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 9, September 2016

A Survey on K-Means Clustering Algorithms for Large Datasets

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: In this article we present a detailed survey on K-Means clustering algorithms for huge datasets like high dimensional dataset etc. Our study gives an overview of different clustering scheme of data mining. The k-means algorithm and its variations are very well recognized to be speedy clustering algorithms. Nevertheless, they are susceptible to the choice of initial points and are unproductive for solving clustering troubles in very large datasets. Currently, incremental schemes have been developed to determine difficulties with the choice of initial points. The global k-means (GKM) and the fast global k-means (FGKM) algorithms are based on such a scheme. They iteratively append one cluster center at a time. Numerical experiments show that these algorithms considerably improve the k-means algorithm. Nevertheless, they require buzzing the whole affinity matrix or computing affinity matrix on all steps of algorithms. This creates both algorithms time consuming and memory demanding for clustering even moderately large datasets. We give comparative study of different k means algorithms to understand them effectively.

Keywords: High Dimensional dataset, clustering, K means, GKM.

1. INTRODUCTION

Data Mining is a field of computer science and it is the collection of techniques for efficient automated discovery of hidden patterns or previously unknown, valid, novel, valuable, and understandable patterns from huge collections of history data, where the data could be stored in databases, data warehouses, or other information repositories.

Clustering is a technique of data mining. It is an unsupervised learning technique, which does not rely on predefined model and output classes. The input objects are just classified based on some observed criteria, which may not be predefined. The process of clustering is collecting a set of entities within a set of disjoint groups, known as clusters so that the objects in the similar cluster have elevated relationship, but are extremely different with objects in further clusters. This system is utilized in lots of domains, for instance human science (environmental science, zoology etc.), health sciences (psychoanalysis, pathology etc.), public sciences (sociology, archaeology etc.), soil sciences (geography, geology etc.), and engineering [1]. Figure 1 explains the system of clustering with four necessary phases:

- Feature selection: This phase selects unique characteristics from a set of contenders while feature extraction employs a few changes to produce helpful and new characteristics from the unique ones. Both are extremely critical to the success of clustering functions. Graceful choice of characteristics could deeply shrink the workload and simplify the succeeding design process.
- Clustering algorithm selection: This phase is typically merged with the choice of an equivalent proximity determination and the building of a principle function. Patterns are clustered according to whether they are similar to each other. Clearly, the proximity determination straight influences the arrangement of the resulting clusters. Approximately all clustering algorithms are explicitly or implicitly connected to a few definition of proximity determination. Several algorithms yet work straight on the proximity matrix. Once a proximity determination is selected, the building of a clustering principle function creates the partition of clusters an optimization trouble, which is well described mathematically, and has rich solutions in the literature [2].
- Cluster validation. Cluster validation refers to procedures that evaluate the results of cluster analysis in a quantitative and objective fashion. A clustering structure is "valid" if it is "unusual" in some sense. Given a data set, each clustering algorithm could always generate a division, no matter whether the structure exists or not. Moreover, different approaches usually lead to different clusters; and even for the same algorithm, parameter identification or the presentation order of input patterns may affect the final results. Therefore, effective evaluation standards and criteria are important to provide the clients with a degree of confidence for the clustering results derived from the utilized algorithms. These assessments should be objective and have no preferences to any algorithm. Also, they should be valuable for answering questions

DOI 10.17148/IJARCCE.2016.5912

like how many clusters are hidden in the data, whether the clusters obtained are meaningful or just an artifact of the algorithms, or why we choose some algorithm instead of another.

- Results interpretation. The ultimate goal of clustering is to provide clients with meaningful insights from the original data, so that they could effectively solve the troubles encountered. Experts in the relevant fields interpret the data partition. Further analyzes, even experiments, may be required to guarantee the reliability of extracted knowledge. Cluster analysis is not a one-shot process. In many circumstances, it needs a series of trials and repetitions. Moreover, there are no universal and effective criteria to guide the selection of features and clustering schemes. Validation criteria provide some insights on the quality of clustering solutions. But even how to choose the appropriate criterion is still a trouble requiring more efforts.

