International Journal of Research in Engineering, Science and Management
Volume-1, Issue-12, December-2018
www.ijresm.com | ISSN (Online): 2581-5792

K- Means Clustering Using the Optimal Cluster Finding Technique

Anita Bishnoi[1], Vinod Todwal[2]
[1]*M. Tech. Student, Department of Information Technology, RCEW, Jaipur, India*
[2]*Associate Professor, Department of Information Technology, RCEW, Jaipur, India*

Abstract**: K-means is a simple technique to analyze the data by forming the clusters on the basis of similarity among the data. In K-means algorithm there are n no. of data and k no. clusters are formed to group the data where k<n. Where the no. of clusters are defined randomly and according to the no. of clusters the no. of centroids are formed, this approach does not produce the optimal results and fall into local optima.**

To get rid of this problem, we find a formula which determine the no. of clusters at run time and give the optimal results. In this paper we used mehlanobis and Euclidian distance formula to find the minimum distance between the data sets while forming the clusters and also applied the concept of Ant Colony Optimization. The results are produced in both 2-dimention and 3-dimension.

Keywords**: Clusters, Centroids, optimal no. of clusters.**

1. Introduction

K-means algorithm is most widely used to analyze the data by forming the clusters on the basis of most similar data in a single cluster and maximum difference in datasets among another clusters.

In this technique the k no. of clusters are previously defined which are to be formed by n no. of data by using appropriate distance formula, where k must be less than the no. of data sets i.e. n. The centroid is selected on the basis of no. of clusters to be formed. The centroids are selected to form the optimal no. of clusters where no. of iteration is performed. In this research paper we are having literature survey in next part and then proposed work is discussed that how the no. of clusters are selected and no. of iterations formed and then comes to the conclusion with the comparative results with the base paper. The results also produce the 2- D samples where X and Y axis are used. While in 3-D samples the result is shown in 3-dimension i.e. X, Y, Z.

2. Literature survey

Abhijit kane et.al. [1]: He determine the no. of clusters for k-means algorithm. In this paper he produced an algorithm which defines the no. of clusters by using a formula that depends upon the volume (no. of data elements) of the cluster. In this algorithm, the different no. of clusters are assumed and then compared the results produced by that particular no. of cluster with another no. of cluster. This algorithm is useful only when no. of datasets are less, it is not suitable for large datasets.

Yanfeng Z hang [2] . Xiaofei Xu used an agglomerative fuzzy k-means approach to build decision cluster classifiers, In this paper they used mathematical model for decision cluster classifier and tried to find the no. of clusters by using logistic & fuzzy methods, In this research some real datasets are used and compared those results with different approach. The algorithm can also control the density level of cluster.

Xiaolong Wang, Yiping Jiao, Shumin Fei [3]: This paper proposed an estimation of clusters number and initial centers of K-means clustering. The method is based on watershed method, which divides the data relative dimensional density distribution into multiple regions. Each regional center is selected as an initial K-means center, and the number of region is set as cluster number. Case study shows the performance of such method is beneficial to the selection of K-means clustering initial parameter. To operate K-means algorithm on large data set is less efficiency, but with initial parameter estimation, the algorithm enhanced both in veracity and running speed.

Jianpeng Qi, Yanwei Yu, Lihong Wang,and Jinglei Liu[4]: In this work we propose a novel optimized hierarchical clustering method incorporated with three optimization principles. K∗ initial centers effectively improves the probability of obtaining best local optima, and multi-round top-n nearest clusters merging approaches the optimal result gradually. The top-n and update principle optimizations update feature values of clusters by previous clusters or moved objects instead of re-computation from scratch. And the pruning strategy reduces significantly the adjusting searching space for each points in k-means iteration. An interesting direction for future work is to leverage modern distributed multi-core cluster of machines for further improving the scalability of our algorithm.

3. Proposed work

In Our work we proposed two formula for finding the no. of clusters for the large amount of data which does not fall into local optima.

Depending upon the sample of numbers the number of cluster formed is controlled. This was required because the number of sample might not fall into a particular range or there is no fixed data available with time and development it will increase and then the number of clusters will be required to get altered every iteration.

