IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-5, May 2019 E-ISSN: 2347-2693

Rapid Clustering Algorithm for Optimizing Cognate Data of Online Database

B.S. Rawat [1*], K. Kumar [2], R.K. Mishra[3], S.S Bedi[4]

[1, 2, 3] Deptt. Of Computer Applications, IFTM University, Moradabad Uttar Pradesh, INDIA
[4]Deptt. of CSIT, IET, MJP Rohilkhand University, Bareilly Uttar Pradesh, INDIA

DOI: https://doi.org/10.26438/ijcse/v7i5.10761082 | Available online at: www.ijcseonline.org

Accepted: 19/May/2019, Published: 31/May/2019

***Abstract*—** Clustering is one of the main diagnostic method in data mining, widely used in cluster analysis having higher efficiency and scalability when dealing with large data sets. So far, numerous useful clustering algorithms have been developed for large databases, such as Connectivity based clustering [1], Centroid based clustering [2], Distribution based clustering[3] and Density based clustering[4]. K-means clustering algorithm was proposed by MacQueen [5] which is a Centroid based cluster analysis method. However there are some limitations of standard K-means algorithm: initialization of cluster centers, how K-means clustering algorithm calculates the distance between each data objects and cluster centers in each iteration. This paper proposes an improved K-means algorithm which first preprocesses the data and then arranges the dataset in a sequential order thus reducing the number of iterations and complexity. In preprocessing, the noisy data is removed and the resultant data undergoes the improved process of sorting and clustering which controls the computing of distance with each data object to the cluster centers iteratively, saving the execution time. Experimental results show that the improved method can effectively advance the speed of clustering and accuracy, reducing the computational complexity of the K-means.

Keywords— Data mining, Clustering, K-means, improved K-means.

I.INTRODUCTION

Clustering is considered as one of the important techniques in data mining and is an active research topic for the researchers in the field of online shopping, medical systems, hotels and many more. Clustering partitions a set of objects into clusters such that pattern of objects within a group are more similar to one another, than patterns of objects in different clusters. So far, numerous useful clustering algorithms have been developed for large databases, such as K-MEANS [6], CLARANS [7], BIRCH [8], CURE [9], DBSCAN [4], OPTICS [10]. These algorithms can be divided into several categories, out of which prominent categories are partitioning, hierarchical and density-based. All these algorithms try to challenge the clustering problems treating huge amount of data in large databases. However, none of them are the most effective. In density-based clustering algorithms, which are designed to discover clusters of arbitrary shape in databases with noise, a cluster is defined as a high-density region partitioned by low-density regions in data space. DBSCAN (Density Based Spatial Clustering of Applications with Noise) [4] is a typical density-based clustering algorithm. In this paper, we present a new algorithm which overcomes the drawbacks of K-means clustering algorithms used in online databases.

Online database is a record of logically related information, recorded in computer files in a uniform form to facilitate easy and efficient retrieving of data by means of internet or other communication networks. Records in online databases are further divided into various fields (product name, brand name, price range etc.) for categorizing, searching and

Retrieving information. Hence online database provide information on different fields of study with great ease, accuracy and speed.

Some of well known online databases of different fields are :

A) **Marketing**: Snapdeal, Flipkart, Amazon etc.
B) **Educational**: Science Direct, ProQuest, Cambridge University Press etc.
C) **Medical**: AIIMS, APPOLLO, MAX etc
D) **Hotels:** Hyatt, Taj, Oberoi etc

Online database deal with a variety of information in various forms and formats, comprising textual information, bibliographic information, numeric and multimedia information including text, audio, images and video. The online database acts as a association between the generators of information and the seekers of that information. This paper includes four parts:1) Introduction, 2) K-means Algorithm, 3) Literature Review, 4) Proposed Methodology

and Algorithm of Improved K-means Algorithm 4) Experimental Results and 5) Conclusion through experimenting data sets.

II. K-MEANS ALGORITHM

K-means is a well known Centroid based, simple unsupervised clustering algorithm in data mining, widely used for clustering large sets of data. It follows a simple procedure of classifying a given data set, which is discussed below.

2.1 Modus Operandi of K-means Algorithm

K-means algorithm was applied to solve the problem of large cluster [11]. It is a partitioning clustering algorithm, which classify the given data objects into k different clusters that are compact and independent. The algorithm consists of two phases. First phase selects k centers randomly, where the value of k is fixed. The next phase is to take each data object to the nearest center [11] by Euclidean distance method. When all the data objects are included in k clusters, then calculate the average of each cluster. This iterative method will repeat, until a constant mean is achieved.