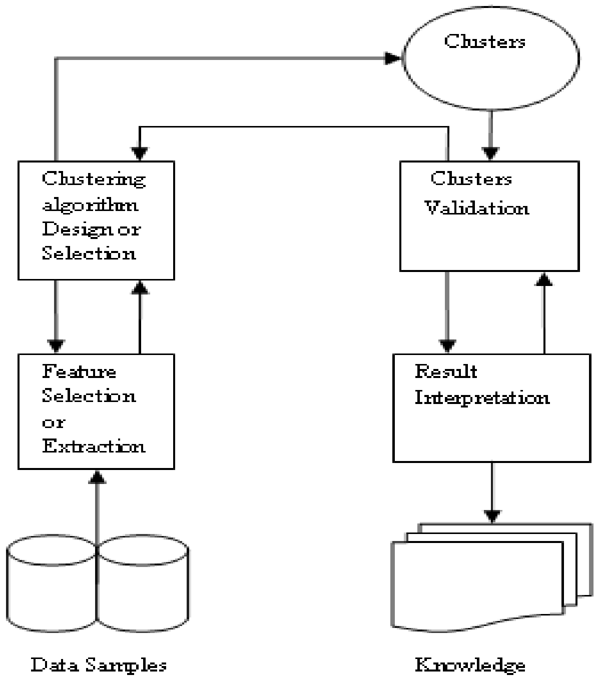
Clusters
Clustering algorithm Design or Selection
Clusters Validation
Feature Selection or Extraction
Result Interpretation
Data Samples
Knowledge

Fig. 1: Process of Clustering

2. CLUSTERING TECHNIQUES IN DATA MINING

Clustering is a dynamic field of research in several fields of data mining such as statistics, pattern recognition, and machine learning data mining. A variety of clustering techniques have currently emerged and successfully applied to real life data mining troubles. Every clustering technique to include three main components:

- The measure utilized to assess similarity or dissimilarity between pairs of objects.
- The particular strategy followed in order to merge intermediate results. This strategy obviously affects the way the final clusters are produced, since we may merge intermediate clusters according to the distance of their closest or furthest points, or the distance of the average of their points.
- An objective function that needs to be minimized or maximized as appropriate, in order to produce final results.

In the ways of cluster identification point of view, we could classify the clustering algorithms as follows:

- Partitioning clustering: These algorithms select some seeds as the representatives of the clusters. Every object in the dataset is assigned to the most similar seed to form clusters. The goodness of the clusters (and thus the seeds) is evaluated by an objective function. Different sets of seeds are tried, and the set that yields the best objective function value is reported. There are two major types of partitioning algorithms: k-means [3] and k-medoids [4]. They differ by the choice of seeds: k-means algorithms utilize centroids as seeds, while k-medoids algorithms select objects from the dataset as seeds.
- Hierarchical clustering: There are two main types, agglomerative and divisive. Agglomerative algorithms treat each data object as a singleton cluster. The algorithms repeatedly identify the two most similar clusters and merge them into a larger cluster until certain stopping criteria are reached. Conversely, divisive algorithms put all objects into a single cluster initially. Each time a cluster is divided into two smaller clusters so that dissimilar objects are separated to different clusters. In general agglomerative algorithms are more popular as divisive algorithms could have exponential time complexity [5]. Regardless of the two important categories of partitioning and hierarchical clustering algorithms, several other algorithms have developed in cluster analysis and are dependent on specific troubles or specific data sets

DOI 10.17148/IJARCCE.2016.5912

available. We shortly specify some of them below, and select only ones that are suitable for clustering of huge amount and large dimensional data.

- Density-based clustering: These algorithms group objects according to specific density objective functions. Density is usually defined as the number of objects in a particular neighborhood of a data object. In these approaches, a given cluster continues growing as long as the number of objects in the neighborhood exceeds some parameter. CAN (Density-Based Spatial Clustering of Applications with noise) [6], OPTICS (Ordering Points to Identify the Clustering Structure) [7] and DENCLUE (DENsity-based CLUstEring) [8].

