Global K-Means (GKM) Clustering Algorithm: A Survey

Arpita Agrawal
PG Scholar
PIES, Bhopal, INDIA

Hitesh Gupta
Head, CSE department
PIES, Bhopal, INDIA

ABSTRACT

K-means clustering is a popular clustering algorithm but is having some problems as initial conditions and it will fuse in local minima. A method was proposed to overcome this problem known as Global K-Means clustering algorithm (GKM). This algorithm has excellent skill to reduce the computational load without significantly affecting the solution quality. We studied GKM and its variants and presents a survey with critical analysis. We also proposed a new concept of Faster Global K-means algorithms for Streamed Data sets (FGKM-SD). FGKM-SD improves the efficiency of clustering and will take low time & storage space.

Keywords
Clustering, K-means, GKM, FGKM, Streamed Dataset

1. INTRODUCTION

The clustering concept has been introduced before a long time ago. Clustering is a process in that we classify objects that are one way or another common in properties. The main aim of the clustering is to offer a combination of similar records. The term classification and clustering is confusing, but they have the difference that in classification objects allocates in predefined classes while in the clustering classes is created. In database management, clustering is a process where, physically stored information is similar to logical information. To make efficient search and rescue in database, several disk admittance to be reduced. Objects having similar properties are grouped in the same class of objects.

In the field of data mining, image processing, pattern recognition, machine learning, vector quantization, etc. Data clustering is used often, whose goal is to partition data into similar groups. The well known clustering algorithm is K-means, in that similar groups are formed through reducing the clustering error described as the sum of the squared Euclidean distances among each dataset point and the corresponding cluster center [3]. This k-means algorithm has two problems. First its results too much depends on the initial positions of the cluster centers [4], which resulting as easily stuck to the local optimal solutions, and the second problem is only linearly separable clusters can be discovered by this algorithm.

Number of solutions were proposed to overcome these drawbacks of k-means algorithm included simulated annealing, genetic algorithms, etc. Because of impractical implementation these techniques had not accepted much and almost applications used the k-means algorithm with multiple restarts [5]. In recent years these some versions are defined by researchers to improve the efficiency of k-means algorithm.

A method for initializing the K-means algorithm proposed which starts by at random contravention [8] of the data. It executes the K-means clustering algorithm by starting at the same set of initial seeds, Algorithm will executed 10 times on the randomly chosen sets. Every run should be initialized using the K final centroid locations from one of the run of 10 subsets. The K center location that we get from this run will be used to initialize the K-means algorithm for the complete data set.

The Global K-Means algorithm (The GKM algorithm) [2] is the incremental approach of clustering. We can dynamically add one cluster center at a time using deterministic global search procedure from suitable initial positions. It is consists of N (Where N is the size of the dataset) executions of K-means algorithm. Experimental results of the algorithm show that GKM algorithm considerably out performs the K-means algorithms.

In this paper we present a study of GKM and its variants which have overcome the basic problems of GKM. Our study is focused on the limitations of GKM & its variants and we also proposed a new algorithm for clustering-A Faster Global K-Means algorithms for Streamed Datasets (FGKM-SD).

In the following section we formerly define K- Mean's algorithm. In Section III we describe Global *K*-Means algorithm. In Section IV related work of GKM has been defined. We evaluate GKM & its variants by critical analysis in Section V. The motivation of our work is shown in section VI and section VII we explain our proposed work. Finally we conclude in our paper in section VIII.

2. THE K-MEANS ALGORITHM

K-mean clustering algorithm states, clusters are entirely reliant on the choice of the initial cluster centroids. K data elements are selected as initial centers; then distance of all data elements is deliberate by Euclidean distance formula. Data elements having less distance to centroids are stimulated to the appropriate cluster. The process is sustained until no more alteration occurs in clusters. The following figure 1 shows the steps of the K-mean clustering algorithm [7]. Following are the algorithmic steps for K-mean algorithm [6].

Algorithm 1:

INPUT: Number of desired clusters *K* Data objects D= {d1, d2…dn}

OUTPUT: A set of *K* clusters

Steps:

- Randomly elevate K data objects (as initial centers) from data set D.
- Repeat;
- Calculate the distance of each data object d_i (1 <= i<=n) from all *k* clusters C_j (1 <= j<=k) and then assign data object d_i to the nearest cluster.
- For each cluster j (1 <= j<=k)
- Recalculate the cluster center until no change in the center of clusters.

