Cluster Estimation For Pattern Classification Problem Using K-Means Clustering

Ritu Vasishta
Department of Computer Science
Chandigarh University, Mohali , Punjab

Dr*. Jitender Kumar
Department of Computer Science
Chandigarh University, Mohali , Punjab

Abstract: **Clustering using K-means algorithm is very common way to understand and analyze the obtained output data. When a similar object is grouped, this is called basis of Clustering. There is K number of objects and C number of cluster in to single cluster in which k is always supposed to be less than C having each cluster to be own centroid but the major problem is how is identify the cluster is correct based on the data. Formulation of cluster is not a regular task for every tuple of row record or entity but it is done by a iterative process. Each and every record, tuple, entity is checked and examined and similarity dissimilarity is examined. So, this iterative process is seeming to be very lengthy and unable to give optimal output for cluster and time taken to find the cluster.**

To overcome the drawback challenge, we are proposing a formula to find the clusters at the run time so this approach can give us optimal results. The proposed approach use Euclidian distance formula as well melanosis to find the minimum distance between slots as technically we called as clusters and the same approach.

Keywords: **Data Clustering, Centroids, Data Mining, K-Means**

I. INTRODUCTION

Clusterise the data based on similarity measure is a very common practice. Finding the difference between object of one cluster to another cluster is another measure practice to find the distance. It denotes to extracting or mining knowledge from large quantities of data. Reminisce that excavating of gold from mountains having rocksand sand is called as gold mining. A. Khadem, , M. Sharifi[1] said mining of data should have been further suitably termed information gathering or mining from data, which is inappropriately slightly long. Knowledge mining, a shorter word, may not reject the emphasis on miningover data. However, it is word typifying procedure that discovers a lesser set of valuable nuggets from a prodigious pact of raw factual. Thus, such a misnomer which transfers both data and mining became a prominent choice. There are many other words carrying a similar or somewhat dissimilar sense to data mining, such as information mining from databases, knowledge mining, data arrangement analysis, data archaeology, and data scouring.

II. LITERATURE SURVEY

established a method in which the number of clusters, also known as related data objects groups, is estimated using a mathematical formula based on the cluster's data elements or objects. Initially, this method assumes a specific number of slots termed clusters, which are then compared to the resultant cluster slots formed by a specific number of clusters. The algorithm's disadvantage is that it is only useful when the number of datasets is small. The problem is that the algorithm is ineffective when dealing with enormous datasets.

Yanfeng Z hang and XiaofeiXu[3] suggested a fuzzy k-means classifier strategy for building decision cluster classifiers. The strategy is an agglomerative one. employs a mathematical model to create a cluster classifier, as well as certain logistic and fuzzy algorithms, and then uses them to find the number of clusters on real datasets before comparing the findings to other approaches. The algorithm's key benefit is that it allows you to manage the cluster's density level.

In this clustering technique, C. Blum, M. Samples [4][10][11] provide an approach for estimating the total number of clusters termed slot and initial centroid of the clusters. The strategy presented in this research was based on the watershed method, which has the ability to partition the density distribution of data into many regions. Every regional data centre detects the first K-means centre, and the cluster number is detected by the region number. We may conclude that the performance of the above-mentioned method is only acceptable and advantageous for initial parameter selection in the event of clustering of data elements and objects.

JianpengQi, Lihong Wang, and Jinglei Liu[5] present an approach known as innovative optimal hierarchical clustering method, which is based on three concepts. This method greatly boosted the likelihood of obtaining the greatest local optima as well as the top-n closest clusters.

III. PROPOSED WORK

Sections I and II explained the fundamentals of the k-means algorithm, as well as its numerous implementations and applications. Zhang, Lui, Rui Tang, Simon Fong [6][7][14] summarize the K-means approach sectionalizes data and the types of clusters, with the number of clusters varying depending on the user's needs and the length of the data. According to the study, the size of the c

should not exceed the size of the database. We propose two mathematical methods in this research study to detect the total number of similar slots for a large number of data items that do not fall into the category of local optima.

