IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-6, June 2019 E-ISSN: 2347-2693

Resolving Issues of Empty Cluster Formation in KMEAN Algorithm Using Advanced Approach

Saumya Kumar[1*], Neetu Verma[2]

[1,2]Deenbandhu Chhotu Ram University of Science and Technology

Corresponding Author: Kumarsaumya22@gmail.com

DOI: https://doi.org/10.26438/ijcse/v7i6.443448 | Available online at: www.ijcseonline.org

Accepted: 08/Jun/2019, Published: 30/Jun/2019

***Abstract*:** The k-means algorithm has been known as clustering techniques. It is used in various fields and domains such as medical imaging as well as biometrics fields etc. Although there are several optimums clustering mechanism in existence, the objective of paper is to discuss the clustering technique especially Kmean clustering. It has been observed that there are many researches already done in field of K-MEAN clustering. The issues related to Kmean clustering would be discussed in this research. Research has introduced the more effective and optimised cluster mechanism to classify the data set into various clusters.

Keyword: Clustering, Fuzzy, K-MEAN Clustering

I. INTRODUCTION

In the clustering, a set of objects are grouped in such a way that objects in the same group are more similar than other groups. In the context of databases, the clustering is known as the capability of several servers or instances to connect to a single database. Instance is known as group of processes and memory. It is communicating with a database. Such are the set of physical files which actually store to the data. The process of grouping a set certain objects into objects belonging to same classes are known as the clustering. In other words, it can be said that the aim of clustering is to separate groups with same traits and assign them into clusters. It is essential that the Documents and the files that locate within a cluster should be similar. The Documents that are located in separate clusters are dissimilar from each other.

II. APPLICATION OF CLUSTERING

Clustering is used in almost all the fields. of clustering are discussed:

Partitioning methods: Partition methods have the capability to improve the iterative relocation mechanism. It has been used in mining objects from one graph to another. The aim of partition clustering algorithm is to divide and breakdown the data points into K partitions

Hierarchical clustering: Such type of clustering is creates clusters that have a certain order from top to bottom in predetermined way. Files and folders on storage medium are arranged and stored in hierarchal manner

Fuzzy clustering: Fuzzy clustering is a form of clustering in which each data point belongs to more than one cluster. Set of objects are grouped in such a way that objects belonging to same group are categorised similar than other groups.

Density-based clustering: Density based clustering algorithm plays a vital role in searching non linear shapes and structures based on the density.

Model-based clustering: Model-based is used for designing an unknown distribution acting as a mixture of simpler distributions, which is sometimes called as basis distributions.

III. K MEAN CLUSTERING

K-means algorithm has been known as a data mining tool. It is also known as a machine learning tool. It has been applied to make a cluster observation into groups of related observations. There is not any prior knowledge required related to relationships. To take the sampling, the algorithm shows the category in which the data is located. Clustering has been applied on various kinds of dataset and various number of clusters are formed. This number is represented by the value k which represents clusters formed. The k-means clustering algorithm is applied in various fields such as medical imaging, biometrics, as well as related fields. This clustering has the capability to get knowledge about the data more willingly than to give instructions.

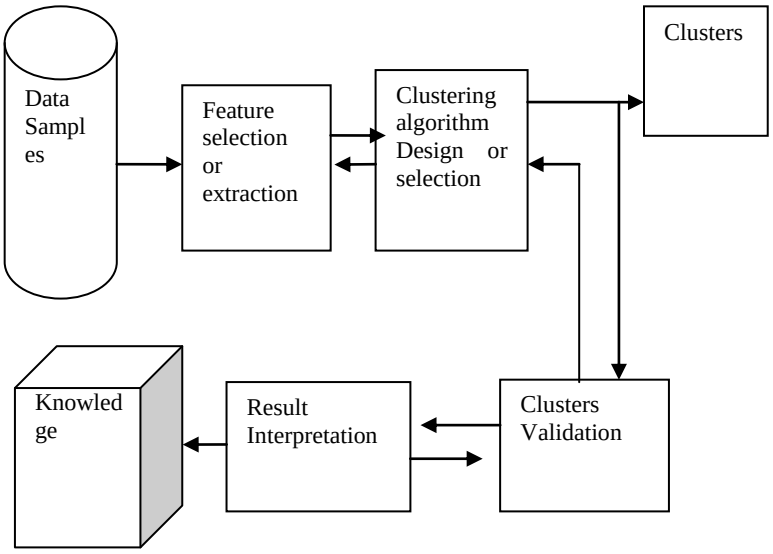

Fig 1. Steps in Clustering

IV. LITERATURE REVIEW

There are several researches done in the field of K-MEAN clustering mechanism.

