K-MEAN MACHINE LEARNING ALGORITHM

SIDDHARTH NANDAKUMAR CHIKALKAR

Bachelor of Computer Applications, Vivekananda collage, Kolhapur, Maharashtra, India.

---***---

ABSTRACT *-The k-mean algorithm which is partition-based cluster analysis method. In large data set k-mean algorithm has higher efficiency and scalability and fast coverage in mass of data set. This algorithm is easy to implementation. Clustering has been intensively studied in machine learning and data processing communities. Most of other clustering algorithm cannot handle efficiently clustering task but k-mean work efficiently.it is the one in every of the simplest and popular unsupervised machine learning algorithm this unsupervised algorithm make inference from data set using only input vector without regarding know, or labelled outcomes.*

Keywords: *machine learning, unsupervised clustering algorithm, data mining, cluster analysis.*

1. Introduction

The k-means clustering algorithm is the most popular clustering tool used in scientific and industrial applications the k-means algorithm is best suited for data mining because of its efficiency in processing large data sets. The large amount of data collected and stored in databases and it increases the need of effective analysis method. One of the primary data analysis tasks is clustering. The main goal of clustering algorithm is to group the object of database into sets of meaningful subclasses. Clustering help us to understand data in unique way.it is the faster way to analysis data and understand the huge amount of data. Clustering is the one of the common things in data mining and knowledge discovery. Clustering can help the business to manage their data better. for example, in retail store the data clustering help to analysis customer shopping behaviour sales campaign and customer retention .in case of insurance companies clustering is deployed in the field of ford detection, risk fact identification .and in case of banking sector data clustering help to customer segmentation credit scoring and analysing customer profitability. So, this is the data clustering use in various business. Even the amazon uses the recommended system which is using clustering it show you the recommended list of products according to your last purchase history. And on of other is Netflix it recommends you the movie based on your watch history whatever you are watch it show you some similar movie related to it. All the concept behind is the clustering. K-means is the clustering algorithm who's main is group similar element or data point into a cluster. 'k' means the number of the cluster.in this paper we see how k-means algorithm work on clustering. Application of clustering in real world scenarios is customer segmentation, document clustering, image segmentation, recommended engines these are some application were data clustering is used.in this paper we present how k-mean algorithm actually work and mathematics behind the k-mean. And their challenges and advantages.

2. WHAT IS THE CLUSTERING

The k-mean clustering algorithm is used for data clustering while data mining. This is the most common method .it is the unsupervised learning algorithm. clustering is one of the process which dividing the whole data set into groups also known as cluster based on same pattern in data set. Basically, it is the process to find a meaningful structure or patter in large data set and grouping of inherent in a set of examples. All data group which is divided by clustering method are similar to each other. there is different method we can clustering like distribution-based method, centroid based method, connectivity-based method etc. centroid based method it is the example of k-mean alogorithm.it is the one of irritative clustering algorithm .there are many application of clustering like customer segmentation and it isn't just limited for banking this strategy is used for every sector, e-commers, sport, advertising, sales etc. next is document clustering we need to cluster similar document together. Clustering help us to making group of similar documents are in the same cluster.one of other is image segmentation we can also use clustering to perform image segmentation. We can apply clustering to create clusters having similar pixels in the same group. And other is recommendation engines. Clustering is also be used in recommendation engines like music app. we want to have our favourite singer song in one playlist. There are many more application is used recommendation engines like Netflix etc.so the clustering is more important and we can make clustering using of k-mean algorithm. The priority of the actual fact is that the information is often complicated mismanaged and noisy. So, the k-mean algorithm not applicable for each situation.

2.1 HOW DOSE THE K-MEAN CLUSTERING WORK

The algorithm starts with initial estimates of k centroid which we can randomly selected from data set. The main goal of this algorithm is finding the group in the data. The algorithm works iteratively to assign each data point to one of *K* groups based on the features that are provided. Data points are clustered based on feature similarity. Let see how the k-mean algorithm is work. For example, this is the data we have to cluster the data set x = {10,20,30,35,65,45,56,78,89,95,80} now decide the how much cluster we want means the k value. Suppose we decide k value = 3.

