Clustering of Boarding Patterns in Public Transport

Tuğba Güler
Akdeniz University
Dept. of Computer Science Engineering
Antalya, Turkey
tugbaguler0441@gmail.com

Abstract— Data mining, data analysis powered unsupervised clustering algorithm is one of the fastest growing research areas due to the availability of large amount of data analysis. Extracts useful information based on the new optimization performance of the clustering algorithm. Clustering is the division of a dataset into number of meaningful subsets and is an unsupervised classification. Machine learning relies on extracting and mining invisible, meaningful data from vast amounts of data. Finding hidden patterns, clusters can be a supported unsupervised learning. K-means is one of the best unsupervised learning strategies among all partitioning-based clustering strategies.

Keywords— K-means clustering, clustering, clustering methods, unsupervised learning, center-based clusters, data mining, public transportation.

I. INTRODUCTION

Cluster analysis method is one of the main analytical methods in data mining, the clustering algorithm method will directly affect the clustering results. The main purpose of using cluster analysis is to find groups of objects such that objects in a group are similar or related to each other and different or unrelated from objects in other groups.

In this project, a data set in which "Line at certain time intervals", "Boarding Time" and "Passenger Numbers" information is given has been studied.

Ways of clustering boarding patterns were investigated. Based on the data, some primary analyzes were made and a strategy was developed to cluster them.

In order to develop a strategy, first of all, the data set should be examined in detail. Controls such as whether the data in the data set exhibit a homogeneous distribution in the data set should be made. Therefore, the data set was first examined in detail.

At the same time, while clustering, it was checked whether the data in the data set could be clustered with each other. As a result of the studies on the given data set, it was seen that the number of passengers could be clustered successfully. Thus, by calculating the sum of the number of passengers in terms of lines and boarding time, clustering was performed with the k-means clustering method. Among the clustering types, center-based clusters were used.

II. OBJECTIVE

In this project, ways of clustering boarding patterns were investigated. By making some primary analyzes of the given data, a strategy was tried to be developed to cluster them.

In the given data set, there is information about how many people use public transport daily. In addition, there is information about which lines are used and boarding time. The aim of this project is to relate these data to each other and to obtain meaningful results.

III. TYPES OF CLUSTERS [1]

Information on some clustering methods is given below. And among these techniques, the most suitable method for the given data set was determined as center-based clusters.

A. *Well-Separated Clusters*

A cluster can be an arrangement of points, so any point in a multi-cluster may be closest (or very similar) to each distinct point in the cluster, unlike any other point that is not in the cluster.

B. *Center-Based Clusters*

A cluster can be the placement of objects in a multi-cluster where an object is specified highest (more similar) to the "center" of the cluster rather than the middle of the other cluster. The center of a cluster is usually a centroid.

C. *Contiguous Clusters*

A cluster can be a location of points, so a point in a multi-cluster may be closest (or very similar) to at least one or additional distinct points in the cluster, unlike any other point not in the cluster.

D. *Density-Based Clusters*

A cluster may be a region of heavy points separated from different high-density regions by low-density regions.

IV. LITERATURE REVIEW

K. A. Abdul Nazeer et al. [2] proposes a k-means algorithm for different value sets of initial centroids, produces different sets.

The final cluster quality in the algorithm depends on the choice of initial centroids. The original k-averaging algorithm has two steps: first to determine the centroids first and second to assign data points to the nearest clusters and then recalculate the clustering mean.

Soumi Ghosh et al. [3] presents a comparative discussion of the two clustering algorithms. These are center-based K-Means and representative object-based FCM (Fuzzy C-Means) clustering algorithms. This discussion is based on performance evaluation of the efficiency of clustering output by applying these algorithms.

Binu Thomas et al. [4] gave a comparative analysis between k-means cluster algorithmic program and fuzzy cluster algorithmic program. The researcher also discusses that fuzzy cmeans algorithmic program tools-tools may be a partial-priority {based} cluster algorithmic program, whereas Fuzzy cmeans may be a non-partially based cluster algorithmic program. Fuzzy cmeans basically works in 2 ways. In the first method, cluster centers are calculated and in the second, information points are assigned to the cluster center calculated with the help of Euclidean distance. This method is like typical k-means with touch separation. In fuzzy means, the information item in the set is assigned a zero-to-one algorithmic program membership. A membership of 0 indicates that the information object is not a member of the set, while a membership indicates the degree to which this information represents a set. The problem faced by the fuzzy c-means algorithmic program is that the information points in each cluster are limited to one advertising the membership value. The algorithm faces disadvantages along with handling outliers. On the other hand, comparison with k-means shows that the fuzzy algorithmic program is economical in extracting hidden patterns and information from natural data with outliers.