Applied K-Means

The clustering approach employed in this research is simple, and we begin by explaining the first version of the algorithm. The algorithm steps begin with selecting C initial centroids, where C is a user-defined limit, i.e. the number of clusters necessary. Each element is assigned to a local centroid, and a cluster is a collection of similar elements assigned to the same centroid. A. Dubey, A. Choubey, Sun, Liu [8][15] stated the cluster's center point is then modified depending on the elements assigned to the slots, which are referred to as cluster groups. This procedure is repeated until no elements or objects have been changed.

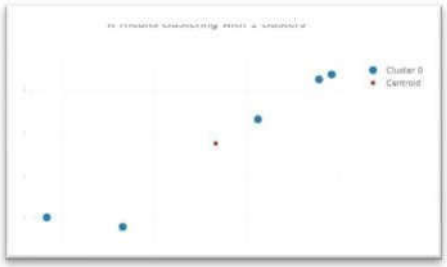
Process Flow:

1. Choose C points or elements or objects as primary centroids.
2. : recurrence
3. : Procedure C clusters by passing on each point or elements or object to its nearest centroid.
4. : Cluster or slot centroid is recalculated until Centroids are unchanged

Assigning Elements or object of data set to the Nearest neighbor called Centroid

We'd like proximity live, which quantifies the notion of "closest" for the particular knowledge under consideration, to assign some extent to the nearest centre of mass. As per M. Dorigo, M. Birattari[9] knowledge points in Euclidean space, Euclidean (L2) distance is commonly used, whereas trigonometric function similarity is much more acceptable for documents. The Manhattan (L1) distance, for example, is frequently used for geometer knowledge, whereas the Jacquard live is frequently used for documents.

C. Blum, M. Sampels [10][11] stated in there paper the algorithm assesses the similarity of every purpose to every center of mass repeatedly, the similarity measures employed for K-means are usually rather simple. However, in other



circumstances, such as when the data is in low-dimensional Euclidean space, it is possible to skip computing many of the similarities, allowing the K-means method to run much faster.

Handhayani, Wasito [12][13] K-means is another method for speeding up K-means computations by minimizing the number of similarities computed.

The k-means approach has been inverted, and new features such as cluster calculation and density fixing have been added. The key difference is that we changed the distance formula somewhat to compensate for the fact that the previously employed Euclidean function took longer.

So the quantity of samples does not fall into a specific cluster group range, the number of clusters formed is controlled based on the sample. This control is also necessary because no fixed data is accessible throughout time, and development time will increase, necessitating an increase in the number of slots for each iteration of the algorithm.

The two equations used in this work are as

Equation one : when samples is less than 10 lac and the other for more than 10lac.

$$= \sqrt{k}/3$$

Where k = number of sample points,c= cluster

f sample >10,000,00 then , number of slots called cluster

shall be = $k^{0.4}$

Centroid Formation : We have two separate formulas for different ranges of clusters because we have to build the number of ideal clusters of data components so that they do not lie in the local optima range. Here's an example of how the formula works for a wide range of data sets.

Example1. Let we have a set having data elements and objects not more than 10, 00,000.

Dataset n=12345, number. Of groups of similar data elements obtained by formula I = √3 is 37,

Now if we use formula II i.e. = $^{0.4}$ is 43, where 37 is minimum number of ideal data cluster where we can group elements.

Example 2: If set of data elements is greater than10, 00,000.So here,
Set of data say n=1111111, and number groups of similar data elements also say clusters generated by using mathematical formula II is as below :

= $\sqrt{}_3$ is 351, while by the next formula is 262,

So the minimum number of ideal groups or clusters is found by the equation as =$^{0.4}$.

IV. RESULTS

This section shows the possible results obtained from above experimental work simulation of the compiled code.

A. First Iteration

Iteration having 5 data points and the centroid result to 1 cluster, having 1 centroid set .

