IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified

Vol. 5, Issue 12, December 2016

An Enhanced Global K-Means Algorithm for Cluster Analysis

Jitendra Pal Singh Parmar[1], Prof. Shivank Kumar Soni[2], Prof. Anurag Jain[3]

PG Scholar, CSE, RITS, Radharaman Institute of Technology and Science, Bhopal, India[1]

Asst. Prof., CSE, RITS, Radharaman Institute of Technology and Science, Bhopal, India[2]

HOD, CSE, RITS, Radharaman Institute of Technology and Science, Bhopal, India[3]

Abstract: Recently, so many methods have been invented to determine issues with the choice of starting points in K-Means clustering algorithm. The Global K-Means and the Fast Global K-Means algorithms both are basis of such methods. They frequently insert one cluster centre at a time. The Weighted Fuzzy C-Means algorithm is as well extremely admired for fuzzy basis data clustering. However these all clustering methods are immensely influenced through the extreme environment of high dimensional data values. Every data in the dataset has compound characteristics and the cost of some characteristics might be so huge that the significance of additional characteristic costs might be entirely overlooked in the clustering procedure. The complexity of utilizing high dimensional datasets in clustering process is well known. To resolve these difficulties and to get better clustering algorithm for huge high dimensional datasets we proposed an algorithm "an enhanced global k-means (EGKM) algorithm for cluster analysis". To calculate the performance of the both FGKM and EGKM algorithms we use six datasets: Letter, Car, Iris, Kddcup, Nursery, Ozone and Spambase. Our trial study shows that EGKM performs much better than FGKM for every high dimensional datasets.

Keywords: GKM, FGKM, K Means, Cluster Analysis.

I. INTRODUCTION

Data mining has engrossed a great deal of consideration in the information engineering and in the public as a sum total in current years, owing to the broad accessibility of enormous amounts of facts and the forthcoming must for converting such facts into constructive knowledge and information. The facts and information acquired may be applied for applications ranging as of scientific exploration, analysis of market, client custody, to manufacture control and fraud recognition.

Classification is the procedure of finding a function (or replica) that distinguishes and describes data concepts or classes, intended for the reason of being competent to make use of the replica to forecast the category of objects whose class label is unidentified. The derived model is based on the examination of a collection of training records (i.e., data items whose class label is well-known.

Cluster Analysis 'What is cluster analysis?' It is a sort of Complementary Prediction and Classification, which analyse class labelled data items, clustering analyses data items without consulting a known class label. In broad-spectrum, the class labels are absent in the training data merely for the reason that they are not acknowledged to set in motion with. The clustering can be applied to engender such labels. The objects are grouped or clustered based on the standard of maximizing the intra-class similarity and minimizing the inter-class similarity. That is, clusters of objects are formed so that objects surrounded by a cluster have soaring correspondence in contrast to one another, however are very disparate to objects in previous clusters. Every cluster which is produced can be viewed as a category of objects, from which conventions can be copied. The clustering can also facilitate taxonomy formation, i.e., the association of interpretations into a hierarchy of classes that group similar events together.

The need to infer and mine likely inferences from high-dimensional datasets has led over the past decades to the development of dimensionality reduction and data clustering techniques. Systematic and technical applications of such clustering methodologies include among others computer imaging, data mining and bio-informatics. One of the widely used and studied statistical methods for data clustering is the K-means algorithm, which was first introduced in and is still in use nowadays as the prototypical example of a non-overlapping clustering approach. The applicability of the K-means algorithm, however, is restricted by the requirement that the clusters to be identified should be well-discreted and of a usual, curved-produced geometry, a prerequisite that is often not met in rehearsal. In this perspective, two essentially discrete approaches have been proposed in the past to address these restrictions.

A basic problem that regularly gets invoked in a great variety of fields such as pattern recognition, image

DOI 10.17148/IJARCCE.2016.51285

processing, machine learning and statistics is the clustering problem [1, 2]. In its rudimentary form the clustering issue is defined as the problem of discovering cognate groups of data points in a provided data set. For each partition is known a cluster and can be stipulated as an area in which the density of objects is locally higher than in other regions. The unblended type of clustering is partition clustering which aims at dividing a given data set into distinct subsets (clusters) so that specific clustering criteria are modified. The most broadly applied criterion is the criterion of clustering error for which each point calculates its squared distance from the analogous cluster centre and then takes the total of these gaps for all locations in the data set. A famous clustering method that diminishes the error in clustering is the k-means algorithm. Even though, it is a local search method and it is quite popular that it suffers from the serious deficiency that its performance majorly depends on the initial starting conditions [3].

