New K-Means Clustering Algorithm To Reduce Empty Clusters

M.Sivamani[1], S.M.Jagatheesan[2]
[2] Associate professor
[1, 2] Gobi Arts And Science College

Abstract- *Data mining is defined as a process used to extract usable data from a larger set of any raw data. It implies analysing data patterns in large batches of data using one or more software. Data mining has applications in multiple fields, like science and research. As an application of data mining, businesses can learn more about their customers and develop more effective strategies related to various business functions and in turn leverage resources in a more optimal and insightful manner. This helps businesses be closer to their objective and make better decisions. This process of retrieving useful data in the understandable form is data mining. Clustering is an important data analytic technique which has a significant role in data mining application. Clustering is the method of arranging a set of similar objects into a group. Partition based clustering is an important clustering technique. This technique is centroid based technique in which data points splits into k partition and each partition represents a cluster. A widely used partition based clustering algorithm is k- means clustering algorithm. But this algorithm also has some limitations. These limitations can be reduced by some improvements in existing algorithm. Empty clusters can happen when using K-means clustering algorithm, if the random initialization is bad, the number of K is inappropriate, the number of K is more than the number of data points in the data set. The original K-means algorithms is not designed to handle this situation. Empty clusters while running K-means, it will drop those clusters in the next iteration. So you may end up with fewer final clusters than initially gave to the algorithm. To avoid this situation, It has try different K or improve initialization of the initial cluster centers and reduce Empty clusters.*

Keywords- Data Mining, Clustering, New K-Means. Empty clusters.

I. INTRODUCTION

Data mining is an interdisciplinary subfield of computer science. It is the use of automatic data analysis techniques to uncover previously undetected relationship among data items [1]. Data mining finds valuable information hidden in large volumes of data. Data mining is the analysis of data and the use of software techniques for finding patterns and regularities in sets of data. It allows users to analyze data from various angles and dimensions, classified it and precise the relation recognized [2]. Data mining can be done by passing through different phases and using supervised and unsupervised learning. There are different types of techniques in data mining process. i.e. classification, clustering and regression. Data mining technologies have been applied successfully in many areas like marketing, telecommunication, fraud detection, financial data analysis, medical and many other scientific applications [5].

Clustering:

Clustering is the essential aspect of data mining. It is the technique of grouping of data in different groups by of their similarity measures. It means data items in the same group that are called cluster are more similar to each other than to those in other groups. Clustering is an unsupervised learning. A good clustering method will produce high superiority cluster with high intraclass similarity and low interclass similarity [4]. It is a common technique for statistical data analysis used in many fields of machine learning, image analysis, pattern recognition, bioinformatics and image retrieval.

II. LITERATURE REVIEW

Clustering techniques mainly used two algorithms: Hierarchical algorithm and Partition algorithm. In the hierarchical algorithm, the dataset is divided into smaller subset in a hierarchical manner whereas in partition algorithm dataset is partitioned into the desired number of sets in a single step. K-means clustering is most popular partition algorithm [2]. It uses in many application for producing the accurate result because of its simplicity of implementation. Rishikesh et al. [7] present a review on some variations of k-means algorithms. These algorithms remove some common problems of the basic k-means algorithm like a number of iteration in the algorithm, selection of initial cluster center, defining the number of cluster and clustering large data set. All these problems and algorithms to find the solution of these challenges discussed in this review paper. M. K. Pakhira et al. [8] present a modified k-means algorithm in this research paper. This m k-means algorithm provides a concept to modify the center vector updating procedure of the basic k-means. The only difference between basic k-means and m k-means is at the center computation step. The performance comparison of both algorithms by the rate of convergence and quality of the solution. K. A. Abdul Nazeer et al. [9] present an improved k-means algorithm which provides a systematic method for finding initial centroid and for assigning data points to the cluster. But this method also has a common problem of k-means that is the value of a number of clusters is still required to give as input. Y. S. Thakre et al. [10] discussed the evaluation of performance of the k-means algorithm with multiple databases and with various distance metrics. The k-means algorithm evaluated for a different number of clusters for recognition rate. This paper works help to choose suitable distance metric for a particular application. Nimrat Kaur et al. [11] proposed an algorithm using matrix method for assigning the data points to the cluster. This method requires less number of iterations to get good quality clusters

K-means Algorithm

K-means clustering is a well-known partitioning method. K-Means Clustering algorithm is an idea, within which there is a need to classify given dataset into K clusters; the value of K Number of clusters is defined by user which is fixed. In this first centroid of each cluster is selected for clustering and then according to chosen centroid, data points having a minimum distance from given cluster, is assigned to that particular cluster. Euclidean Distance is used for calculating the distance of data point from the particular centroid.

