Fabricating Proficient Drive along by Employing K-Means Clustering Algorithm

Saurabh Anand
Software Engineer
J.C Penney

Pallavi Singh
Software Engineer
J.C Penney

Pooja N. Desai
Software Engineer
J.C Penney

ABSTRACT

During data analysis, often data needs to be grouped together based on similar looking or behaving. As the real world data features modulate with the Big data, where the data is unlabeled, the task of dividing the population or data points into a number of groups with similar points is of prime necessity. This method of identifying similar groups of data in a data set is called clustering. In simple words, the aim is to segregate groups with similar traits and assign them into clusters. This paper presents the importance of the K-means Clustering algorithm to understand the inner structure of the data to obtain the areas wherein based on the number of car rides booked in an area, optimum pickup point can be found using K-Means Clustering Algorithm.

General Terms

Bigdata

Keywords

Clustering, K-Means, Big Data, Pickup optimization

1. INTRODUCTION

Decision making on most of the data involves data representation (Features and Similarity) and learning either by classification of labeled data or clustering by unlabeled data. Most Big Data have unlabeled data. Clustering involves given a collection of unlabeled object, find the meaningful groups. However, clustering involves many challenges such as how to measure similarity, number of clusters, cluster validity, and outliers. Data clustering in simple terms is defined as organizing a collection of n objects into a partition or hierarchy (nested partitions). In 2011, data clustering returned 6100 hits for 2011(google scholar).

Clustering is the key to Big Data problem:

1) Not feasible to “label” large collection of objects
2) No prior knowledge of the number and nature of groups (clusters) in data
3) Clusters may evolve over time and Clustering provides efficient browsing, search, recommendation, and organization of data.There are various clustering algorithms like K-Means [1], Kernel K-Means, Nearest Neighbour and much more but no clustering algorithm is optimal. One of the algorithms discussed in this paper is K-Means clustering algorithm.

1.1 K-Means Algorithm:

K-means clustering comes under unsupervised learning, which is used when the data is unlabeled (i.e., data which doesn’t have a defined specific group).

The main aim of this algorithm is to find groups or clusters in the data, with the number of groups or clusters represented by the variable K based on a center point for each of the clusters. The idea is that the cluster center will naturally migrate to the center of its members given the dataset over multiple iterations as the cluster are recentered based on its point memberships. After successfully applying K-Means algorithm the centroid for all the clusters and data can be labeled according to using these centroids. Each data point is attached to a single clustering group.

Clustering looks at the data which can be grouped together. The choosing of K clusters depends on labels to which each data can belong. Each of cluster centroid represents that clustering group. By determining the centroid feature weight, qualitative determination of what kind of data each group cluster represents can be done.
The K-means algorithm employs continuous refinement to produce a final result data set. The algorithm takes input such as the number of clusters K and the data set to be categorized into similar groups or clusters. The dataset consists of features for each data point. The algorithms initially start by estimating K centroids, which can be randomly generated or randomly selected from the given data set. The algorithms then use below two steps in the iteration:

1. Data assigning step:
In this step, data points are assigned to a centroid by calculating the squared Euclidean distance which gives the nearest distant centroid for a data point. Each centroid defines one of the clusters. More formally, if *ci* is the collection of centroids in set *C*, then each data point *d* is assigned to a cluster based on below formula:

$$\underset{c_i \in C}{\arg\min}\ dist(c_i, x)^2$$

where $dist(\cdot)$ is the standard (L_2) Euclidean distance. Let the set of data point assignments for each i^{th} cluster centroid be S_i.

2. Centroid update step:
In this step, the centroids are computed again after data assignment is over. This is done by taking the mean of all data points assigned to a particular centroid's cluster for each centroid as shown in the formula below.

$$c_i = \frac{1}{|S_i|}\sum_{x_i \in S_i} x_i$$

The algorithm continues to iterate between steps one and two until a stopping condition is met (i.e., no data points change clusters, the sum of the distances is minimized, or some maximum number of iterations is reached).

This algorithm guarantees convergence to a result. The result can be a local optimum (i.e. not necessarily the best outcome which can be achieved), which means if the Algorithm is evaluated continuously, after each iteration the subsequent runs may give better data clustering or grouping when the algorithm is started with random centroids initially.