Fig 1.1 2D cluster graph on 5 data points

- Data Elements :5
- Dimension:2
- Cluster Formed:1
- Centroid:(132.29477558457307 , 55.198750630212444)

B. Second Iteration

Fig 1.2 shows the resultant graph obtained in third iteration having 50 data points and the centroid obtained to be to be 3 clusters , having 3 centroids sets.

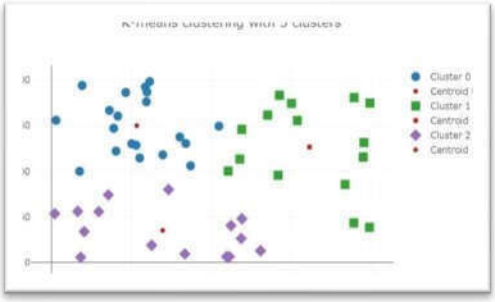


Fig 1.2 2D cluster graph representation over 50 datapoints using python

C. Third Iteration

In the third iteration with 500 data algorithmic output be 8 clusters having 8 centroids.

- Data Points:500
- Dimension:2
- Cluster Formed:8

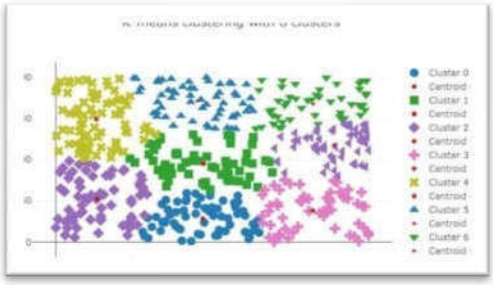


Fig. 1. 3 Cluster graph representation with large data using python

D. Forth Iteration

Resultant graph obtained for 3 dimensional data in iteration having 5 data. The centroid estimated by algorithm comes out to be 1 clustershaving 1 centroids.

- Data Elements :5
- Dimension:3
- Resultant Slots or Cluster:1

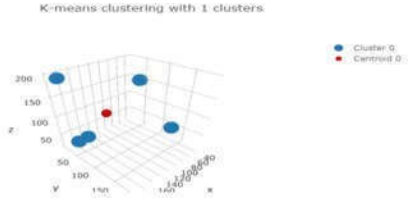


Fig. 1.4 3D pictorial view of Cluster graph

E. Fifth Iteration

obtained for 3 dimensional data in iteration having 10 data points. The centroid estimated by algorithm comes out to be 2 clusters having 2 centroids.

- Data Elements:10
- Dimension:3
- Resultant Slots or Cluster:8

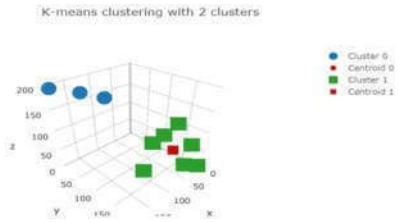


Fig1.5 3D cluster graph representation over 10 data points using python

F. Sixth Iteration

Resultant graph obtained for 3 dimensional data in iteration having 50 data points. The centroid estimated by algorithm comes out to be 3 clusters having 3 centroids.

- Data Elements:50
- Dimension:3
- Resultant Slots or Cluster:3

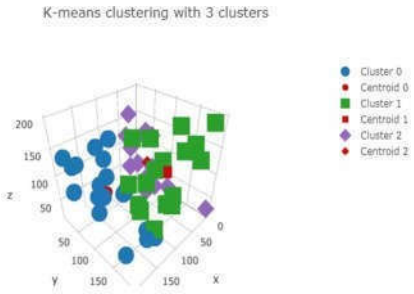


Fig1.6 3D cluster graph representation over 50 data points using python

VI. CONCLUSION

The proposed work of finding the optimal cluster and selecting the centroid, as we can see from the above experimental and compiled code work, is extremely important. The proposed application of a new formula, as discussed in section II, to a widely used algorithm known as k-means, in order to improve the algorithm's performance and make it ideal for clustering. The problem we ran into was that the algorithm became slow as the number of clusters increased above 20 lacs, and the system's requirements increased as well.