Old k-means algorithm consists of following steps:

1. Initialization: In this first step data set, the number of clusters and centroid that we defined for each cluster.
2. Classification: The distance is calculated for each data point from centroid and data point having the minimum distance from the centroid of a cluster is assigned to that particular cluster.
3. Centroid Recalculation: Clusters generated previously, centroid is again repeatedly calculated means recalculation of the centroid.
4. Convergence Condition: Some convergence conditions are given as below:
   4.1 Stopping when reaching a given or defined the number of iterations.
   4.2 Stopping when there is no exchange of data points between clusters.
   4.3 Stopping when a threshold value is achieved.
5. If all of above conditions are not satisfied, then go to step 2 and the whole process repeat again, until given conditions are not satisfied.

New k-means clustering Algorithm:

The K-means are partition based clustering algorithm. Here each clusters centroid values are represented by the mean value of the objects in the cluster .

Input:

K:The number of clusters
D:A data set containing n objects.
Output:A set of reduced Empty clusters
Method:
1.First choose k objects from D as the initial cluster centers;
2. Repeat
3.Assign each object to the cluster;
Repeat

Calculate the distance for each centroids from a data point and the data point having minimum distance from the centroid of a cluster is assign to that particular cluster centroid and calculate the mean value for that cluster centroid

Update the cluster mean to that centroid
Until end of data points
4.Untill no change in cluster values

1. Initialization:

   In this first step data set, the number of clusters and centroid defined for each cluster.

2. Classification: The distance is calculated for each data point from centroid and data point having minimum distance from centroid of a cluster is assigned to that particular cluster.

3. Centroid Recalculation: Clusters generated previously, centroid is again repeatedly calculated means recalculation of centroid.

4. Convergence Condition: Some convergence conditions given as below:

   4.1 Stopping when reaching a given or defined the number of iterations.
   4.2 Stopping when there is no exchange of data points between clusters.
   4.3 Stopping when a threshold value is achieved.

5. If all of above conditions are not satisfied, then go to step 2 and whole process repeat again, until given conditions are not satisfied.
6. Reduce of Empty Cluster: Clusters generated previously are rechecked. Clusters where no data points are allocated to a cluster under consideration during assignment phase are Reduced.

Proposed system Update the centroid of each cluster is an easy process.

Recomputed the mean value of the objects for each cluster.

Until centroids do not change and can get results for reduce empty cluster.

This method is relatively scalable and efficient in processing large data sets because the computational complexity of the algorithm is O(nki), where n is the total number of objects, k is the number of clusters, and i is the number of iterations, normally k < n and i < n.

Empty clusters:

Empty cluster problem is one of the drawbacks of the K-Means method. Proposed a new technique to reduce the computational cost and efficiency of the method.This New k-means algorithm will be able to reduce empty clusters and also have better performing results than compare to basic K-Means clustering algorithm.In situations, where the new k-means is used as an integral part of some higher level application, this empty cluster problem may produce anomalous behavior of the system and may lead to significant performance degradation. This modified version of the K-Means algorithm that efficiently reduce the empty cluster problem. K-Means clustering can give results rapidly and better performance than old K-Means or traditional K-Means clustering Algorithm.

One of the major problems of the k-means algorithm is that it may produce empty clusters depending on initial center vectors. The fundamental data clustering problem may be defined as discovering groups in data or grouping similar objects together and reduce empty cluster.

K-means focuses on the assignment of cluster centroid selection so as to improve the clustering performance by New K-Means clustering algorithm. Empty clusters are reduced by using this New K-Means algorithm. Empty clusters are reduced into two, while using this new k-means clustering algorithm

Result Analysis :

We showed that our proposed algorithm is capable to solve the empty clusters problem. The New k-means algorithm removes the existing limitations in efficient kmeans with better quality of data clustering and at the same time to reduce Empty cluster into two. Clustering is the process of grouping objects that belongs to the same class. Similar objects are grouped in one cluster and dissimilar objects are grouped in another cluster. One major problem with K-means clustering is that empty clusters are generated during execution, if within case no data points are allocated to a cluster under consideration during assignment phase. The databases are considered from real life projects. The algorithms are applied on databases. The number of clusters generated in case of old k-means is more because in this case empty clusters are generated. In the case of new k-means numbers of clusters are less than the previous case because empty clusters are reduced here and memory size also decreases in this case. The number of records increase, number of empty clusters also increases in old k-means algorithm. When new k-means algorithm is applied, number of clusters gets reduced because empty clusters have been reduced in this case.