Different approaches to improve the efficiency of the k-means algorithm have been proposed [4, 5], of which incremental ones are among the most triumphant. In these approaches clusters are computed step by step by solving all intermediate clustering problems. The Global K-Means algorithm (GKM) proposed in [1] and the modified Global K-Means algorithm (FGKM) proposed in [6] are incremental clustering algorithms.

II. LITERATURE SURVEY

In this chapter we discussed about different preceding workings that had been previously anticipated by many researchers. A number of general approaches are also discussed here that work proficiently in the field of fast data mining, rapid clustering, K-means (KM), Fuzzy C-means (FCM), Global K-means (GKM), Fast Global K-means (FGKM), etc.

In “[10]”, this research paper emphasises on the geometric concept of Probabilistic Neural Network (PNN) for the complications of pattern categorization with EM (Expectation Maximization) preferred as the algorithm of training. This method gets about the complication of random initialization that signifies, the client has to pre described the quantity of clusters via experiment and miscalculation. GKM is utilized to determine this and to contribute a deterministic amount of clusters utilizing a selection standard. On above of that, FGKM was experimented as a replacement for GKM, to diminish the calculation time acquired. Potential purpose of this method is as an analysis miniature that might be utilized in the employment industry to observe the situation of aids, such as instruments, and to categorize them into their error modes emphasises on the input vectors obtained from sensors sets on the appliance.

In “[11]”, authors represented k-Means algorithm and its variations and found all these algorithms are perceptive to the selection of start positions and incompetent for determining clustering problems in high dimensional datasets. In this research paper, a novel edition of the GKM algorithm is recommended known as the modified global k-means (MGKM) algorithm. A start position for the kth cluster centre is calculated by diminishing a secondary cluster service. The experimental results on high dimensional datasets exhibit the supremacy of the introduced algorithm, nevertheless, it involves more calculation time than the GKM algorithm.

In “[12]”, the Kernel k-means method is an augmentation of the ordinary k-means method that only recognizes nonlinearly distinguishable clusters. With the intention to defeat the flaws related with this technique, in this research work authors recommended the global kernel k-means algorithm, which is a deterministic and incremental method based on kernel clustering. Their technique inserts one cluster at every phase via a universal investigation process containing of numerous implementations of kernel k-means from appropriate selections. This algorithm is doesn’t rely on cluster selection, recognizes nonlinearly distinguishable clusters and, because of its incremental environment and investigate process, it situates by optimal explanations preventing insufficient local minima. Additionally an amendment is recommended to diminish the calculation rate that doesn’t appreciably influence the result superiority. This introduced global kernel k-means algorithm maps data positions from input field to a higher dimensional characteristic field by the utilization of kernel operation and enhances the clustering inaccuracy in the characteristic field through establishing closed optimal minima. The key improvements of this technique are its deterministic character that constructs it autonomous of cluster selection, and the skill to recognize nonlinearly distinguishable clusters in input field.

In “[13]”, in this research paper authors craft investigation of two modification of k-means method known as GKM and X Means algorithms above colon dataset. On colon dataset it has to categorize it into two comparable groups. Precision of GKM is somewhat in excess of precision of X Means algorithms. Amounts of tests to attain a universal and constant optimal result are less for both the GKM and X Means algorithms. Acceleration of implementation is quick for X Means algorithms in comparison to GKM algorithm. X means doesn’t need starting value of no. of clusters (K). Actually x means repeatedly choose an appropriate value of k itself utilizing lower and upper bound. Starting choice of cluster centres doesn’t influence the excellence of clusters. Concurrence rate of both algorithms is enhanced in comparison to k-means.