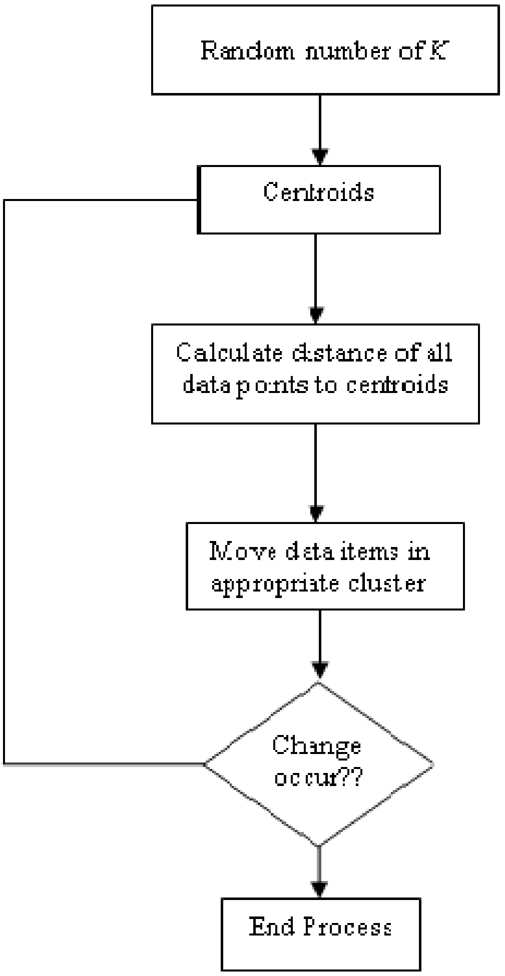
Random number of K
Centroids
Calculate distance of all data points to centroids
Move data items in appropriate cluster
Change occur??
End Process

Fig.1: K-Means clustering process

O (*nkt*) is the time complexity of K-mean Clustering.

Where n represent the number of objects, K represent number of clusters and t represent iterations. Limitations of K-means Algorithms are:

- It's requiring a user to give out the number of clusters at first, and its sensitiveness to initial conditions.
- While using the K-mean algorithm the computational time increases on implementing in large amount of data.
- On using a large amount of data set the storage space increases in K-mean algorithm

3. THE GLOBAL K - MEANS ALGORITHM

The Global K-Means (GKM) algorithm minimized clustering error using deterministic useful global clustering algorithm. The k-means algorithm as a local search procedure, this algorithm follows incremental approach to solve M cluster clustering problem, all intermediate problems with 1, 2… M −1 clusters are sequentially solved [2].

The GKM method provides optimal solution for clustering problem with M clusters through a series of local searches (using K-means algorithm). For each local search, M-1 cluster centers must be initially placed in their optimal positions corresponding to the clustering problem with M-1 clusters. Then placed the remaining M^{th} cluster center at several positions within the data space. If for M=1 the optimal solution is known, we can iteratively apply the above procedure to get the 2^{nd} optimal solutions for all k-clustering problems K=1, 2…, M.

This effective method is not only deterministic but it also does not depend upon any initial conditions or empirically adjustable parameters. Above mentioned points are the significant advantages of overall clustering approaches.

The global k-means algorithm successively computes the clusters. For first iteration, the centroid of set A is computed. Similarly for computing k-partition, k-th iteration of this algorithm uses k-1 clusters centers from the previous iteration. We can describe the global k-means algorithm for the computation of q ≤ m clusters in a data set A are as follows.

Algorithm 2: The global k-means algorithm

Step 1 (Initialization) Compute the centroid x^1 of the set A:

$$x^1 = \frac{1}{m}\sum_{i=1}^{m} a^i, a^i \in A, i = 1, \dots\dots, m$$

And set k = 1.

Step 2 Set k = k + 1 and consider the centers

$x^1, x^2, \dots, x^{k-1}$ from the previous iteration.