III. CONCLUSION

K-Means clustering can give results rapidly and better performance than old K-Means or traditional K-Means clustering Algorithm.The new k-means is used as an integral part of some higher level application, this empty cluster problem may produce anomalous behavior of the system and may lead to significant performance degradation. One of the major problems of the k-means algorithm is that it may produce empty clusters depending on initial center vectors. The fundamental data clustering problem may be defined as discovering groups in data or grouping similar objects together and reduce empty cluster. K-means focuses on the assignment of cluster centroid selection so as to improve the clustering performance by New K-Means clustering algorithm. Empty clusters are reduced by using this New K-Means algorithm. Empty clusters are reduced into two, while using this New k-means clustering algorithm.The new K-Means algorithm that efficiently reduce the empty cluster problem.

REFERENCES

[1] Sharmila, R.C Mishra, “Performance Evaluation of Clustering Algorithms”, International Journal of Engineering Trends and Technology, 2013

[2] Pranjal Dubey, Anand Rajavat, “Implementation Aspect of K-Means Algorithm For Improving Performance”, Proceedings of 28th IRF International Conference, Pune, India, 7th June 2015

[3] Kehar Singh , Dimple Malik, Naveen Sharma, “Evolving Limitations in k-means Algorithm in Data Mining and their Removal”, International Journal of Computational Engineering & Management, April 2011

[4] Amandeep Kaur Mann, Navneet Kaur, “Review Paper on Clustering Techniques”, Global Journal Of Computer Science and Technology Software & Data Engineering, 2013

[5] S. D. Gheware , A. S. Kejkar, S. M. Tondare, “Data Mining: Task, Tools, Techniques and Applications”, International Journal of Advanced Research in Computer and communication Engineering, Oct 2014

[6] Anshul Yadav, Sakshi Dhingra, “A Review on K-Means Clustering Technique”, International Journal of Latest Research And Trends in Technology, July-Aug 2016

[7] Rishikesh Suryawanshi, Shubha Puthran, “Review of Various Enhancement for Clustering Algorithm in Big Data Mining”, International Journal of Advanced Research in Computer Science and Software Engineering, 2015

[8] Malay K. Pakhira, “A Modified k-means Algorithm to Avoid Empty Clusters”, International Journal of Recent Trends in Engineering, May 2009

[9] K. A. Abdul Nazeer, M. P. Sebastian, “Improving the Accuracy and Efficiency of K-means Clustering Algorithm”, Proceedings of the World Congress on Engineering, July 2009

[10] Y. S. Thakre, S. B. Bagal, “Performance Evaluation of K-means Clustering Algorithm with Various Distance Metrics”, International Journal of Computer Application, Jan 2015

[11] Nimrat Kaur Sidhu, Ranjeet Kaur, “Redefining and Enhancing K-means Algorithm”, International Journal of Innovative Research in Computer and Communication Engineering, May 2013

[12] Michael Hahsler, "Introduction to rules – A computational environment for mining association rules and frequent item sets" Journal of Statistical Software, 2005

[13] Michael Hahsler, “A Probabilistic Comparison of Commonly Used Interest Measures for Association Rules”, 2015

[14] Hipp, J., Güntzer, U., Nakhaeizadeh, G., "Algorithms for association rule mining --- a general survey& comparison". ACM SIGKDD Explorations Newsletter, 2000

[15] Pei, Jian, Han, Jiawei, Lakshmanan, Laks V. S., “Mining frequent item sets with convertible constraints”, in Proceedings of17th International Conference on Data Engineering, April 2–6, 2001

[16] Agrawal, Rakesh, Srikant, Ramakrishnan, “Fast algorithms for mining association rules in large databases”, in Bocca, Jorge B., Jarke, Matthias, Zaniolo, Carlo; editors, Proceedings of 20th International Conference on Very Large Data Bases (VLDB), Santiago, Chile, September 1994, pages 487-499

[17] Shafeeq Ahamed, K. S. Hareesha, “Dynamic Clustering of Data with Modified K-Means Algorithm”, International Conference on Information and Computer Networks, 2012

[18] Vrinda Khairnar, Sonal Patil, “Effcient clustering of data using improved k-means algorithm: A Review”, Imperial Journal of Interdisciplinary Research, 2016

[19] F. Bayat, *et al.*, “A non-parametric heuristic algorithm for convex and non-convex data clustering based on equipotential surfaces,” *Expert Systems with Applications*, vol. 37, pp. 3318-3325, 2010.

[20] E. Rasmussen, “Clustering algorithms,” in *Information retrieval*, B. F. William and B.-Y. Ricardo, Eds., ed: Prentice-Hall, Inc., 1992, pp. 419-442.

[21] E. W. Forgy, “Cluster analysis of multivariate data: efficiency versus interpretability of classifications,” *Biometrics*, vol. 21, pp. 768-769, 1965