Choosing K

The algorithm described above finds the clusters and dataset labels for a particular pre-chosen K. To find the number of clusters in the data, the user needs to run the K-means clustering algorithm for a range of K values and compare the results. In general, there is no method for determining the exact value of K, but an accurate estimate can be obtained using the following techniques.

One of the metrics that is commonly used to compare results across different values of K is the mean distance between data points and their cluster centroid. Since increasing the number of clusters will always reduce the distance to data points, increasing K will always decrease this metric, to the extreme of reaching zero when K is the same as the number of data points. Thus, this metric cannot be used as the sole target. Instead, mean distance to the centroid as a function of K is plotted and the "Elbow Point" [2] as shown in Fig1 below, where the rate of decrease sharply shifts, can be used to roughly determine K.

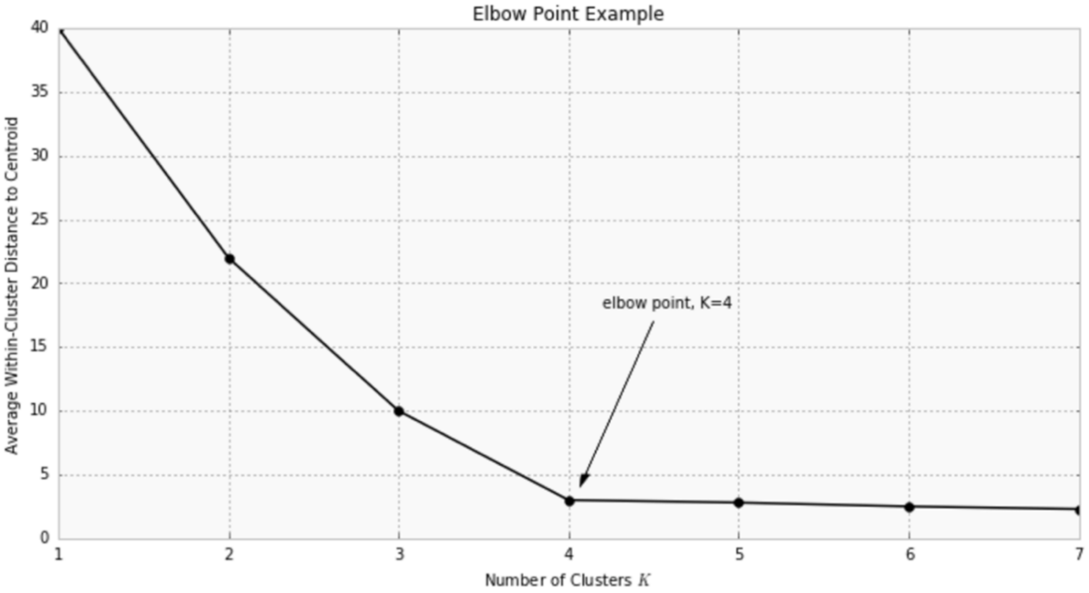

Fig 1: Determination of Elbow Point

A number of other techniques exist for validating K, including cross-validation [3], information criteria [4], the information theoretic jump method [5], the silhouette method [6], and the G-means algorithm. In addition, monitoring the distribution of data points across groups provides insight into how the algorithm is splitting the data for each K.

Why is the K-Means algorithm popular?

Having presented the K-Means algorithm, let's now briefly discuss its characteristics in more details. Why is K-Means popular? The main reason is that it is a very simple algorithm. It is easy to implement and also easy to understand.

Some drawbacks of the K-Means algorithms is that the final result depends on how the initial prototypes are selected randomly. If lucky, nice clusters can be found that somewhat makes sense. However, if the initial points are not chosen in an appropriate way, the result may not be meaningful. There exists various solutions to this problem such as running K-Means several times, and then to choose the best set of clusters.

Another limitation of K-Means is that the user must explicitly specify the number of clusters to be found (the K parameter). But finding the best value for the K parameter may require trying several values. To address these limitations, various extensions of K-Means have been proposed. Besides, many other clustering algorithms have been proposed.

2. BACKGROUND STUDY

The implementation of k-means clustering is easily done even on large datasets. While using heuristics like Lloyd's algorithm [7], it can be easily achieved by k-means clustering. The successful implementation of k-means can be observed in various fields like computer vision, geostatistics, agriculture or even market segmentation. The general use of k-means is to find a starting configuration or as a preprocessing step.

