, City of New York, Police Department, NYPD.

[8] The R Project for Statistical Computing, The R Foundation.