Step 3 Each point of A is the starting point for the k-th cluster center, To obtain m initial solutions with k points $(x^1, \dots, x^{k-1}, a)$; k-means algorithm is applied to each of them; keep the best k-partition obtained and its centers $x^1, x^2, \dots, x^k$.

Step 4. (Stopping criterion) If k = q then stops, otherwise go to Step 2.

This version of the algorithm is not applicable for clustering on middle sized and large data sets.

4. RELATED WORK

Researchers are continuously working on global K-means (GKM) algorithm and they introduced numbers of techniques which has more simplicity and efficiency than GKM. Several of them are described here:

The modified global k-means algorithm [9] was developed for clustering in gene expression data sets which is effective for solving clustering problems in gene expression data sets. This algorithm computes clusters incrementally and to compute k-partition of a data set it uses k − 1 cluster centers from the previous iteration. Computation of the starting point for the k^{th} cluster center is the key point. Starting point is calculated by minimizing so-called auxiliary cluster functions.

For Autonomous Cluster Initialization of Probabilistic Neural Network [10] an approach was demonstrated. In this approach statistical based Probabilistic Neural Network (PNN) was used for pattern classification problems with Expectation – Maximization (EM) chosen as the training algorithm. Global algorithm solves the problem of random initialization. Initially, user needs to predefine the number of clusters using trial and error method. Global K-means provides a deterministic number of clusters using a selection criterion. This model was well tested with Fast Global k-means to ensure their correct classification and computational times.

Result shows that FGKM provided relatively close accuracy and improved computational time.

K-Means algorithm and its variations are the fastest clustering algorithms but they are sensitive to the choice of starting points and inefficient for solving large data sets clustering problems. Recently new version of the k-mean algorithm has been developed, which is known as global k-means algorithm [11].

Global k-means is the incremental algorithm that allows us to add one cluster center at a time and uses each data point as a candidate for the *k*-th cluster center. Experimental results show that the global *k*-means algorithm considerably outperforms the *k*-means algorithms. New version of this algorithm is proposed in this paper, it uses minimizing an auxiliary cluster function to compute the starting point for the *k*-th cluster center. Numerical results of these experiments (i.e. 14 data sets) demonstrate the superiority of the new algorithm, however it required more computational time than global k-mean algorithm.

Based on colon dataset, global k-means and x-means algorithms were analyzed [12]. Comparison was made in respect of accuracy and convergence rate**.** Accuracy of global k-means is slightly more than accuracy of x-means. Number of trials to reach a global and a stable optimum solution is less for both the algorithms. Speed of execution is the fastest for x-means in comparison to global k-means.

The recent version of GKM algorithm presented a new method of how the next new cluster center is created by using some idea of K-mediods clustering algorithm suggested by Park and Jun [13]. This algorithm will help not only to reduce the computational load of the GKM without affecting the performance of it, but also avoid the influence of the noisy data on clustering result. It requires much less calculation amount and shows less computational complexity. The distance between each pair of objects is computed only once, which contributes to the excellent feature. At the same time, the selection of the next cluster initial center can avoid the impact of noisy data on the clustering result.

A New version of GKM algorithm uses auxiliary cluster function to compute the starting point for the *k*-th cluster center by minimizing it [14]. The difference between the FGKM algorithm and new version is in the way of starting point for the *k*-th cluster center is obtained. A local minimizer used as starting point for the *k*-th cluster center. This algorithm not only used to compute the cluster incrementally but also to compute *k*-partition of a data set, it uses $k-1$ cluster centers from the previous iteration.

Incremental approach is the recent thing that is developed to resolve difficulties with the choice of starting points. The modified global k-means and the global k-means algorithms are based on such an approach that they iteratively add one cluster center at a time. Numerical experiments show that these algorithms considerably improve the k-means algorithm. However, they require storing the whole affinity matrix or computing this matrix at each iteration this makes both algorithm's time consuming and memory demanding for clustering even moderately large data sets. An auxiliary cluster function introduced to generate a set of starting points lying in different parts of the data set. This method exploits information gathered in previous iterations of the incremental algorithm to eliminate the need of computing or storing the whole affinity matrix and thereby to reduce computational effort and memory usage. This algorithm computes clusters incrementally, using the k-1 cluster centers from the previous iteration to solve the k-partition problem. An important step in this algorithm is the computation of a starting point for the k-th cluster center. This starting point is computed by minimizing the so-called auxiliary cluster function. Unlike the modified global k-means algorithm the proposed algorithm does not rely on the affinity matrix to compute the starting point.