The k-means algorithm uses cluster analysis for the partition the input data set into k partitions. The k-means is one among the simplest clustering algorithm. Various methods have been implemented with multiple methods to instantiate the center. Amongst the major clustering approaches, the sum-of-squares criterion is prominently used and K-means is the name proposed to it and is well known to all. This algorithm was proposed by several scientists in different forms and in different assumptions. Researchers are analyzing new methods which are more efficient than the existing one and brings better-segmented result. The k-means algorithm was introduced by Stuart Lloyd as a technique for pulse-code modulation. The credited for its real-time uses is attributed to James MacQueen by idea which goes back to Hugo Steinhaus.

3. PROBLEM STATEMENT

Currently, a lot of problems is faced by industries which provide transport services to people through cabs and other vehicles on roads. One of the challenges faced by the management of these companies is optimizing the cost of deploying and making available cab services to people especially in big cities where geographical area is huge. A service provider is faced with the problem of a limited number of cabs to cover the whole city with so many people booking cab every moment and pick up points are at random locations. The service provider needs to manage all the bookings with a limited number of cabs and find the best way to deploy car at best feasible and optimal location in the city to cover all the current bookings at a point of time.

4. PROPOSED SOLUTION

Consider a person using a mobile app for booking a cab in his locality, the mobile app internally uses Google Map APIs to get the Latitude and Longitude of the location. The Lat-Long locations can be converted to X,Y coordinates in a plane with some reference point as Origin of the Plane. Now Each of the bookings represents a location on XY Plane. Suppose a total of K cars is available at any moment for booking. K-means clustering algorithm can be employed to find the optimal geographical location on google maps where the individual cars can be positioned to cover all the bookings at that moment.

5. DEMONSTRATION TO EVALUATE PERFORMANCE

For the experimental purpose, initially (0,0) (assume origin) is taken as some known location in Google Maps and all other locations on the map are derived with reference to this position on Google Maps. Suppose 10 cars needs to be deployed to start with in a locality and 50 booking locations with reference to the origin. As K-means clustering is run on the 50 location (X,Y) points as input data along with a total number of clusters as the number of cars (10 in this case) were fed to the code, optimal clustering patterns based on the Euclidean distances to clusters are found. The K-means algorithm finally gave 10 cluster points when there was no further change in centroids of individual clusters. The 10 cluster points finally obtained are the optimal geographical location to deploy the 10 cars for optimally providing the cab service to all the people at a time. Also as the number of clusters (here number of cars in this case) increased, the distribution also became less dense as the locations points which were densely grouped in a single cluster were able to migrate to intermediate clusters which were not there before.

EXPERIMENTAL DATA

SET 1

Number of Position Points (customer pickup coordinates in XY plane): 5

{1.0, 1.0} – Customer 1
{1.5, 2.0} - Customer 2
{4.0, 1.2} - Customer 3
{5.2, 1.0} - Customer 4
{7.5, 5.0} - Customer 5

Number of Clusters to be formed: 4

Resultant Individual cluster centers:

Centroids of cluster initialized at: (1.0, 2.0) (3.0, 3.0) (4.0, 4.0) (6.0,7.0)

Resultant Individual cluster centers:(1.0, 1.0) (4.0, 1.0) (4.0, 4.0) (7.0, 5.0)

Cluster 0 includes:(1.0, 1.0) (1.5, 2.0)

Cluster 1 includes:(4.0, 1.2) (5.2, 1.0)

Cluster 2 includes: NONE

Cluster 3 includes: (7.5, 5.0)

Graphical Analysis

The graph below in Fig 2 and Fig 3 shows respectively the results before (Initial cluster centers initialized in k-means algorithm to begin with) and after applying k-means (when k-means Algorithm completed its execution on the dataset)

Initial analysis using centroids

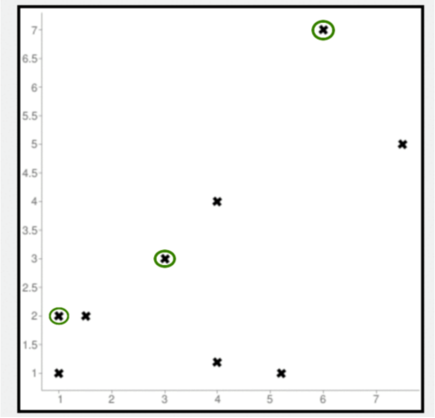

Fig 2: Initial Cluster Centers

Visually, you can see that the *K*-means algorithm splits the 4 groups based on the distance feature. Each cluster centroid is marked with a circle and each cluster is marked with different color.
Red points depict Cluster 0