Shafeeq et al. [5] present a modified K-means algorithm to improve the cluster quality and to fix the optimal number of clusters. As input number of clusters (K) given to the K-means algorithm by the user. But in the practical scenario, it is very difficult to fix the number of clusters in advance. The method proposed in this paper works for both the cases i.e. for the known number of clusters in advance as well as an unknown number of clusters. The user has the flexibility either to fix the number of clusters or input the minimum number of clusters required. The new cluster centers are computed by the algorithm by incrementing the cluster counter by one in each iteration until it satisfies the validity of cluster quality. This algorithm will overcome this problem by finding the optimal number of clusters on the run.

Junatao Wang et al. [6] propose an improved K-means algorithm using a noise data filter in this paper. The shortcomings of the traditional k-means clustering algorithm are overcome by this proposed algorithm. The algorithm develops density-based detection methods based on characteristics of noise data where the discovery and processing steps of the noise data are added to the original algorithm. By pre-processing the data to exclude these noise data before clustering data sets the cluster cohesion of the clustering results is improved significantly and the impact of noise data on the K-means algorithm is decreased effectively and the clustering results are more accurate.

Shi Na et al. [7] present the analysis of shortcomings of the standard k-means algorithm. As K-means algorithm must calculate the distance between each data object and all cluster centers in each iteration. This repetitive process effects the efficiency of the clustering algorithm. An improved k-means algorithm is proposed in this paper. A simple data structure is required to store some information in every iteration which is to be used in the next iteration. Computation of distance in each iteration is avoided by the proposed method and saves running time.

Kohei Arai et al. [8] have projected hierarchical k-means which mixes k-means and hierarchical algorithmic programs. The strategy executes k-means for a few mounted ranges of times then applies a hierarchical algorithmic program on centroids obtained as a result of executions of k-means. The centroids, therefore, obtained from the hierarchical algorithmic program are then used as initial centroids for k-means. However, authors have recommended that their technique works higher (in terms of speed) as compared to ancient k-means for advanced cluster task (large numbers of information set and lots of dimensional attributes)

T. Gonzalez et al. [9] technique picks up the initial center of mass arbitrarily and also the remaining centroids are selected because of the information that has the best minimum distance to the antecedently designated centroid. This technique was originally developed as a 2-approximation to k-center cluster drawback.

F.A. Ramadan et al. [10] proposes an economical increased k-means algorithmic rule to beat issues in existing k-means. Original means are known because of their ease, simplicity, speed of convergence, and flexibility to thin information. Despite its large number of benefits, it suffers from sure disadvantages. These drawbacks are the formatting of centroids, problem to converge to native minimum i.e updating of centroids until the native minimum isn't fount & execution of recurrent whereas loops. All these issues are handled by the projected k-means cluster algorithmic rule. The enhanced algorithmic rule first assigns datasets to their highest center of mass and so works out distance with different centroids. In the next step, the 2 distances are compared and if the new distance tiny is little than the previous distance then the data point is touched to the new cluster otherwise it is small then it's allotted to the same cluster. This method can save a great deal of your time and improve potency. This algorithmic rule uses 2 new functions. The first one is distance() perform that's accustomed work out the distance between every data point and its nearest cluster head. The second is distance new() perform accustomed work out the distance between data points and different remaining clusters.

Tajunisha et al. [11] Performance Analysis of k-Means with different initialization methods for high dimensional data.

Uses Principal Component Analysis (PCA) for dimension reduction and to find initial cluster centers. The variable with the highest Eigenvalue calculated using PCA is taken as the first principal component along which partitioning is done, based on which k subsets are formed and k median values are taken as initial k centers.

V. METHODOLOGY

Clustering is an important and fundamental concept of the data mining field used in various applications. In Clustering, data are divided into various classes. These classes represent some important features. This means classes are the container of similar behavior of objects.

The objects which behave or are closer to each other are grouped in one class and non-similar objects are grouped in a different class. Clustering is an unsupervised learning process.