A Fast global k-means clustering algorithm is introduced by making use of the cluster membership and geometrical information about a data point [15]. This algorithm is referred to as MFGKM. The algorithm uses a set of inequalities developed to determine a starting point for the j^{th} cluster center of global k-means clustering. Adopting multiple cluster centers election (MCS) for MFGKM, another clustering algorithm was also developed called MFGKM+MCS. MCS determines more than one starting point for each step of cluster split; while the available fast and modified global k-means clustering algorithms select one starting point for each cluster split.

The fast global k-means (FGKM) clustering algorithm is one of the most effective approaches for resolving the local convergence of the k-means clustering algorithm. Numerical experiments show that it can effectively determine a global or near global minimizer of the cost function.

However, the FGKM algorithm needs a large amount of computational time or storage space when handling large data sets. To overcome this deficiency, a more efficient FGKM algorithm, namely FGKM+A [17], is developed in this paper. In this development, firstly apply local geometrical information to describe approximately the set of objects represented by a candidate cluster center. On the basis of the approximate description, As a result of the acceleration, the FGKM+A algorithm not only yields the same clustering results as that of the FGKM algorithm but also requires less computational time and fewer distance calculations than the FGKM algorithm and its existing modifications.

Multi-Granulations nearness approximation space is a new generalized model of approximation spaces, in which topology neighborhoods are induced by multi probe functions with many category features [16]. In this paper, by combining global k-means clustering algorithms and topology neighborhoods, two k-means clustering algorithms are proposed, in which AFS (Axiomatic Fuzzy Sets) topology neighborhoods are employed to determine the clustering initial points.

The AFS global k-means algorithms are introduced, in which the distance based on the AFS topology neighborhood is employed in the step of determining initial cluster centers. The algorithms are independent of the initial conditions, which allow working with many category features. They were tested with data sets containing many category features. Although the proposed algorithms have an accuracy of equal or better quality compared against the k-means, the global k-means and fast global k-means on some data sets.

5. CRITICAL ANALYSIS

Many researchers' concern is to apply GKM Algorithm for increasing computation efficiency of clustering algorithm, clear and fast identification initial starting point. So here we analysis some of previous work done on GKM algorithm.

Table 1: Analysis of GKM methods

Method	Data set used	Problem	Solution
GKM [2]	Iris Synthetic Image -segmentation	Local search procedure, Depends on the initial starting conditions	Clusters can be obtained using a series of local searches It does not depend on any initial positions for the cluster center Providing excellent results in terms of the mean square clustering error criterion
Modified GKM [9]	Boston Lung Cancer Novartis multi-tissue Leukemia 3 more data sets of genes	Converge only to local minima	Computes clusters incrementally Compute k-partition of a data set it uses k − 1 cluster centers from the previous iteration. Starting point is computed by minimizing so-called auxiliary cluster function
EM-based PNN using GKM [10]	Medical , Iris Benchmark dataset Westland	Random initialization	A deterministic number of clusters using a selection criterion
MGKM [11]	German towns Bavaria postal 1 Fisher's Iris Plant Pima Indians TSPLIB1060 Image Segmentation Ionosphere Congressional Voting Records	Unsupervised classification of patterns, Unable to locate either a global minimize or a local minima	Computes clusters incrementally Works better with data sets, which do not have well separated clusters.
Simple & fast GKM [13]	Soybean-small Iris ,Wine Segmentation Liver Disorders Artificial Pima Indians Diabetes	Heavy computational load	Use the idea of K-medoids clustering algorithm and reduced heavy computational load
MFGKM [14]	one synthetic data six real data sets	Computational complexity	a set of inequalities developed to determine a starting point for the j^{th} cluster center of global k-means clustering method uses CGAUCD (codebook generation algorithm using code word displacement)
AFS global K-means [16]	Iris , New thyroid Wine Haberman BCW(O), Pima Mammo, ballon	Initial starting conditions	Distance based on AFS topology neighborhood is employed in the step of determining initial cluster centers. independent of the initial conditions
FGKM+A [17]	Handwritten digits Statlog , Musk Isolet , Coil Letters , Shuttle Corel image	large amount of computational time or storage space	Local geometrical information to describe approximately the set of objects represented by a candidate cluster center.