Supposing that the target object is x and x_i indicates the average of cluster C_i then criterion function is defined as:

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} |x - x_i|^2$$

E is the sum of the squared error of all objects in database. The distance of criterion function is Euclidean distance, which is used for determining the nearest distance between each data object and cluster center. The Euclidean distance between one vector $x=(x_1, x_2 \ldots x_n)$ and another vector $y=(y_1, y_2 \ldots y_n)$, The Euclidean distance $d(x_i, y_i)$ can be obtained as follow:

$$d(x_i, y_i) = \left[\sum_{i=1}^{n} (x_i - y_i)^2 \right]^{1/2}$$

The process of K-means algorithm as follow:

Input: Number of desired clusters k, and a database D= {d_1, $d_2 \ldots d_n$} containing n data objects.

Output: A set of k clusters

Steps:

1) Randomly select k data objects from dataset D as initial cluster centers.

2) Repeat;

3) Calculate the distance between each data object d_i(1<=i<=n) and all k cluster centers c_j(1<=j<=k) and assign data object d_i to the nearest cluster.

4) For each cluster j(1<=j<=k), recalculate the cluster center.

5) Until no changing in the center of clusters.

The K-means clustering algorithm always converges to local minimum. Before the K-means algorithm converges, calculations of distance and cluster centers are done while loops are executed a number of times, where the positive integer t is known as the number of K-means iterations. The precise value of t varies depending on the initial starting cluster centers [12]. The distribution of data points has a relationship with the new clustering center, so the computational time complexity of the K-means algorithm is O(nkt), where n is the number of all data objects, k is the number of clusters and t is the number of iterations.

III. LITERATURE REVIEW

1) Zhe Zhang et.al [13] discussed and improved K-means at two points: optimization of initialization and improvement on global searching capability. The improved clustering algorithm increased the searching probability around the best centroid and enhanced the stability of the algorithm. The experiment on two groups of representative dataset proved that the improved K-means algorithm performs better in global searching and is less sensitive to the initial centroid.

2) Malay K. Pakhira [14] worked on one of the major problems of the K-means algorithm that is, it produces empty clusters depending on initial center vectors. For static execution of the K-means, this problem is considered insignificant and can be solved by executing the algorithm for a number of times. In situations, where the 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. Pakhira presented a modified version of the K-means algorithm that efficiently eliminates this empty cluster problem. Here, a new center vector computation strategy enables us to redefine the clustering

process and is found to work very satisfactorily, with some conditional exceptions which are very rare in practice.

3) Shi Na et.al [15] analyzed that the K-means clustering algorithm has to calculate the distance between each data object and all cluster centers in each iteration, which decreases the efficiency of clustering. Hence, Shi Na proposed an improved model of K-means algorithm which requires a simple data structure to store some information in every iteration. This information would be used in the next iteration. The improved method avoids computing the distance of each data object to the cluster centers repeatedly, saving the running time. Experimental results show that the improved method can effectively improve the speed of clustering and accuracy, reducing the computational complexity of the K-means.

4) K. Mumtaz and Dr. K. Duraiswamy[16] worked on spatial data that has been collected from various applications ranging from geo-spatial data to bio-medical knowledge. The spatial data collected is increasing exponentially and exceeds human's ability to analyze it. Recently, clustering has been recognized as a primary data mining method for knowledge discovery in spatial database. The database can be clustered in many ways depending on clustering algorithm employed, parameter settings used, and other factors. Multiple clustering can be combined so that the final partitioning of data provides better clustering. In this paper, a novel density based K-means clustering algorithm has been proposed to overcome the drawbacks of DBSCAN and K-means clustering algorithms. The result obtained is an improved version of K-means clustering algorithm. This algorithm performs better than DBSCAN while handling clusters of circularly distributed data points and slightly overlapped clusters.

5) Juntao Wang and Xiaolong Su[17] proposed and worked on some of the major deficiencies of K-means: initialization of the number of clusters K , arbitrarily selection of the initial cluster centers and the algorithm is influenced by the noise points. Juntao presented an improved K-means algorithm using noise data filter which developed density-based detection methods based on characteristics of noise data where the discovery and processing steps of the noise data are added to the original algorithm. By preprocessing the data, Juntao excluded the noisy data before clustering, reducing the impact of noise data on K-means algorithm. The results obtained are more accurate as compared to original K-means.