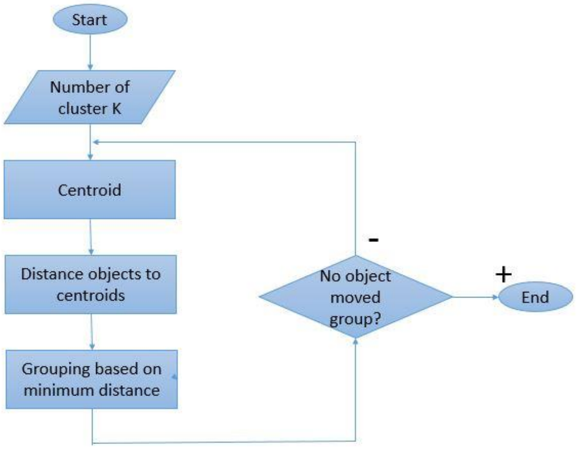
Start
Number of cluster K
Centroid
Distance objects to centroids
Grouping based on minimum distance
No object moved group?
-
+
End

Figure 1. K-means clustering algorithms

K-means clustering technique is a technique of clustering which is widely used. This algorithm is the most popular clustering tool that is used in scientific and industrial applications. It is a method of cluster analysis that aims to partition observations into k clusters in which each observation belongs to the cluster with the nearest mean.

K-means clustering: K-means clustering is an unsupervised clustering technique in which data points are given as input and without and predefined results it generates clustering results.

- The generic algorithm is very simple as presented in figure.1.
- Select K points as initial centroids.
- Repeat
- Form K cluster by assigning each point to its closest centroid.
- Recomputed the centroid of each cluster until the centroid does not change

It follows a simple and easy way of classifying a given dataset through a pre-fixed number of clusters, i.e., k number of clusters. The main idea is to define k centers, one for each cluster. These centroids should be placed logically as different locations cause different results. Therefore, it is a better choice to place them as far apart as possible. The next step is to take every point belonging to a given dataset and associate it with the nearest centroid. When there is no waiting spot, the first step is completed, and an early batch is made. At this point, it is necessary to recalculate k new centers as centers of clusters originating from the previous step. After these k new centers, a new binding must be made between the same data points and the nearest new center point. In this way, a loop is created. As a result of this cycle, k is continued until the centers of gravity change their positions step by step, and until no more changes are made, that is, the centroids no move.

VI. RESULTS

As a result of the studies on the given data set, it has been seen that the number of passengers can be clustered successfully. Thus, the sum of the number of passengers in terms of line and boarding time was calculated and clustering was done with the k-means clustering method. The graphs obtained as a result of this clustering are shown below.

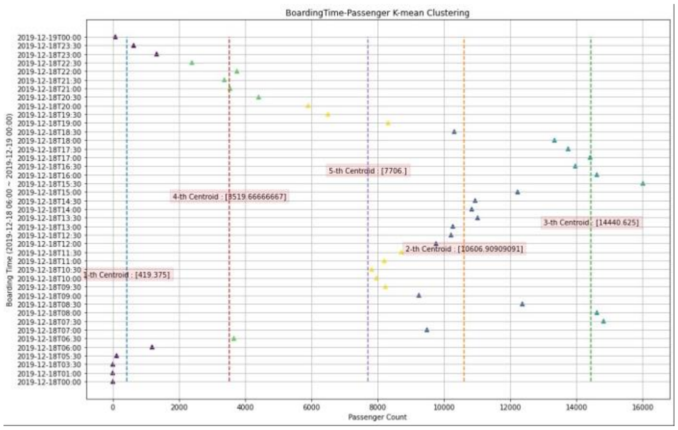

Figure 2. Boarding Time and Passenger K-Means Clustering

Figure 2 shows the clustering graph of boarding time and the number of passengers. When we examine the graph in detail; it is seen that the value of k is chosen as 5. And the centroid value of each cluster is shown in the graph.

While the number of passengers is very low at 00:00, 01:00, 03:30, 05:30 hours, it is seen that too many passengers use the bus at 15:30. The hour with the most passengers, that is, the most crowded, is at 15.30.

Other busy hours are 07:30, 08:00, 16:00, 16:30, 17:00. If a generalization is made by looking at this graph, it can be said that the number of passengers is much higher during commuting and departure times.

The third centroid part of the graph shows the time zones where the bus is used by the most passengers.

The first centroid is the one that shows the hours when the least number of passengers use the bus. When you look at the second centroid, you can see the frequently used bus times. For example, 13:30, 14:00, 14:30, 15:00.