Yellow points depict Cluster 1

Blue points depict Cluster 3

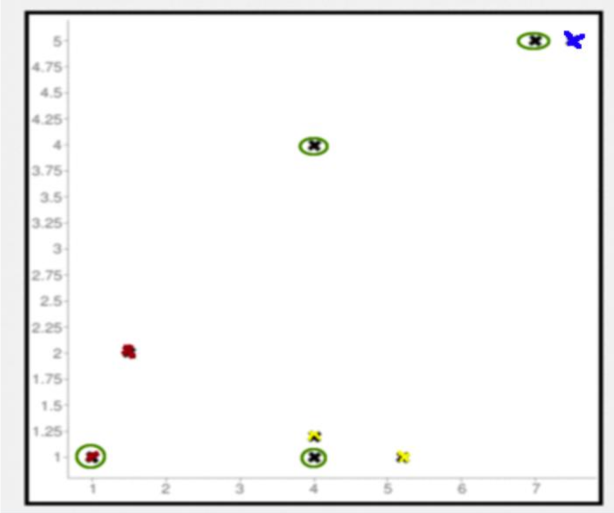

Fig 3: Final centroids of individual clusters after applying K-Means Algorithm

SET 2:

Number of Position Points (customer pickup coordinates in XY plane): 10

– Customer 1 to Customer 10

{1.0, 0.5}
{2.5, 2.0}
{4.0, 1.2}
{3.0, 4.0}
{7.5, 5.0}
{3.0, 1.0}
{2.5, 2.0}
{3.0, 1.2}
{5.2, 1.0}
{6.5, 5.0}

Number of Clusters to be formed: 4

Resultant Individual cluster centers:

(1.0, 0.0) (3.0, 1.0) (3.0, 4.0) (6.0, 5.0)

Cluster 0 includes: (1.0, 0.5)

Cluster 1 includes: (2.5, 2.0) (4.0, 1.2) (3.0, 1.0) (2.5, 2.0) (3.0, 1.2) (5.2, 1.0)

Cluster 2 includes: (3.0, 4.0)

Cluster 3 includes: (7.5, 5.0) (6.5, 5.0)

Centroids of clusters initialized at: (1.0, 1.0) (2.0, 3.0) (3.0, 4.0) (6.0,7.0)

Graphical Analysis:

The graph below in Fig 4 and Fig 5 shows the results before (Initial cluster centers initialized in k-means algorithm to begin with) and after applying k-means respectively (at the end of k-means Algorithm).

Initial analysis using centroids

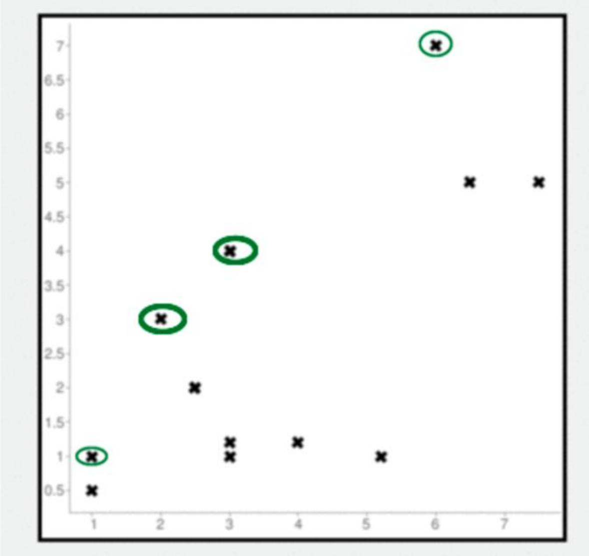

Fig 4: Initial Cluster Centers

Visually, you can see that the *K*-means algorithm splits the 4 groups based on the distance feature. Each cluster centroid is marked with a circle and each cluster is marked with different color.