6) Navjot Kaur et.al[18] discussed about the experimental results of K-means clustering and its performance in case of execution time . Navjot observed that K-means clustering algorithm takes more time for execution which highly effected its performance. So in order to reduce the execution, time they used the Ranking Method and shown that how clustering is performed in less execution time as compared to the traditional method. This work makes an attempt at studying the feasibility of K-means clustering algorithm in data mining using the Ranking Method.

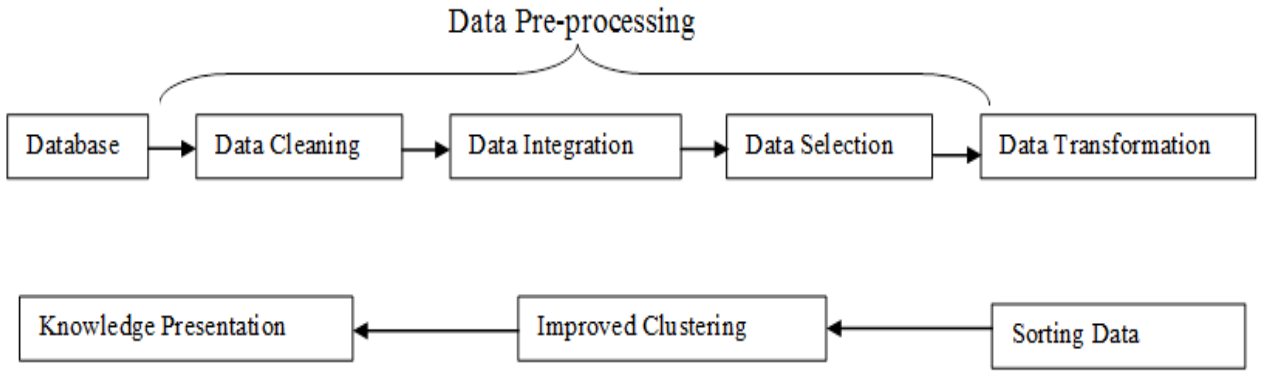

Fig. (1): Data Mining Model

7) Shyr-Shen Yu [19] proposed tri-level K-means algorithm and bi-layer K-means algorithm that removed the vulnerability of K-means to noisy data and susceptibility to initial cluster centers. While the data in a dataset S are often changed, after a period of time the trained cluster centers cannot precisely describe the data in each cluster. In this paper, Shyr-Shen Yu proposed an online machine learning based tri-level K-means algorithm to solve this problem.

Noisy data, outliers, and the data with quite different values in a same cluster may decrease the performance of a pattern matching system. Bi-layer K-means algorithm can deal with above problems. Experimental results demonstrate that both algorithms can provide much better accuracy of classification than tradition

3.1 Drawback of K-means Algorithm

We can see from the above analysis, that K-means algorithm calculates the distance from each data object to every centroid (cluster center) in each iteration and then compare them for the smallest distance. Assuming that cluster C formed after the first j iterations, the data object x is assigned to cluster C, but in a few iterations, the data object x is still assigned to the cluster C. In this process, after several iterations, we calculate the distance from data object x to each cluster center and find that the distance to the cluster C is the smallest. So in the course of several iterations, K-means algorithm calculate the distance between data object x to the other cluster center, which takes up a long execution time thus affecting the efficiency of clustering.

IV. PROPOSED METHODOLOGY AND ALGORITHM OF IMPROVED K-MEANS ALGORITHM

In this section we proposed an improved K-means method and experimented on very large datasets like SD, FK and VKM which were available online. Subsequently we applied K-means algorithm to our datasets and compared the result of both algorithms. This section unfolds the architecture, algorithm and algorithm flow of improved K-means algorithm.

4.1 Proposed Methodology

The proposed architecture is divided into three phase model. In first phase, the data is collected from our retail smart store. Then, the second phase is data preprocessing that starts with the data cleansing which involves removing the noisy data first, so the incomplete, missing and irrelevant data are removed and formatted according to the required format. In third phase, the data is sorted using quicksort algorithm followed by the last and fourth phase which generate the clusters and shape them as desired. Fig.1 illustrates the whole process.

4.2 Proposed Algorithm

K-means algorithm calculates the distance from each data object to all the centers of k clusters, which takes up a lot of execution time especially for large-capacity databases. For the shortcomings of the above K-means algorithm, we propose an improved K-means method. The main motive of the new algorithm is sorting the given dataset in ascending order which decreases the number of iterations that were needed earlier to calculate the distance from the data object to the other k clustering centers. This results in decreased complexity of earlier algorithm, thus increasing the efficiency and decreasing the time of the clustered dataset.
The process of new improved K-means algorithm as follow:

Input: Number of desired clusters, k, and a database D={d_1, d_2…d_n} containing n data objects.