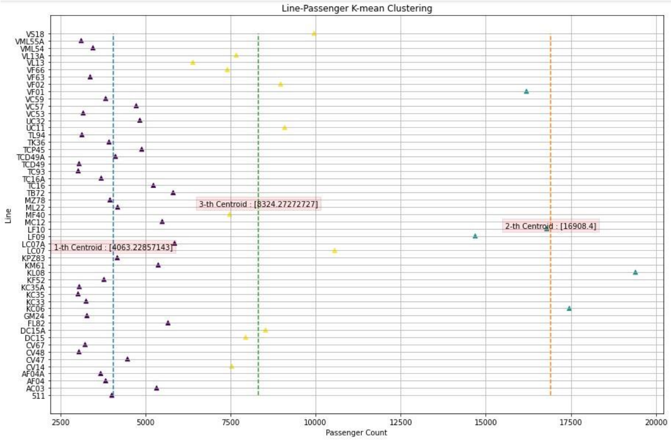

Figure 3. Line and Passenger K-Means Clustering Choosing K Value is 3

Figure 3 shows the clustering graph of line and number of passengers. In this graph, the k value is chosen as 3. Looking at this graph, the second centroid gives us information about the buses that carry the most passengers. The vehicle carrying the most passengers is the KL08. After that, the buses carrying the most passengers are KC06, LF10, VF01, and LF09 respectively.

Buses such as LC07, VS18, UC11, VF02, DC15A are located on the right side of the 3rd centroid, while DC15, VL13A, CV14, VF66, VL13 are located on the left side. Although the buses located on the right side of the center carry more passengers, when we consider their distance from the center, they are in the same cluster.

Other buses within the lines are in the first centroid. The number of passengers carried by TB72 and LC07A is very close to each other, almost the same, and they are the two vehicles carrying the highest number of passengers in this cluster.

Looking just to the right of the centerline shows KPZ83, ML22 and TCD49A. The values of KPZ83 and ML22 are much closer to each other, and it can be said that they carry almost the same number of passengers.

There are too many tools in the first centroid, and it is a little difficult to compare. In this case, clearer inferences can be made by updating the k values. The graphs obtained when the k value is updated to 5 and 7 in the graph below are shown below.

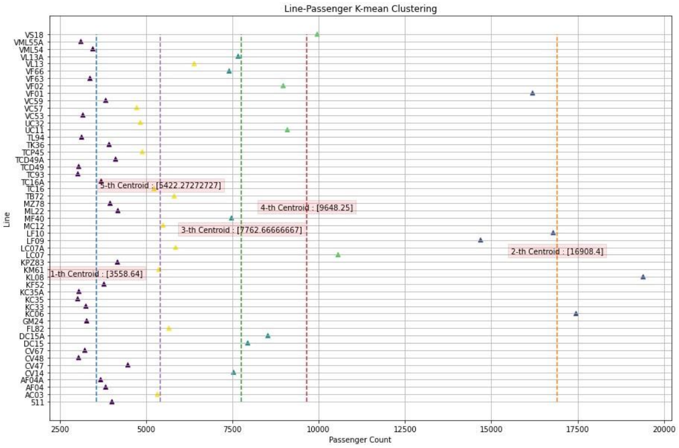

Figure 4. Line and Passenger K-Means Clustering Choosing K Value is 5

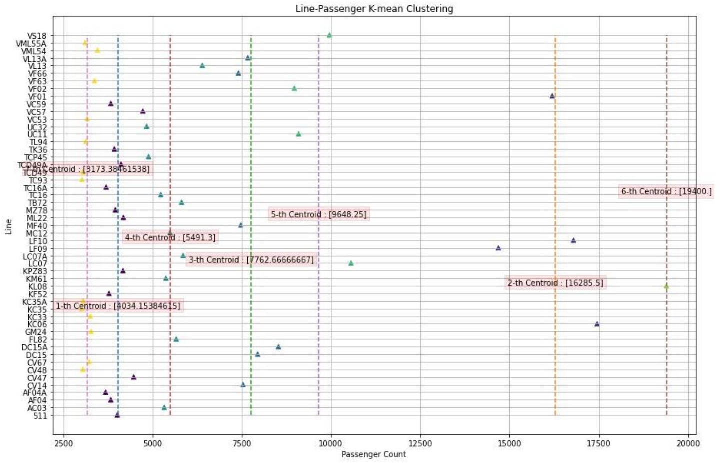

Figure 5. Line and Passenger K-Means Clustering Choosing K Value is 7

When we look at these graphs, the second centroid did not change when the k value was 3 and 5. When the k value is updated to 7, it is seen that KL08 is now in the sixth centroid.

If a general deduction is made from those graphs:

- It shows that the daily behavior of passengers in public transport depends on different time zones. Considering the boarding time with the highest number of passengers it can be said that passengers prefer public transportation when going to work and returning home from work.
- In addition, information can be obtained about how the travel habits of the passengers are related to the bus line. Data in the data set belong to the province of Antalya, located in the south of Turkey. If the route of the bus lines in Antalya is investigated, inferences can be made about the socioeconomic characteristics of the passengers. However, additional route research is needed for this inference.

VII. DISCUSSIONS

Research has been done on k-means and some data have been obtained.