In “[14] The GKM algorithm recommended by Likas in 2003 is an incremental method of clustering that effectively inserts single cluster centre at a time by a deterministic universal investigate process containing of N (through N being the amount of the dataset) executes of the K-means method from appropriate start points. But this

DOI 10.17148/IJARCCE.2016.51285

method has a serious calculation load. In this research paper, authors recommended a novel description of the GKM. The exceptional characteristic of this method is its advantage in time of execution. It acquires fewer execution times than the GKM algorithms. This improvement is because of that authors enhanced the technique of constituting the subsequent cluster centre in the GKM algorithm.

They described a novel operation to choose the excellent candidate centre for the subsequent cluster progressive by the scheme of K-medoids method. Trials on recognized UCI datasets illustrate that the introduced algorithm might considerably decrease the calculation time without influencing the accomplishment of the GKM method.

In “[15]”, the Multi Granulations proximity estimation field is a novel universal model of estimation fields, in those topology regions are provoked by multi search operations with numerous group characteristics. In this research paper, through joining GKM clustering method and topology regions, authors introduced two k means algorithms, in that AFS topology regions are occupied to establish the clustering start positions. The introduced method might be enforced to the datasets with arithmetic, Boolean, grammatical ranking level, sub favourite relations characteristics. In this research paper, the AFS GKM method are introduced, which is based on AFS topology region in the phase of determining start cluster centres. The methods are autonomous of the start situations that permit enforcing with numerous group characteristics.

III. PROPOSED WORK

Here In this segment we discuss about suggested technique “An Enhanced Global K Means (EGKM) Algorithm for Cluster Analysis”. The EGKM algorithm is recommended for large datasets. Thinking a dataset X as {x1, x2, x3, . . . , xN}, at first the method divides X into M clusters as C = C1, C2, . . . , CM and determine all cluster’s centre so that the value operation (objective operation) of variation quantify is reduced or limited than a definite brink value. The objective operation is similar as given in equation 1. The cluster centres are described as m1, . . . , mM.

evertheless for the given dataset, initially it discrete dataset into large pieces X1, X2, . . . , Xs according to the arrival time of data, and the dimension of all large pieces are find out by core memory of the dispensation system, let n1, n2, . . . , ns be the data integers of large pieces X1, X2, . . . , Xs correspondingly . Caused by its dataset setting, a weighted time w (t) is puts on all data signifying the data impact expansion on the clustering method, and

$$\int_{t_0}^{t_c} w(t)dt = 1.$$

Where t0 is the start time of arrival of data and tc is the recent time.

The core idea of EGKM is enhancing the weighted clustering centres through recurrence till the charge operation acquires a fulfilling outcome or the quantity of recurrence is to a lenience. Additionally, through the procedure, we assign the individual a stable weight as 1.

The complete process is showed as follow:
1) Bring in the large piece Xl (1 ≤ l ≤ s).
2) Revise the cluster centroids weight.

• If l = 1: employ FGKM algorithm to get cluster centroids

$$w'_i = \sum_{j=1}^{n_1} (u_{ij}) w_j \qquad\qquad 1 \le i \le M$$

Where wj = 1, ∀ 1 ≤ j ≤ n1.

• If l > 1:

$$w'_i = \sum_{j=1}^{n_l+M} (u_{ij}) w_j \qquad\qquad 1 \le i \le M$$

Where wj = 1, ∀ M + 1 ≤ j ≤ nl + M.

The weight of centroid w_i then revises as wi = w’i
3) Revise cluster centroids as revised in FGKM algorithm.
4) Calculate objective operation:

$$E(m_1,\ldots,m_N) = \sum_{i=1}^{c+n_1} \sum_{k=1}^{M} I(x_i \in C_k)*w_k*||x_i - m_k||^2$$

End if objective operation is decreased or contemplate on a definite cost, or its enhancement over prior recurrence is under a definite brink, or recurrences arrive at a definite patience value.
5) Calculate a fresh U utilizing Equation 4. Go to step 2.
6) If l = s then end, otherwise go to step 1.

U could be computed as:

$$u_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{d_{ik}}{d_{jk}}\right)^{2/(m'-1)}}$$

Where uik is the membership cost of the kth data x_k in the ith cluster since Global K Means and the Fast Global K Means methods are on the whole dataset, data might include a extremely huge dataset, so permitting Global K-Means and the Fast Global K-Means algorithms to contract with data straight might use major quantities of CPU time to touch, or outcome in an insufferable recurrence amount. Our intended EGKM algorithm decreases such complications and gets much smaller implementation time and memory utilization as contrasted to both GKM and FGKM algorithms.