In 2000, S. S. R. Abidi et al [1] discussed a data mining strategy. They have provided the relation between neural networks and K-means clustering. They have proposed solution with respect to data clustering. A technique has been proposed in the research work. They have introduced a technique to perform knowledge discovery on the data set through data clustering. The output result of the research is presented in the form of DCW. It is a data mining application.

In 2010, D. Napoleon et al [2] discussed an efficient and improved K-Means clustering algorithm. They have identified ways to minimise the time complexity. There were chances of wastage of time during uniform distribution data points. Accurateness of the algorithm has been identified during various execution of the program on the input data set. In the research work, the implementation work related to K-Means algorithm has been proposed. The motive of the research work is to convert the existing K-means clustering algorithm in more efficient and proficient algorithm. Thus the clustering without any complexity would be accessible. The algorithm‘s correctness is described at the time of several execution of the program. These programs perform with input data points. As the research result, the defined efficient K-Means happens to performs in much better way than the traditional K-Means algorithm.

In 2010, S. H. Ganesh et al[3]proposed on a novel priority mechanishm which is based data mining. They have discussed the advanced K-means clustering. It aids in detecting the protein sequence from dataset. To fulfill this objective, the research has made an efficient K-means clustering. These proposed algorithms have the capability to make the detection of the proteins sequence. The sequence is related to the dataset of frequent item set. The research work has implemented the sample of protein sequences taken from the PDB.

In 2010, I. S. Stinging, et al[4] explained visualization of K-means clustering on web-based OLAP operations. The clustering algorithm which has been applied was K-means. There was one objective of this implementation work considered by the researcher. It has been also highlighted by the researchers. The purpose of the research work was to formulate a visualization module. The proposed work is capable to assist the hotspot clusters resulted from OLAP operations.

In 2010, S. Na, et al [5] researched on k-means Clustering Algorithm.They have used modified k-means Clustering. The kmeans clustering is both ,a mining tool and also a machine learning tool. It has been applied to make a cluster observation into groups of related observations. The research work has discussed the standard k-means clustering algorithm. They also analyzed the loopholes in standard algorithm.

In 2011, K. A. A. Nazeer, et al[6] optimized the K-means Clustering. The k- algorithm is known as data mining tool and is also known as a machine learning tool. They made a research work with the use of O(n log n) Heuristic Method. It has been used to finding better Initial Centroids.

In 2011, R. V. Singh et al[7] proposed the data clustering with modified K-means algorithm. In the research work, the k-means clustering algorithm is described. It has been known as a data mining tool. It has been applied to make a cluster observation into groups of related observations. In the research work they have discussed a data clustering concept. The research has been made with the use of modified K-Means algorithm.

In 2011, D. V. S. Shalini, et al [8] discussed on frequent mining patterns on stock of data. The hybrid (mixed) clustering has been used in the research work. An algorithm has been proposed in the research work. It is an algorithm for mining and discovery of knowledge patterns from huge stock of data. The proposed system considered the factors that affect the products sale.

In 2012, N. Aini Abd Majid, et al[9] did research on K-means clustering pre-analysis. The research work has presented a new application. The research has highlighted the feature and advantages of the k-means clustering. K-mean clustering has been known as a machine learning

tool. It is capable to use in creation of the cluster observation into groups of related observations. K-means is a clustering tool. It has been used in order to determine precisely.

In 2012, T. Soni Madhulatha et. al. [10] discussed an overview on clustering methods. It is known as a data mining tool. It is also known as a machine learning tool. It has been applied to make a cluster observation into groups of related observations. This paper has covered the clustering algorithms, benefits and its applications. The research paper has concluded the limitations of clustering algorithm also.

In 2013, W.Sarada, Dr.P.V.Kumar [11] reviewed clustering techniques and performed comparative study of those techniques. There is confusion between the clustering and classification. They have explained that grouping certain objects into classes of similar objects is termed as clustering. In other words, they told that the aim of clustering is to separate groups with same traits and assign them into clusters. It is essential that the Documents and the files that locate within a cluster should be similar. The Documents that are located in separate clusters are dissimilar from each other

In 2013, hoj Raj Sharmaaet. al.[12] presented paper on clustering algorithms. Data mining has been known as the process. In this procedure, they analyzed the data from various different domains and summarized it into required and useful information.