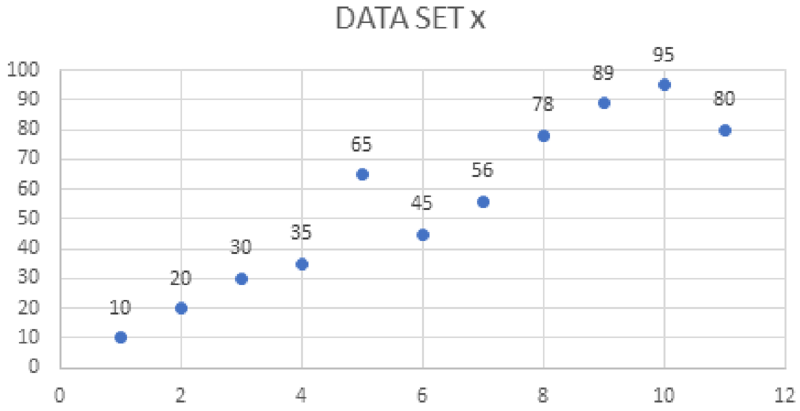

Figure 1. Data set x without clustering

This is the graphical presentation of data set x. the first step is choose any random number as centroid in data set x. we elect 30,65,78 this is centroid value. Now assign all remaining value which is just too near 30,65,78. Now k1=30, k2=65, k3=78. This is the randomly selected centroid. After assigning all value we got k1={10,20,30,35,45} and k2={65,56}and k3={78,89,95,80} this is our three cluster .now the second step is we have to calculate their mean of every clusters .after calculating their mean .k1=28 k2=60.5 and k3=85.5 .this is new centroid now assign all value again, to their nearest centroid. After assigning all value to new centroid we got k1= {10,20,30,35}, k2= {65,45,56}and k3= {78,89,95,80}. Now again calculate the mean of those newly formed data set. Now the new centroid after calculating their mean we got k1=23.75, k2=55.33 and k3=85.5. Again, assign all value to their closest centroid. After calculation the k1{10,20,30,35} k2={45,65,56} and k3= {78,89,95,80}.now we got the same pattern two time .all data point remain in same cluster now we will stop here ,we can say that algorithm isn't learning any new pattern and it's sign to stop

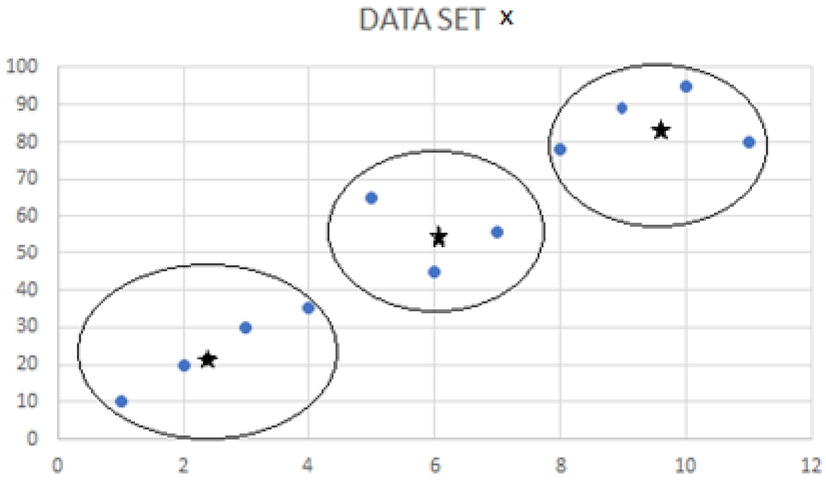

Figure 1.2 data set x after clustering

There are essentially three stopping criteria that can be adopted to stop the K-means algorithm:1) centroid of newly formed clusters do not change 2) points remain in the same clusters.3) maximum number of iterations are reached. We can stop the algorithm if the centroid of the newly formed clusters is not changing. Even after multiple of iteration if we are getting the same centroid for all clusters. Then we have to stop. This is the working of k-mean algorithm.

3. ADVANTAGES

Some of major advantages of k-mean clustering algorithm listed below.

- if the data set variable huge then the k-mean computation most times faster then hierarchical clustering , if we keep k small.
- Easy to implement and understandable we can easyliy implement k-mean for large data .
- We can generalizes of cluster of different shape and size such as elliptical cluster.
- K-mean algorithm is good in capturing structer of data if cluster have spherical like-shape.
- Use simple principles without the need for any complex statistical terms
- Once clusters and their associated centroids are identified, then it is easy to assign new objects to a cluster based on the object's distance from the closest centroid
- K-Means may produce tighter clusters than hierarchical clustering.