IV. EXPERIMENTAL ANALYSIS

To compute the supremacy of our proposed algorithm EGKM, We implemented both FGKM and EGKM algorithms in NetBeans IDE (java) to trial the efficiency of both the methods. Our trial study demonstrates that EGKM executes much enhanced than FGKM for datasets having high dimensional. We employ the experiments on six datasets: Letter, Car, Iris, Kddcup, Nursery, Ozone and Spambase. All these datasets are accessible in UCI repository.

A. Execution Time

We consider the time of execution for EGKM by inserting the implementation of all data fragment. As the dataset is high dimensional so there is no assurance about the size of arrival of the data. Therefore we avoid the data entrance interruptions from execution time. Table I, II and III demonstrates the evaluation of execution time for both algorithms FGKM and EGKM for Letter, Kddcup, and Nursery datasets. The outcomes obviously demonstrate that EGKM decreases execution time on an average of 65.53%.

Table I: Comparison of execution time and memory utilization of Kddcup Dataset for both FGKM and EGKM algorithms

KDD Cup	EXECUTION TIME		MEMORY UTILIZATION	
No. of Clusters	**FGKM**	**EGKM**	**FGKM**	**EGKM**
5	953	187	46000	37640
10	1609	313	51430	41680
15	2252	643	55628	45720
20	2922	1204	59687	49760
25	3515	2000	63436	53800

Table II: Comparison of execution time and memory utilization of Letter Dataset for both FGKM and EGKM algorithms

LETTER	EXECUTION TIME		MEMORY UTILIZATION	
No. of Clusters	**FGKM**	**EGKM**	**FGKM**	**EGKM**
5	1547	203	29200	21160
10	2609	344	35872	26000
15	3830	641	38457	30840
20	4860	1187	42758	35680
25	5922	2000	47861	40520

Table III: Comparison of execution time and memory utilization of Nursery Dataset for both FGKM and EGKM algorithms

NURSERY	EXECUTION TIME		MEMORY UTILIZATION	
No. of Clusters	**FGKM**	**EGKM**	**FGKM**	**EGKM**
5	985	140	20200	11240
10	1640	313	24688	15280
15	2265	453	29467	19320
20	2890	844	33571	23360
25	3469	1657	38645	27400

Changing situation of execution time: It has been examined that both FGKM and EGKM algorithms demonstrate minor difference in execution time for several runs with similar constraints. This is due to the hypothesis acquired for FGKM. However the differences aren’t extremely high and could be adequate. Nevertheless for precise outcomes we ran both algorithm several times (with similar dataset and constraints) and acquired average of them as concluding execution time.