Aim of the research is to calculate the efficiency of several data mining algorithms on diabetic dataset. Along with this they also determined the optimum algorithm. The efficiency and accurateness of the data mining is determined on the basis of many factors surrounding test mode, distance function and parameters.

In 2014, Shweta Srivastavaet. al. [13] presented the clustering techniques analysis for microarray data international.

In the research work, the embedded approach related to gene selection and clustering technique. Such techniques are efficient to perform the sample clustering/. In the research work, the clustering techniques are compared with each other. The comparison is been done on the basis of various parameters.

In 2015 Muhammad Husain Zafar et. al. [14] proposed clustering based study of classification algorithms. The research work has provided the study on the clustering techniques. They also provide the comparison of several clustering algorithms. Several clustering algorithms are available to analyze the data. The research paper has provided the required information and compares various clustering mechanisms. These algorithms are as follows K-Means, Farthest First, DBSCAN, CURE, Chameleon algorithm.

In 2015, N. Claypo et al [15] analyzed the opinion mining for restaurant reviews. For this purpose the K-Means is used in integration and association with MRF feature selection. In this Research, they proposed an opinion mining on reviews of Thai restaurant. MRF feature selection and K-Means clustering together are used in research work. Proposed method starts with preprocessing of text for the break down of reviews into various words and eliminating the stop words.

In 2016, S. Kapil, et al [16] wrote clustering algorithm on data using genetic algo. Here, K-means clustering is optimized. That's why the genetic algorithm has been used. In this research paper, k-means clustering has been optimized. For this purpose the genetic algorithm has been used. The reason to use this algorithm is the possibility of overridden of k-means. The result of the research work related to the k-means and genetic k-means has been analyzed. Output has indicated that the K-means combined with GA algorithm suggests new advancements and improvement in various research domain.

In 2016, S. Kapil et al[17] evaluated the performance of K-means clustering. The several distance factors and metrics also has been discussed. The results of the research work have shown the effect of such distance function applied on k-means clustering.
Objective of study is to monitor the k-means clustering and several distance functions that are applied in k-means. For example Manhattan distance function values and Euclidean value function. The experiment shows the effect of such distance function on k-means clustered data set . The comparisons of such distance functions are compared using various iteration, with sum squared errors and time elapsed in building the full model.

In 2016 , R. Ahlawat, et al[18]analyzed the factors affecting pattern of enrollment in Indian universities. They have utilized the k-means clustering for this purpose. The primary objective of research work is to observe the patterns of enrollment in universities of India..

In 2016, V. Baby et al[19] did research on distributed threshold k-means clustering. It has been used for privacy purpose in preserving the data mining. The research work has proposed an efficient and distributed threshold value indicating privacy-value in k-means algorithm.

In 2016, A. Saini, et al[20] gave a new model. The proposed approach was related to clustering in case of humongous amount of data which is called as big data. Amount of data gathered from various different sources is

increasing at exponential rate. So due to this, It has resulted in the need for more efficient algorithms to analyze large amount of datasets quickly . The research paper, presented an algorithm that is capable to overcome the disadvantages of previous algorithms used. Their work has provided a way to select a initial seeding in very less time, providing fast and accurate cluster analysis increasing the efficiency over large datasets.

In 2016, J. Qi, Y. Yu, et al [21] discussed an Effective K-Means Clustering Algorithm. The research work has proposed an advanced k-means clustering method. Research has proposed hierarchical advancement initialized by k* cluster centers. It has been done in order to reduce the risk of selecting seeds in random fashion .

In 2018, A. R. Condrobimo, B et al [22] proposed on data mining technique with cluster analysis. In their research work they have used the K-means algorithm for particular LQ45 index. The objective is to apply data mining techniques and along with it perform the analysis.

In 2018, M. Aryuni, et al [23] researched on Customer Segmentation in XYZ Bank. They have used the K-Means method of clustering

In the research work, Knowledge Discovery techniques has been applied. The performances of both methods were observed , analyzed and concluded. The result showed that based on intra distance, K-Means method outperformed and gave better results than K-Medoids method. K-Means with Davies-Bouldin index, performed better than K-Medoids.