We may deduce from the experimental scenario that K-means When we need to fragment data on a wide scale and examine it, the clustering technique can be quite useful. Clustering techniques may be quite useful for evaluating large amounts of data, and the ability to convert data into relatable data is a huge benefit.

Many advantages were included in this research work,

as the auto cluster number definition and the generation of clusters of both two-dimensional and three-dimensional sample points. Also, the two-dimensional sample points might go up to twenty lakhs sample points without trouble, albeit anything more than that raises the system requirements, which are difficult to assemble and take time.

REFERENCES

[1] A. Khadem, E. F. Nezhad, M. Sharifi, "Data Mining: Methods & Utilities",Researcher2013; 5(12):47-59. (ISSN: 1553-9865).

[2] Abhijit Kane, Determining The Number of Clusters For a K-Means Clustering Algortihm, Indian Journal of Computer Science and Engineering(IJCSE).

[3] Xiaolong Wang, Yiping Jiao, ShuminFei, Estimation of Clusters Number and Initial Centers of K-means Algorithm Using Watershed Method , 2015 14th International Symposium on Distributed Computing and Applications for Business Engineering andScience.

[4] C. Blum, M. Sampels, Ant Colony Optimization forFOPshop scheduling: a case study ondifferent pheromone representations, Proc. 2002 Congr. on Evolutionary Computation (CEC'02),Vol. 2, IEEE Computer Society Press, Los Alamitos, CA, 2002, pp. 1558–1563

[5] Jianpeng Qi, Yanwei Yu, K-Means: An Effective and Efficient K-means Clustering Algorithm, 2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking,(SocialCom), Sustainable Computing and Communications(SustainCom)

[6] Yanfeng Zhang, XiaofeiXu, Yingqun Liu, Xutao Li, Yunming Ye, An Agglomerative Fuzzy K-means Approach to Building Decision Cluster Classifiers, 2011 Second International Conference on Innovations in Bio- inspired Computing andApplications.

[7] Rui Tang, Simon Fong, Xin-She Yang, Suash Deb, Integrating Nature-inspired Optimization Algorithmsto K-meansClustering,

[8] AnkitaDubey, Dr. AbhaChoubey, A Systematic Review on K-Means Clustering Techniques, International Journal of Scientific Research Engineering & Technology (IJSRET), ISSN 2278 – 0882 Volume 6, Issue 6, June 2017.

[9] M. Dorigo, M. Birattari, and T. Stitzle, "Ant Colony Optimization:Arificial Ants as a Computational Intelligence Technique, IEEEcomputational intelligence magazine, November, 2006.

[10] C. Blum, M. Sampels, An ant colony optimization algorithm for shop scheduling problems,J. Math. Model. Algorithms 3 (3) (2004) 285–308.

[11] C. Blum, M. Sampels, When model bias is stronger than selection pressure, in: J.J. MereloGuervos et al. (Eds.), Proc. PPSN-VII, Seventh Internat. Conf. on Parallel Problem Solving fromNature, Lecture Notes in Computer Science, Vol. 2439, Springer, Berlin, Germany, 2002, pp.893–902.

[12] TenyHandhayani, Ito Wasito, Fully Unsupervised Clustering in NonlinearlySeparable Data Using Intelligent Kernel K-Means, ICACSIS 2014,978-1-4799-8075- 8/14/$31.00 c_ 2014IEEE

[13] M. Li, M. Ng, Y. Cheung, and J. Huang. Agglomerative fuzzy k-means clustering algorithm with selection ofnumber of clusters. TKDE, 20(11):1519–1534, 2008.

[14] 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

[15] Yufen Sun, Gang Liu, Kun Xu, A k-Means-Based Projected Clustering Algorithm, 2010 Third International Joint Conference on Computational Science and Optimization