K-means is one of the simplest algorithms that use unsupervised learning to solve known clustering problems. It works well with large datasets. However, K-Means also has disadvantages:

- They have a strong sensitivity to outliers and noise.
- It does not work well with non-circular cluster shape - several clusters and initial seed value need to be specified beforehand.
- Its ability to exceed the local optimum is low.

Despite these disadvantages, k-means also has advantages. Other clustering algorithms with better features tend to be more expensive. In this case, k-means becomes a great solution for pre-clustering by reducing the space to discrete smaller sub-areas where other clustering algorithms can be applied. K-means is the simplest method. All that needs to be done is to select "k" and run it several times. The smartest algorithms (especially the good ones) are much more difficult to implement efficiently and have many more parameters to adjust.

In this project, the k-means clustering method is used because it is easy and has a low cost. In much larger data sets or if the data set is enlarged by updating, clustering with k-means can be continued easily.

The k-means is a clustering method that can be applied to any data. In addition, k-means does not require prior information on new data and can explore data structures without information on data. Also, the search time for k-means is very short.

VIII. Conclusion

In this article, k-means clustering techniques and methods are reviewed. The K-means algorithm is a well-known and most used algorithm, and therefore, the need for further improvement in various parts of the algorithm has emerged. Choosing a center for outliers, empty sets, and datasets is quite a challenge. Therefore, more research was needed to focus on these mentioned issues. Various techniques related to this subject have been mentioned in the literature review area. Due to the increasing size of data day by day, they need more development.

This article attempts to review a substantial number of papers to deal with the current k-means algorithm. At the same time, by working on the given dataset, it shows that the k-means algorithm can be developed by choosing the center-based appropriately. It has been understood that k-means can be applied quite quickly and easily in exploring data structures and making inferences by dividing the data into subsets. Since there are no outliers in the data set, the clustering process is much easier. By examining the data set, it was seen that the number of passengers could be clustered successfully. By calculating the sum of the number of passengers in terms of line and boarding time, inferences were made by clustering the k-means with the clustering method. Center-based clusters were used among clustering types.

References

[1] Kapil Joshi, Himanshu Gupta, Prashant Chaudhary, Punit Sharma, “Survey on Different Enhanced K-Means Clustering Algorithm”, International Journal of Engineering Trends and Technology (IJETT) – Volume 27 Number 4 - September 2015.

[2] K. A. Abdul Nazeer, M. P. Sebastian,îImproving the Accuracy and Efficiency of the k-means Clustering Algorithm, Proceedings of the World Congress on Engineering 2009 Vol I WCE 2009, July 1 - 3, 2009, London, U.K.

[3] Soumi Ghosh, Sanjay Kumar Dubey, Comparative Analysis of K-Means and Fuzzy C-Means Algorithmsî, International Journal of Advanced Computer Science and Applications, Vol. 4, No.4, 2013

[4]. Raju G, Binu Thomas, Sonam Tobgay and Th. Shanta Kumar“Fuzzy Clustering Methods in Data Mining: A comparative Case Analysis” 2008 International Conference on advanced computer theory and engineering,2008 IEEE

[5] Shafeeq,A., Hareesha,K.,ìDynamic Clustering of Data with Modified K-Means Algorithm, International Conference on Information and Computer Networks, vol. 27 ,2012

[6] Junatao Wang, XiaolongSu,îAn Improved K-means Clustering Algorithm, Communication Software and Networks (ICCSN), 2011 IEEE 3rd International Conference on 27 may,2011 (pp. 44-46)

[7] Shi Na, Liu Xumin, Guan Yong, ìResearch on K-means Clustering Algorithm: An Improved Kmeans Clustering Algorithm, Intelligent Information Technology and Security Informatics,2010 IEEE Third International Symposium on 2-4 April, 2010(pp. 63-67)

[8]. Kohei Arai, Ali Ridho Barakbah, “Hierarchical K-means: an algorithm for centroids initialization for K-means”, Reports of the Faculty of Science and Engineering, Saga University, Vol. 36, No.1, 2007.

[9]. T. Gonzalez, “Clustering to minimize the maximum intercluster distance”. Theoretical Computer Science, Vol. 38,pp. 293–306, 1985.

[10] FAHIM, SALEM A.M, TORKEY F.A, RAMADAN M.A “An efficient enhanced k-means clustering algorithm” Journal of Zhejiang University SCIENCE A ISSN 1009-3095 (Print); ISSN 1862-1775 (Online)

[11] Tajunisha and Saravanan, “Performance Analysis of k-Means with different initialization methods for high dimensional data” International Journal of Artificial Intelligence & Applications (IJAIA), Vol.1, No.4, October 2010.