V. LITERATURE REVIEW IN TABULAR FORM

SNO	YEAR	AUTHOR	TITLE	METHODOLOGY	Result
1	**2000**	S. S. R. Abidi and J. Ong	A data mining strategy for inductive data clustering[1]	inductive data clustering	Implemented mining on (exploratory) data
2	**2010**	D. Napoleon and P. G. Lakshmi	An efficient K-Means clustering algorithm for reducing time complexity using uniform distribution data points[2]	K-Means	The time taken by proposed methodology is less than the elapsed time taken in Traditional method of K-Means algorithm
3	2010	S. H. Ganesh and C. Chandrasekar	A novel priority based data mining algorithm using improved K-means clustering for detecting protein sequence from dataset[3]	Improved K-means clustering	Detect the pattern sequence of protein from given dataset
4	**2011**	K. A. A. Nazeer, S. D. M. Kumar and M. P. Sebastian	Enhancing the K-means Clustering Algorithm by Using a O(n logn) Heuristic Method for Finding Better Initial Centroids[6]	O(n logn) Heuristic Method	The proposed algorithm generates the better clusters formation in less time
5	2012	T. Soni Madhulatha	An Overview On Clustering Methods [10]	Clustering Methods	Overviewed the clustering methods
6	2013	W.Sarada, Dr.P.V.Kumar	A Review On Clustering Techniques And Their Comparison[11]	Clustering Techniques	Studied and provided the compare several clustering algorithms
7	2013	Bhoj Raj Sharmaa	Clustering Algorithms: Study And Performance Evaluation Using Weka [12]	Clustering Algorithms	Determined the optimum algorithm

8	2014	Shweta Srivastava	Clustering Techniques Analysis for Microarray Data [13]	Microarray	Refined the classification and compare the gene selection and clustering method
9	2015	Muhammad Husain Zafar.A	Clustering Based Study of Classification Algorithms [14]	Classification Algorithms	Researched and compared the K-Means, Farthest First, DBSCAN, CURE, Chameleon algorithm
10	2016	Qi, Y. Yu, L. Wang and J. Liu	*-Means: An Effective and Efficient K-Means Clustering Algorithm[21]	K-Means Clustering Algorithm	Optimized the k-means updation in the dataset of various iterations
11	2018	R. Condrobimo, B. S. Abbas	Data mining technique with cluster anaysis use K-means algorithm for LQ45 index on Indonesia stock exchange[22]	K-means algorithm	Cluster analysis for efficient and quick identifier for each and every member of LQ45 index cluster
12	2018	M. Aryuni, E. Didik Madyatmadja and E.Miranda	Customer Segmentation in XYZ Bank Using K-Means and K-Medoids Clustering[23]	K-Means and K-Medoids Clustering	K-Means outperformed than the K-Medoids algorithm.

VI. PROBLEM STATEMENT

During study of different clustering mechanism, K-mean Clustering is found to be significant mechanism to cluster the data. But it has certain limitations. There remains the issue of formation of empty clusters after applying clustering algorithm on the dataset. This leads to wastage of space due to empty cluster creation.

VII. CONCLUSION

The research work is related to the k-means algorithm. It is certainly considered as one of the simplest techniques and is widely used in various fields such as medical imaging , biometric field etc. The research has been made to calculate and compare the complexity of space consumption between K-Means and Advanced and Optimized K-Means algorithm. The K-Means algorithm is widely as well as frequently used proposed algorithm. Traditional Algorithm is capable to give results for smaller (in size) datasets whereas the optimized method gives the results for larger (in size) data sets and also takes less time during process of execution .This research may be Performed when elapsed time of the K-Means algorithm is greater and larger than Optimized K-Means algo.

REFERENCE

[1]. S. S. R. Abidi, "A data mining strategy for inductive data clustering: a synergy between self-organising neural networks and K-means clustering techniques," 2000 TENCON Proceedings. Intelligent Systems and Technologies for the New Millennium (Cat. No.00CH37119), Kuala Lumpur, Malaysia,

[2]. D. Napoleon, "An efficient K-Means clustering algorithm for reducing time complexity using uniform distribution data points," Trendz in Information Sciences & Computing(TISC2010), Chennai, 2010, pp. 42-45.