4. CHALLENGES

as the wide range of advantages k-mean algorithm come with some challenges.follwing are the some of major challenges for k-mean clustering algorithm .

  - difficult to predict k value .most of cases its hard to find appropiate k vaue we can use elbow method and the silhouette method for detemine k value
  - orderind the data strongrly affect on output
  - it dose not perform clustering well if the geomatric shape is complex.
- k-mean clustering dose not work well cluster (in the orignal data) of different size of different density.
- Clustering outliers. Centroids can be dragged by outliers, or outliers might get their own cluster instead of being ignored. Consider removing or clipping outliers before clustering.
- Initial seeds have a strong impact on the final result.
- Sensitive to scale: rescaling your datasets (normalizing or standardizing) will completely change results. While this itself is not bad, not realizing that you have to spend extra attention scaling your data might be bad.

5. APPLICATIONS

- K-Means clustering is used in a variety of examples or business cases in real life they are listed below.

- Academic performance. K-mean clustering is used in to creating cluster of students according their score and categorized in to grades A, B OR C.

- Diagnostic systems. K-mean used in medical profession to creating smarter medical decision support systems. Especially in the treatment of liver aliments.

- Search engine. Clustering forms the backbone of the search engine. Because when the search is the performed, the search result needs to be in grouped.

- Wireless sensor network. The clustering algorithm plays the role of finding the cluster heads, which collects all the data in its respective cluster

- Segmentation. Clustering used in segmentation like customer segmentation in banks, image segmentation etc.

- Recommendation engines: this type of clustering used in many apps and web. For e.g. in music songs app user want their favourite artiest song in one playlist.

6. CONCLUSION

In this paper we present the how k-mean clustering algorithm is work. K-mean is easy to use and understandable. We can easily use for large data set but we got difficulties to find k value. This paper also explains all challenges and advantages of k-mean algorithm, and k-mean application in real world. K-mean used in data mining and pattern recognition and give very good result so it is efficient algorithm in clustering of data. A good clustering method produce high quality cluster and k-mean is appropriate for good and high-quality cluster. The k-mean algorithm identifies k number of centroid and so assign every data point to nearest cluster while keeping the centroid as small as possible.

REFEERENCES

1. M.Jianliang, S.Haikun and B.Ling, "The Application on Intrusion Detection Based On K-Means Cluster Algorithm", IEEE International Conference on Information Technology and Applications, 2009.

2. Alan Jose, S. Ravi and M. Sambath, Brain Tumor Segmentation Using K-Means Clustering and Fuzzy CMeans Algorithm and Its Area Calculation. In International Journal Of Innovative Research In Computer And Communication Engineering, Vol. 2, Issue 2, March, 2014.

3. A. Mahendiran, N. Saravanan, N. Venkata Subramanian And N. Sairamm, Implementation Of K-Means Clustering In Cloud Computing Environment, Research Journal Of Applied Sciences, Engineering And Technology 4(10): 1391-1394, 2012.

4. Chinedu Pascal Ezenkwu, Simeon Ozuomba, Constance Kalu, Application Of K-Means Algorithm For Efficient Customer Segmentation: A Strategy For Targeted Customer Services, (Ijarai) International Journal Of Advanced Research In Artificial Intelligence, Vol. 4, No.10, 2015

5. M. Ester, H. Kriegel, J. Sander, and X. Xu. A DensityBased 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.

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

7. MashiatFatma, Jaya Sharma, Leukemia Image Segmentation Using K-Means Clustering And Hsi Color Image Segmentation,International Journal Of Computer Applications (0975 – 8887) Volume 94 – No 12, May 2014.

8. B.S.Vamsi Krishna, P.Satheesh, Suneel Kumar R., "Comparative Study Of K-Means And Bisecting KMeans Techniques In Wordnet Based Document Clustering", International Journal Of Engineering And Advanced Technology, Volume-1, Issue-6, August 2012

9. Fayyad, U.M., G. Piatetsky Shapiro, P. Smyth And R. Uthurusamy, Advances In Knowledge Discovery And Data Mining,Aaai Press/The Mit Press, Pp: 573-592, 1996