There have been two equations used in this as the first one is for the sample points less than ten lac and the other one is for more than 10 lac.

$$r = \sqrt{n}/3$$

Where n is the number of sample points and r is number of cluster

$$No.of\ Iteration = {}^{n}C_{r}$$

For more than ten lac sample points the number of clusters will be defined as:

$$r = n^{0.4}$$

Formation of centroid. As we have to form the no. of optimal clusters not to fall into local optima, so there are two formula for different range of clusters where we are trying to find the minimum optimal no. of clusters. Here, we are showing with an example that how these formulas work for different range of datasets.

Example: for less than 10 Lac datasets, let us have

Datasets n=12345, then no. of clusters formed by formula $r = \sqrt{n}/3$ is 37, while by the second formula i.e. $r = n^{0.4}$ is 43, where 37 is minimum no. of optimal clusters where our data can grouped. Similarly, for the datasets more than 10 lac, let n=1111111, and no. of clusters formed by formula $r = \frac{\sqrt{n}}{3}$ is 351, while by the next formula is 262, So, minimum no. of optimal clusters are found by $r = n^{0.4}$.

4. Results

This section deals with the possible results and data sets analysis of large data sets based on the simulation of the compiled code and hence the generated structured dataset of numerous values and indicators.

A. *First Iteration*

Result generated in first iteration consists of 5 data points in two dimensions. However the algorithm calculated the number of centroids that comes out to be 1 cluster, with a set of 1 centroid.

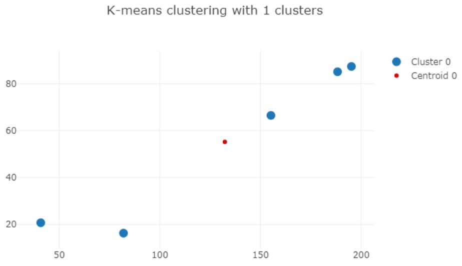

Fig. 1. 1 cluster graph

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

B. *Second Iteration*

Result generated in third iteration consists of 50 data points in two dimensions. However the algorithm calculated the number of centroids that comes out to be 3 clusters, with a set of 3 centroids.

- Data Points: 50
- Dimension: 2
- Cluster Formed: 3
- Centroid:(161.70062405552758, 150.1456957874784), (35.10179087960542,83.5588174561129) (126.72764640058038, 53.384859486789445)

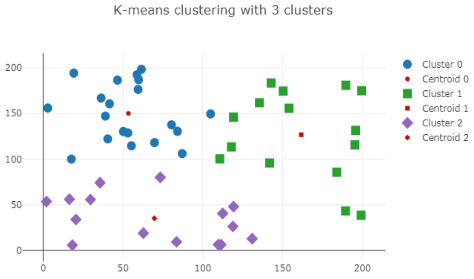

Fig. 2. 3 cluster graph

C. *Third Iteration*

Result generated in iteration 6 consists of 500 data points in two dimensions. However the algorithm calculated the number of centroids that comes out to be 8 clusters, with a set of 8 centroids.

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

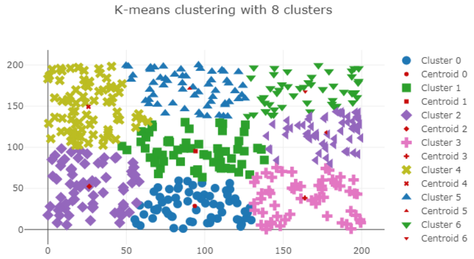

Fig. 3. 8 cluster graph

5. Conclusion

The proposed work on K-mean clustering algorithm is formation of clusters and selection of centroid is very crucial. Implementing new formula on k-means to make it more effective and optimized clustering algorithm. However the efficiency of algorithm slow down as the no. of clusters reaches above 20 lacs and the system requirement goes high.

References

[1] Xiaolong Wang, Yiping Jiao, Shumin Fei, 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 and Science.

[2] Yanfeng Zhang, Xiaofei Xu, 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 and Applications.

[3] Jianpeng Qi, Yanwei Yu, Lihong Wang,and Jinglei Liu, 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)

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

[5] Rui Tang, Simon Fong, Xin-She Yang, Suash Deb, Integrating Nature-inspired Optimization Algorithms to K-means Clustering,

[6] Ankita Dubey, Dr. Abha Choubey, 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.

[7] Teny Handhayani and Ito Wasito, Fully Unsupervised Clustering in Nonlinearly Separable Data Using Intelligent Kernel K-Means, ICACSIS 2014.

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