B. Memory Utilization

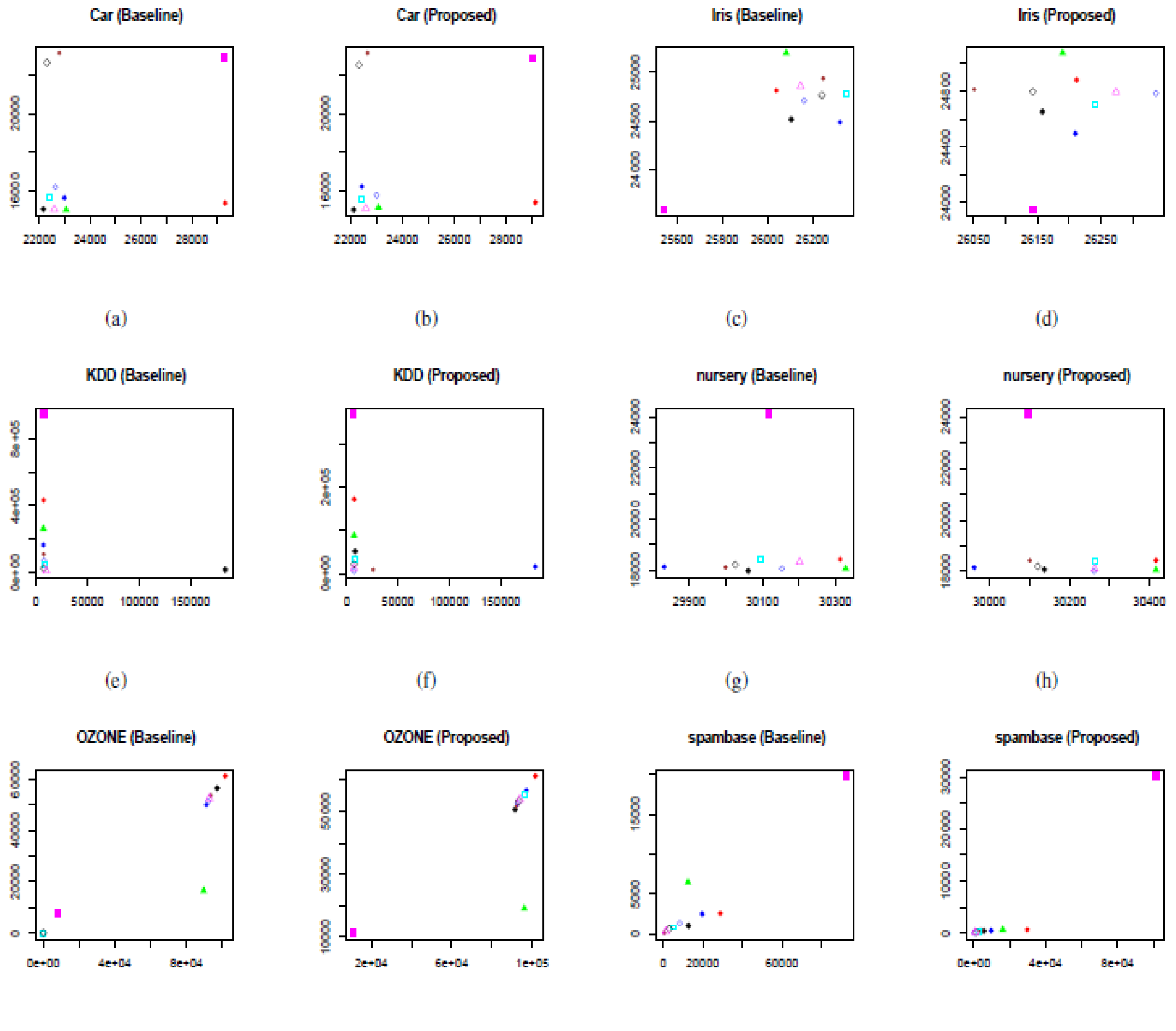
As EGKM method enter dataset as amount of large pieces as a result the proposed method computed the memory utilization of all large piece discretely and obtain the biggest value as the concluding memory utilization for EGKM.

Table I, II and III also exhibits the proportion of enhancement in tenures of memory utilization by proposed

DOI 10.17148/IJARCCE.2016.51285

EGKM method in comparison with the FGKM algorithm. The enhancement is more than 68% for each dataset utilized i.e. Letter, Kddcup, and Nursery. The FGKM employs complete dataset at an instant and therefore it wants adequate memory to seize the complete dataset. Due to this reason the FGKM involves much elevated memory utilization than the proposed EGKM method.

Memory utilization issues: The EGKM method has a membership matrix which isn't present in FGKM method. When the amount of fragments in EGKM is less the memory needed by the membership matrix is high and use the advantage achieved from EGKM. However as the amount of fragments raise the advantage could be scrutinized. In our trials it is supposed the amount of fragments is 50, which is standard for these days' high dimensional dataset.



(a) (b) (c) (d)