Output: A set of k clusters

Steps:

1) *Sort the input data using Quicksort algorithm*

2) *Divide the total number of elements by the number of clusters to be formed, to get the number of elements in a single cluster i.e Number of elements in a cluster = n / k. If , remainder is 0, divide the elements equally(i.e equal to quotient), else If we have a remainder, then divide the elements equally except for the last cluster, since the remaining elements (which is equal to remainder) will be added to the last cluster.*

3) *Calculate the mean of objects in each cluster as the new cluster centers*,

$$m_i = \frac{1}{N}\sum_{j=1}^{N_i} x_{ij} ,$$

i= 1, 2, 3…….k; N is the number of elements of elements of current cluster i.

4) *If new mean of the cluster are obtained , then calculate the distance between each object x_i of two consecutive clusters and new cluster centre (mean) of these two cluster, then assign each object to the nearest cluster.*

5) *For each cluster j(1<=j<=k), recalculate the cluster center using step 3.*

6) *Until no changing in the center of clusters.*

The complexity of the above algorithm is O(nk(t/2)) where t<<<n, which is better than the earlier computational complexity of K-means algorithm i.e. O(nkt).
Algorithm flow of improved K-means method is given below in Fig. 2

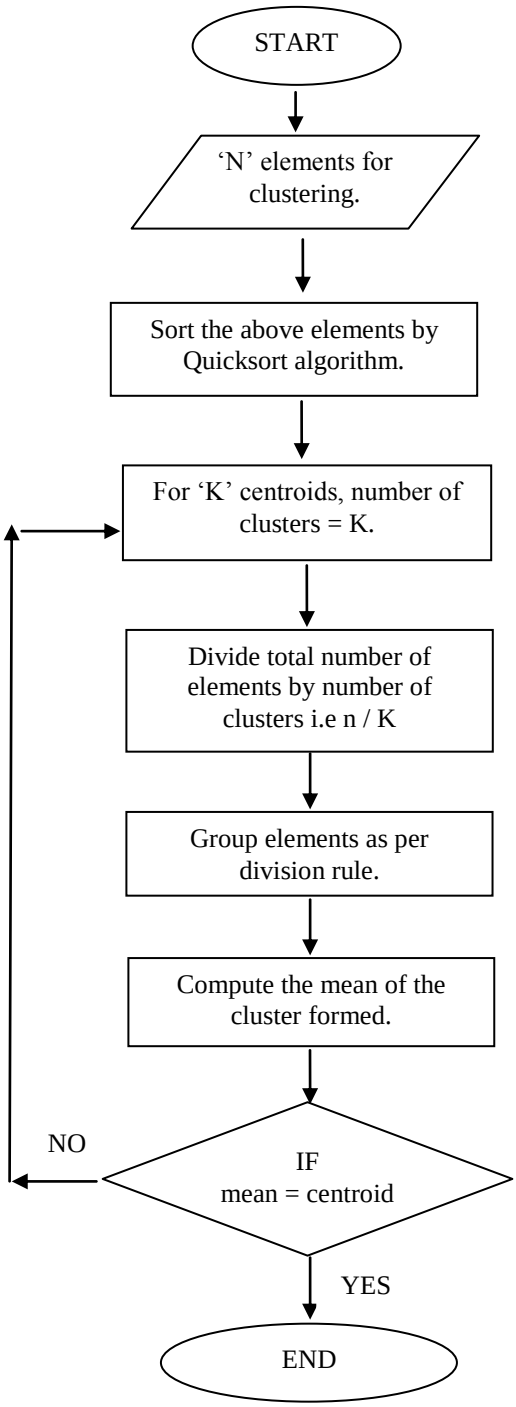

Fig. 2: Proposed algorithm flow.

V. EXPERIMENTAL RESULTS

In this proposed model of improved K-means, we worked on three large datasets SD, FK and VM from the online available repositories to test the efficiency of improved K-means algorithm and the K-means. Before starting the experiment we preprocess our datasets whose result is shown in Fig. 3. Preprocessing removes the noisy, dirty and unwanted data from our datasets, thus, minimizing the complexity of further process. Three simulated experiments have been carried out to demonstrate the performance of the improved K-means algorithm. In three experiments, we compute: number of iterations, accuracy and execution time for both the algorithms. Experimental comparison of improved K-means algorithm with the K-means algorithm in terms of the number of iteration, accuracy and execution time is shown in Table 1, Table 2, Table 3 and Fig 4, Fig 5, Fig 6 respectively. Experimental operating system is Window 10 and tool used for experimenting the datasets is WEKA.