[3]. S. H. Ganesh, "A novel priority based data mining algorithm using improved K-means clustering for detecting protein sequence from dataset," 2010 IEEE International Conference on Computational Intelligence and Computing Research, Coimbatore, 2010, pp. 1-4.

[4]. S. Sitanggang, "K-means clustering visualization of web-based OLAP operations for hotspot data," 2010 International Symposium on Information Technology, Kuala Lumpur, 2010, pp. 1-4.

[5]. S. Na, "Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm," 2010 Third International Symposium on Intelligent Information Technology and Security Informatics, Jinggangshan, 2010, pp. 63-67.

[6]. K. A. A. Nazeer, "Enhancing the K-means Clustering Algorithm by Using a O(n log n) Heuristic Method for Finding Better Initial Centroids," 2011 Second International Conference on Emerging Applications of Information Technology, Kolkata, 2011, pp. 261-264.

[7]. R. V. Singh, "Data clustering with modified K-means algorithm," 2011 International Conference on Recent Trends in Information Technology (ICRTIT), Chennai, Tamil Nadu, 2011, pp. 717-721.

[8]. D. V. S. Shalini, "Mining frequent patterns of stock data using hybrid clustering," 2011 Annual IEEE India Conference, Hyderabad, 2011, pp. 1-4.
[9]. N. Aini Abd Majid, "K-means clustering pre-analysis for fault diagnosis in an aluminium smelting process," 2012 4th Conference on Data Mining and Optimization (DMO), Langkawi, 2012, pp. 43-46.
[10]. T. Soni Madhulatha. AN OVERVIEW ON CLUSTERING METHODS. IOSR Journal of Engineering Apr. 2012, Vol. 2(4) pp: 719-725.
[11]. W.Sarada, A REVIEW ON CLUSTERING TECHNIQUES AND THEIR COMPARISON , International Journal of Advanced Research in Computer Engineering &Technology (IJARCET) Volume 2 Issue 11, November 2013
[12]. Bhoj Raj Sharmaa Clustering Algorithms: Study and Performance Evaluation Using Weka Tool.International Journal of Current Engineering and Technology ISSN 2277 - 4106 © 2013.
[13]. Shweta Srivastava. "Clustering Techniques Analysis for Microarray Data." International Journal of Computer Science and Mobile Computing A Monthly Journal of Computer Science and Information Technology IJCSMC, Vol. 3, Issue. 5, May 2014
[14]. Muhammad Husain Zafar. A Clustering Based Study of Classification Algorithms. International Journal of Database Theory and Application Vol.8, No.1 (2015), pp.11-22.
[15]. N. Claypo "Opinion mining for thai restaurant reviews using K-Means clustering and MRF feature selection," 2015 7th International Conference on Knowledge and Smart Technology (KST), Chonburi, 2015, pp. 105-108.
[16]. S. Kapil, "On K-means data clustering algorithm with genetic algorithm," 2016 Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC), Waknaghat, 2016, pp. 202-206.
[17]. S. Kapil and M. Chawla, "Performance evaluation of K-means clustering algorithm with various distance metrics," 2016 IEEE 1st International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES), Delhi, 2016, pp. 1-4.
[18]. R. Ahlawat, S "Analysis of factors affecting enrollment pattern in Indian universities using k-means clustering," 2016 International Conference on Information Technology (InCITe) - The Next Generation IT Summit on the Theme - Internet of Things: Connect your Worlds,
[19]. V. Baby, "Distributed threshold k-means clustering for privacy preserving data mining," 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Jaipur, 2016, pp. 2286-2289.
[20]. Saini, J. Minocha, "New approach for clustering of big data: DisK-means," 2016 International Conference on Computing, Communication and Automation (ICCCA), Noida, 2016, pp. 122-126.
[21]. Qi, Y. 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) (BDCloud-SocialCom-SustainCom), Atlanta, GA, 2016, pp. 242-249.
[22]. R. Condrobimo, "Data mining technique with cluster anaysis use K-means algorithm for LQ45 index on Indonesia stock exchange," 2018 International Conference on Information and Communications Technology (ICOIACT), Yogyakarta, 2018, pp.
[23]. M. Aryuni, "Customer Segmentation in XYZ Bank Using K-Means and K-Medoids Clustering," 2018 International Conference on Information Management and Technology (ICIMTech), Jakarta, 2018, pp. 412-416.