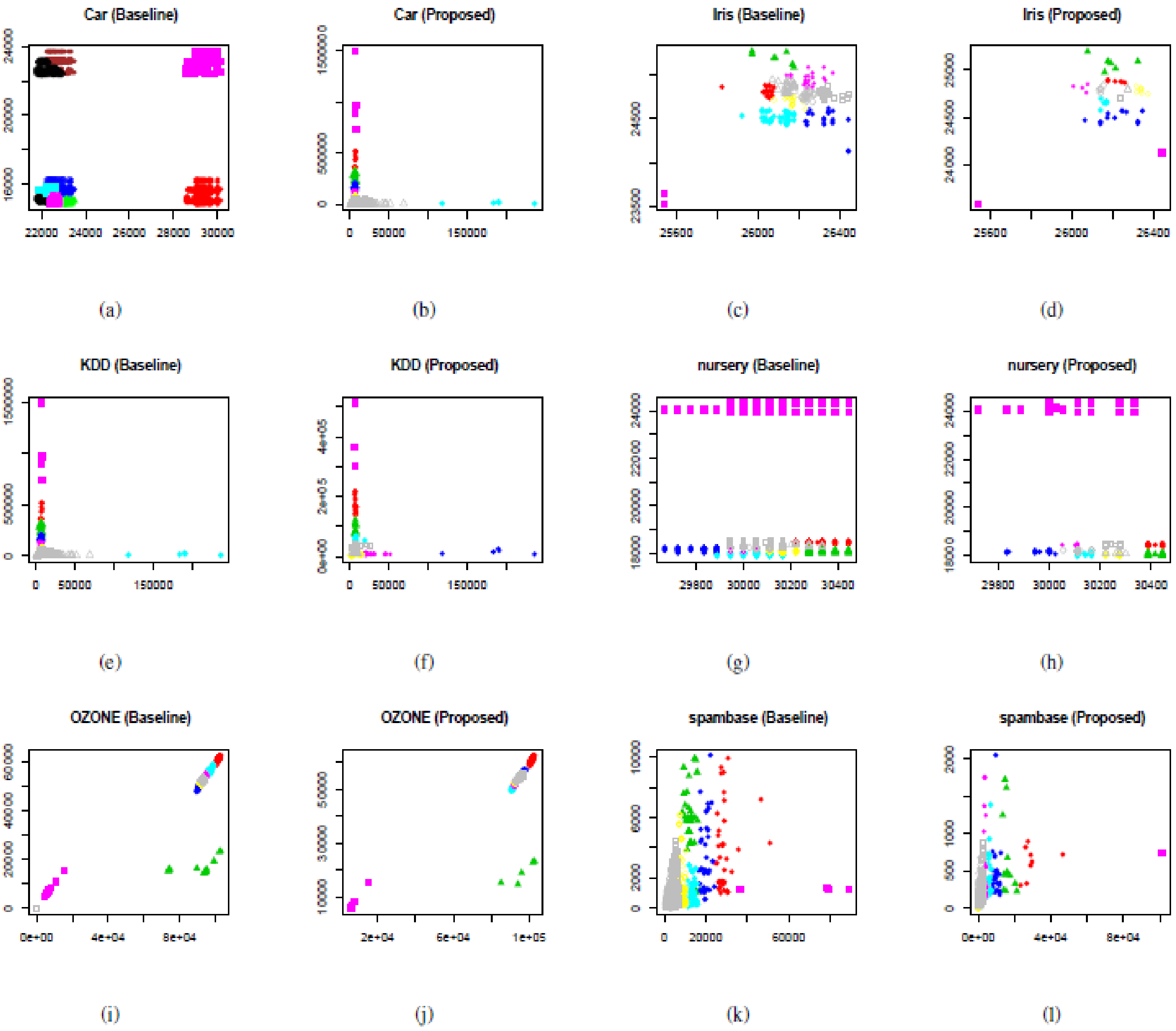
(e) (f) (g) (h)

(i) (j) (k) (l)

Figure 1: cluster centres diagram of FGKM and EGKM method for all six dataset

C. Cluster analysis

This section illustrates the analysis of clustering based on EGKM method. Six special datasets has been utilized for the purpose of analysis. All datasets are implemented for 10 clusters and the outcomes are illustrated for both the algorithms FGKM and EGKM. Figure 2 shows the cluster-centres of each dataset discretely. The graphs of both FGKM and proposed EGKM for each dataset are given one after another and can be recognized by the heading given in each graph. Figure 1 show that the cluster centres are approximately similar in both FGKM and EGKM algorithm. Consequently our algorithm EGKM provides the vast enhancement in time of execution and memory utilization without any key clustering variations. The cluster diagram for all datasets of both FGKM and EGKM algorithms are shown in Figure 2. It should be Noteworthy that the EGKM algorithm splits the dataset in several fragments and thus the cluster dimension of FGKM and EGKM isn't similar. However their prototypes are approximately seems similar