Red points depict Cluster 0

Dark Blue points depict Cluster 1

Green points depict Cluster 3

Light Blue points depict Cluster 4

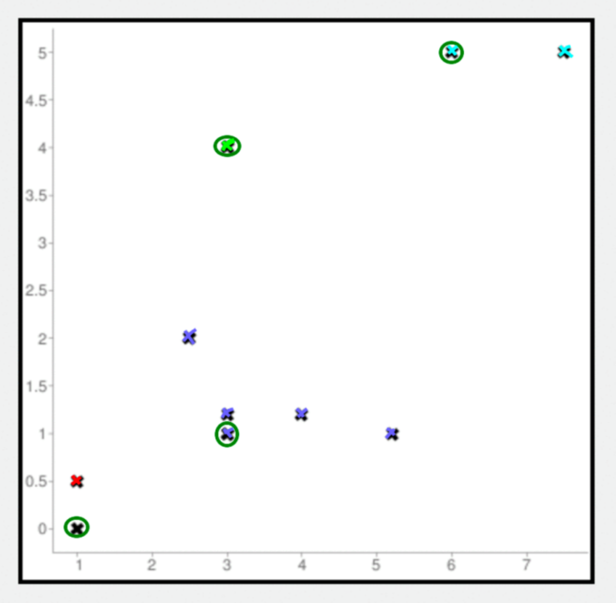

Fig 5: Final centroids of individual clusters after applying K-Means Algorithm

So, similarly, this approach can be used for a larger set of position data of customer's pick up point and also the number of cars available (equal to the final number of clusters to be formed).

6. FUTURE SCOPE AND PROPOSED SOLUTION

The proposed solution can be improved as currently, geographical locations is considered which is assumed to be evenly distributed in geographical space but if most of the location points are clustered around the same initially assumed centroids which was assigned initially to K-means Algorithm, most of the data will be clustered in the same cluster and other clusters may remain empty(as cluster-2 above in Dataset 1).So this can be tackled by recursive K-means algorithm, where empty clusters are removed and applying the K-means algorithm recursively in the clusters with data and assuming initial cluster points lying in the same clusters. This way the unused cars (cluster centers with no booking data) can be used recursively inside clusters which have no empty cab booking location set.

7. CONCLUSION

This paper showed the implementation of K-Means clustering algorithm of Data Mining [8] for finding the optimal pickup points for various customers with a limited number of cabs available to serve them at a point of time. For each subsequent addition of a new cab, the number of cluster increases by one and so the data density in a nearby cluster can decrease subsequently by addition of new cluster centroids. Also, as the position coordinate points increases, the data is clustered more in some clusters and some clusters may remain empty or less dense due to uneven distribution of position points in XY coordinate space as shown in this paper.

To sum it all, K-Means Algorithm can be employed effectively to solve problems which require near to optimal solution as shown in this paper. Though clustering may not give the optimal solution, it can give near to optimal solution and further optimization can be applied to obtain the optimal solution to a problem using clustering algorithms.

8. REFERENCES

[1] Han, J. &Kamber, M. (2012). Data Mining: Concepts and Techniques. 3rd.ed. Boston: Morgan Kaufmann Publishers.

[2] Sudhir Singh, Dr. Nasib Singh Gill, Comparative Study Of Different Data Mining Techniques: A Review, www. ijltemas.in, Volume II, Issue IV, APRIL 2013 IJLTEMAS ISSN 2278 – 2540.

[3] M. Ester, H. Kriegel, J. Sander, and X. Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. Proc. of the 2nd Int'l Conf. on Knowledge Discovery and Data Mining, August 1996.

[4] PERFORMANCE ANALYSIS OF PARTITIONAL AND INCREMENTAL CLUSTERING, Seminar National Aplikasi Teknologi Informasi 2005 (SNATI 2005) ISBN: 979-756-061-6 Yogyakarta, 18 June 2005.

[5] Data Mining and Statistics for Decision Making, Page no. 251, Stephane Tuffey, Wiley Publication.

[6] Performance Evaluation of Incremental K-means Clustering Algorithm, IFRSA International Journal of Data Warehousing & Mining |Vol1|issue 1|Aug 2011.

[7] R C Dubes, A K Jain, "Algorithms for Clustering Data," Prentice Hall, 1988.

[8] M H Dunham, "Data Mining: Introductory and Advanced Topics," Prentice Hall, 2002.