6. MOTIVATION

The great advantage of GKM is that it is improving the way of creating the next cluster center and it defined a novel function to select the optimal candidate center for the next cluster. But GKM also have the problem that it takes much time to compute large data sets. It also takes large space to implement large data sets. When we will apply GKM of high definition (HD) data sets or streamed data sets than this will cause the same problem of storage and time.

7. RESEARCH SCOPE

A faster global K-means clustering algorithm for streaming data sets can be developed. This algorithm will be applicable for those datasets that have streaming behavior. An example of such data sets is live videos on the internet. A faster global K-means algorithm is available as we discussed above but the algorithm is for a normal data set and not for streaming data sets. Our proposed algorithm will improve the efficiency of the FGKM clustering algorithm. The performance of the proposed algorithm is more remarkable as the number of dimensions or clusters of a data set increases. We will develop a strategy to accelerate the speed of clustering.

8. CONCLUSION

In this paper we study different GKM clustering algorithms and examine their advantages and limitation. GKM provides solution to overcome the drawbacks of k-means algorithm but it has its own limitations like slow execution and large space requirement. To reduce these drawbacks of GKM number of solution and methods had been proposed which was efficient in comparison to GKM. We summarized most of them in our critical analysis section. We also proposed a new GKM clustering algorithm “a faster global K-means clustering algorithm for streaming data sets (FGKM-SD)” which accelerates GKM. Our algorithm requires less computing time and fewer distance calculations. It will also take low memory space.

9. REFERENCES

[1] JuanyingXie and Shuai Jiang. A simple and fast algorithm for global K-means clustering, Second International Workshop on Education Technology and Computer Science, pp 36-40 2010

[2] A. Likas, M. Vlassis, and J. Verbeek, “The global k-means clustering algorithm,” Pattern Recognition, vol. 36, pp. 451–461, 2003.

[3] G. Tzortzis and A. Likas “The Global Kernel k-Means Clustering Algorithm” International Joint Conference on Neural Networks (IJCNN 2008), 2008, pp 1978-1985

[4] J.A. Lozano, J.M. Pena, P. Larranaga, An empirical comparison of four initialization methods for the k-means algorithm, Pattern Recognition Lett. 20 (1999) 1027–1040

[5] M.N. Murty, A.K. Jain, P.J. Flynn, Data clustering: a review, ACM Comput. Surv. 31 (3) (1999) 264–323

[6] Na, S. and L. Xumin, 2010. "Research on K-means Clustering Algorithm An Improved K-means Clustering Algorithm," in Third International Symposium on Intelligent Information Technology and Security Informatics (IITSI), Jinggangshan

[7] Wang, J. and X. Su, 2011. "An improved K-means clustering algorithm," in 3rd International Conference on Communication Software and Networks (ICCSN), Xi'an.

[8] P. S. Bradley and U. M. Fayyad, “Refining initial points for k-means clustering,” Proceedings of the Fifteenth International Conference on Machine Learning, Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, 1998, pp. 91–99.

[9] Bagirov, Adil M., and KarimMardaneh. "Modified global k-means algorithm for clustering in gene expression data sets." In *Proceedings of the 2006 workshop on Intelligent systems for bioinformatics-Volume 73*, pp. 23-28. Australian Computer Society, Inc., 2006.

[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 (Slovenia)* 32, no. 2 (2008): 219-225.

[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] Kumar, Parvesh, and SiriKrishanWasan. "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.

[13] 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.

[14] BAGIROV, Adil M., Julien UGON, and Dean WEBB. "Fast modified global k-means algorithm for incremental cluster construction." *Pattern recognition* 44, no. 4 (2011): 866-876.

[15] Lai, Jim ZC, and Tsung-Jen Huang. "Fast global k-means clustering using cluster membership and inequality." *Pattern Recognition* 43, no. 5 (2010): 1954-1963.

[16] 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.

[17] Bai, Liang, Jiye Liang, Chao Sui, and Chuangyin Dang. "Fast global k-means clustering based on local geometrical information." *Information Sciences* (2013).