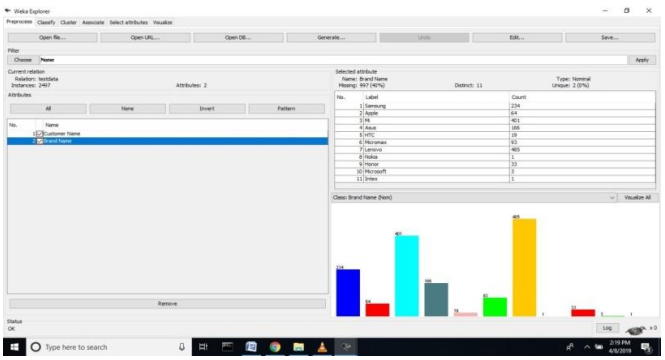

Fig. 3: Preprocessing of dataset

Table 1: Number of Iterations

Name of Dataset	K-means	Improved Algorithm
VM	10	8
SD	16	14
FK	29	27

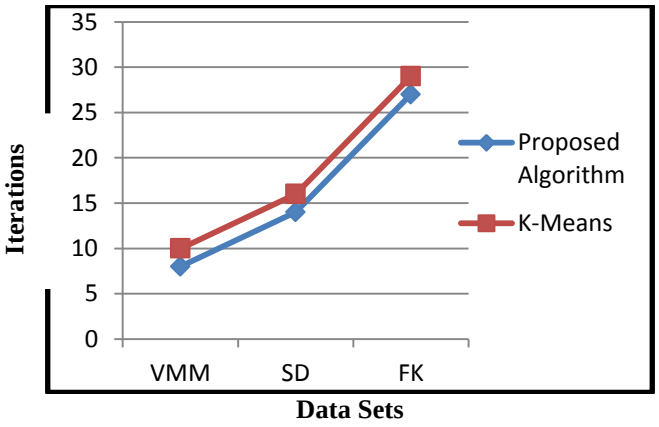

Fig 4: Graphical comparison for Iterations

Table 2: Accuracy

Name of Dataset	K-means	Improved Algorithm
VM	0.78	0.87
SD	0.02	0.04
FK	0.30	0.48

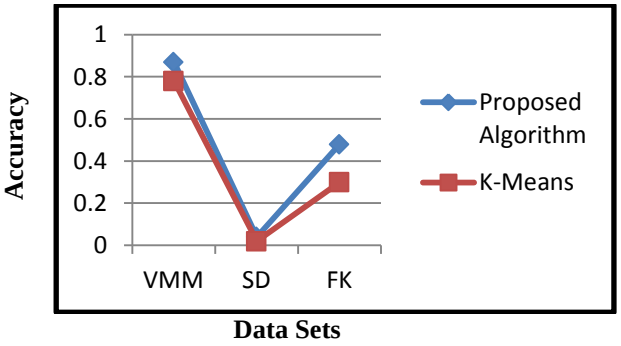

Fig 5: Graphical comparison for accuracy

Table 3: Execution Time

Name of Dataset	K-means	Improved Algorithm
VM	0.082	0.031
SD	0.041	0.038
FK	0.102	0.071

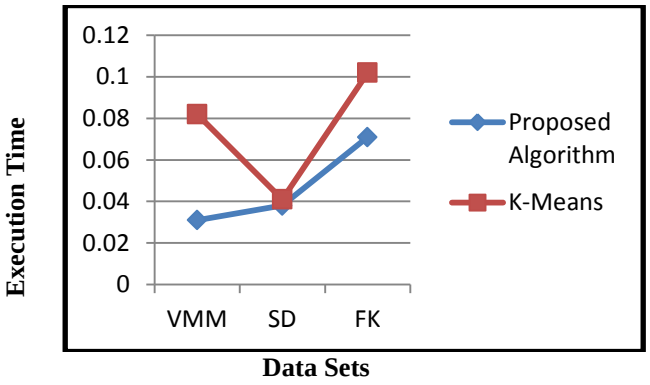

Fig 6: Graphical comparison for execution time

The result of all three experiments proves that the improved K-means algorithm produces better results than K–means in terms of iterations, accuracy and execution time.