(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j) (k) (l)

Figure 2: clusters diagram of FGKM and EGKM method for all six dataset

V. CONCLUSION

Clustering is a broadly analysed problem in various applications such as WSN (wireless sensor networks), pattern recognition, etc. Here we have put our endeavour to implement the FGKM and EGKM clustering algorithms, both. We recognized the broad framework of an Enhanced Global K Means basis consent clustering and provided the corresponding algorithm by boosting the performance of FGKM. We also extended the scope of FGKM to the cases where there exists incomplete basic clustering.

Experiments on six actual world datasets (Letter, Car, Iris, Kddcup, Nursery, Ozone and Spambase) have demonstrated that EGKM is highly efficient w.r.t. time & memory and the similar clustering performance with modern methods. In future the proposed approach could be useful for the K-Medoid algorithm also, which is also a partitioning method. Some other high dimensional datasets could also be utilized for future experiments. Data balancing of global Fuzzy C-means is a possible extension of the proposed work.

REFERENCES

1. A. Likas, M. Vlassis, and J. Verbeek, "The global k-means clustering algorithm," Pattern Recognition, vol. 35, no. 2, pp. 451–461, 2003.
2. S. Theodoridis and k. Koutroumbaspattern, Pattern Recognition, 2nd edition, Elsevier, 2003
3. L. Bai, J. Liang, C. Sui and C. Dang, "Fast global k-means clustering based on local geometrical information," Information Sciences, vol. 245, pp. 168 – 180, 2013.
4. Hongjun Maciej Jaworski, Piotr Duda, and Lena Pietruczuk "On Fuzzy Clustering of Data Streams with Concept Drift" 2012
5. S. J. Redmond and C. Heneghan, "A method for initializing the K-means clustering algorithm using kd-trees," Pattern Recognition Letters, vol. 28, pp. 965–973, 2007.
6. Adil M. Bagirov, Julien Ugon, Dean Webb, "Fast modified global k-means algorithm for incremental cluster construction" in Proceedings of the Centre for Informatics and Applied Optimization, Graduate School of Information Technology and Mathematical Sciences, 2010.
7. V. Ramasubramanian and K. Paliwal, "Fast K–Dimensional Tree Algorithms for Nearest Neighbour Search with Application to Vector Quantization Encoding". IEEE Transactions on Signal Processing, 40: (3) March 1992.
8. A. M. Bagirov, "Modified global k-means algorithm for minimum sum -of-squares clustering problems," Pattern Recognition, vol. 41, pp. 3192– 3199, 2008.

9. T. Kanungo and D. Mount, “An efficient k-means clustering algorithm: analysis and implantation,” IEEE Trans, PAMI, vol. 24, pp. 881–892, 2004
10. Chang, Roy Kwang Yang, Chu Kiong Loo, and M. V. C. Rao. "A global k-means approach for autonomous cluster initialization of probabilistic neural network." Informatica 32, no. 2 (2008).
11. Bagirov, Adil M. "Modified global k-means algorithm for minimum sum-of-squares clustering problems." Pattern Recognition 41, no. 10 (2008): 3192-3199.
12. Tzortzis, Grigorios, and Aristidis Likas. "The global kernel k-means clustering algorithm." In 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), pp. 1977-1984. IEEE, 2008.
13. Kumar, Parvesh, and Siri Krishan Wasan. "Analysis of X-means and global k-means USING TUMOR classification." In Computer and Automation Engineering (ICCAE), 2010 The 2nd International Conference on, vol. 5, pp. 832-835. IEEE, 2010.
14. Xie, Juanying, and Shuai Jiang. "A simple and fast algorithm for global K-means clustering." In Education Technology and Computer Science (ETCS), 2010 Second International Workshop on, vol. 2, pp. 36-40. IEEE, 2010.
15. Wang, Lidong, Xiaodong Liu, and Yashuang Mu. "The Global k-Means Clustering Analysis Based on Multi-Granulations Nearness Neighborhood." Mathematics in computer science 7, no. 1 (2013): 113-124.

DOI 10.17148/IJARCCE.2016.51285