VI. CONCLUSION

K-means algorithm is simple and commonly used for clustering large datasets. In this paper a improved K-means algorithm is developed to overcome the limitation of K-means algorithm. Improved K-means works on large datasets giving high accuracy, fast completion time and low complexity as compared to original K-means. This was possible due to reduced number of iterations as per the results shown above. We have also reduced the computational complexity to O(nk(t/2)) where t<<<n, of the new improved K-means algorithm which is far better than the complexity of original K-means which is O(nkt).

REFERENCES

[1] Jianhua Li and Laleh Behjat, "*A Connectivity Based Clustering Algorithm With Application to VLSI Circuit Partitioning*", IEEE Transactons On Circuis and Systems-II: Express Briefs, Vol.53, No. 5, May 2006.

[2] Lifei Chen , Shengrui Wang, Xuanhui Yan, "*Centroid-based clustering for graph datasets*.", 21st International Conference on Pattern Recognition , November 11-15, 2012.

[3] Xiaowei Xu, Martin Ester, Hans-Peter Kriegel, Jörg Sander , "*A distribution-based clustering algorithm for mining in large spatial databases*", Proceedings 14th International Conference on Data Engineering, 06 August 2002.

[4] Ester M., Kriegel H., Sander J., Xiaowei Xu, "*A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*", KDD'96, Portland, OR, pp.226-231, 1996.

[5] J.MacQueen, "*Some methods for classification and analysis of multivariate observations.",* In Proc. 5th Berkeley Symp. Math. Stat. Prob., 1:281–297, Berkeley, CA, 1967.

[6] Kaufman L. and Rousseeuw P. J., "Finding Groups in Data: An Introduction to Cluster Analysis*"*, John Wiley & Sons, 1990.

[7] Raymond T. Ng and Jiawei Han, "*CLARANS: A Method for Clustering Objects for Spatial Data Mining*", IEEE Transactions On Knowledge and Data Engineering, Vol. 14, No. 5, SEPTEMBER 2002.

[8] Zhang T, Ramakrishnan R., Livny M., " BIRCH: An efficient data clustering method for very large databases", In: SIGMOD Conference, pp.103~114, 1996.

[9] Guha S, Rastogi R, Shim K., " CURE: An efficient clustering algorithm for large databases", In: SIGMOD Conference, pp.73~84, 1998.

[10] Ankerst M., Markus M. B., Kriegel H., Sander J., "OPTICS: Ordering Points To Identify the Clustering Structure", Proc.ACM SIGMOD'99 Int. Conf. On Management of Data, Philadelphia, PA, pp.49-60, 1999.

[11] Fahim A M,Salem A M,Torkey F A, "*An efficient enhanced K-means clustering algorithm*" Journal of Zhejiang University Science A, Vol.10, pp:1626-1633,July 2006.

[12] Sun Jigui, Liu Jie, Zhao Lianyu, "*Clustering algorithms Research*",Journal of Software ,Vol 19,No 1, pp.48-61,January 2008.

[13] Zhe Zang, Junxi Zhang, Huifeng Xue, "*Improved K-means clustering algorithm*", Congress on Image and Signal Processing, IEEE DOI 10.1109/CISP.2008.

[14] Malay K. Pakhira, "*A modified K-means algorithm to avoid empty clusters*" International Journal of Recent Trends in Engineering, Vol 1, No. 1, May 2009 .

[15] Shi Na, Liu Xumin, Guang Yong, "*Research on K-means clustering algorithm: An improved K-means clustering algorithm* " Third International Symposium on Intelligent Information Technology and Security Informatics, IEEE, DOI 10.1109/IITSI.2010.

[16] Mumtaz, Dr. K. Duraiswamy, "*A novel density based improved K-means clustering algorithm- Dbkmeans*" International Journal on Computer Science and Engineering ISSN : 0975-3397 213 Vol. 02, No. 02, 2010, 213-218.

[17] Juntao Wang, Xiaolong Su, "An improved K-means clustering algorithm" IEEE, 3rd ICCSN International Conference on Communication Software and Networks, 2011.

[18] Navjot Kaur, Jaspreet Kaur Sahiwal, Navneet Kaur, "*Efficient K-means clustering algorithm using ranking method in datamining*" ISSN: 2278 – 1323, International Journal of Advanced Research in Computer Engineering & Technology ,Volume 1, Issue 3, May 2012.

[19] Shyr-Shen Yu , Shao-Wei Chu , Chuin-Mu Wang , Yung-Kuan Chan , Ting-Cheng Chang, "*Two Improved K-means Algorithms*", Applied Soft Computing Journal ,2017.