Grid-based clustering: The Grid-based clustering utilizes a grid based multi-resolution data structure in which firstly quantized the original data (object) space into fixed number of cells which form the grid structure and then execute all the operation on the quantized space. Its main uniqueness is the fastest processing time that typically depends on the size of grid space instead of data objects. These schemes utilize a single uniform grid mesh to partition the entire domain into cells and data objects located within a cell. Each grid cell Consists of some statistical information regarding the attributes (such as mean, min, max) are pre-computed and stored. These statistical parameters are valuable for query processing. The main perspective of these algorithms is to divide the dataset into a fixed number of cells and then work with objects regarding to these cells. They do not relocate objects but construct the several hierarchical levels of rectangular cells of objects. It means, they are nearest to hierarchical clustering algorithms but merging of grids depend on a predefined parameter, which is depend on the number of objects that fall into a particular cell of the multi-resolution grid, not a distance calculation. The grid based clustering algorithms are STING (Statistical Information grid) [9], Wave Cluster [10] and CLIQUE (Clustering In Quest)11].

- Model-based clustering: A model is hypothesized for each of the clusters and the idea is to find the best fit of that model to each other. These algorithms choose a model for the data and then find values of model parameters that best fit the data. These may be either hierarchical or partitioning algorithm, based on the model which are hypothesize about the data set and these are refine this model to specify partitioning. These are nearest to density-based clustering algorithms, in that they grow particular clusters so that the preconceived model is improved. Nevertheless, they sometimes start with a fixed number of clusters and they do not all utilize the same concept of density. The model-based clustering algorithms are Self Organizing Map (SOM) Neural Network and Expectation-Maximization algorithms [12].

Clustering is a dynamic field of data mining in which some special requirements of clustering are necessary such as [3]:

- Scalability: It concerns that algorithms should perform well with small or huge amount of data objects. Nevertheless, a huge database could contain millions of data objects.
- High dimensionality of data: Most of all clustering algorithms are excellent at handling low-dimensional data consists of two or three dimensions. Nevertheless a dataset may contain several dimensional or attributes. It is crucial to cluster datasets in high dimensional space, since data may be very sparse and highly skew.
- Interpretability and usability: Most of the time, it is expected that clustering algorithms produce usable and interpretable results. But when it comes to comparing the results with preconceived ideas or constraints, some techniques fail to be satisfactory. Therefore, easy to understand results are highly desirable.
- Minimal requirements for input parameters: Many clustering algorithms require some client-defined parameters, such as the number of clusters, in order to analyze the data. Nevertheless, with large data sets and higher dimensionalities, it is desirable that a scheme require only limited guidance from the client, in order to avoid biasing the result.
- Insensitivity to the order of input orders: The same data set, when presented to certain algorithms in different orders, many lead to dramatically different clustering. The order of input mostly affects algorithms that perform only single scan over the data set, leading locally optimal solutions at every step. Thus, is crucial that algorithms should insensitive to order of input.
- Handling of noise: Clustering algorithms should be able to handle deviations, in order to improve cluster quality. Deviations are defined as data objects that depart from generally accepted norms of behavior and are also referred to as outliers. Deviation detection is considered as a separate trouble.
- Analyze mixture of attribute types: The ability to analyze dataset with mixtures of attribute types, as well as homogeneous ones.
- Find arbitrary-shaped clusters: Different types of algorithms will be biased toward finding different types of cluster structures/shapes and it is not always an easy task to determine the shape or the corresponding bias.

3. THE K-MEANS ALGORITHM

K-means 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. Following are algorithmic steps K-mean algorithm [13].

Algorithm 1:

INPUT: Number of desired clusters K, Data objects $D = \{d_1, d_{2,} \dots d_n\}$.

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 di (1 <= i<=n) from all k clusters C j (1 <= j<=k) and then assign data object di to the nearest cluster.
- For each cluster j (1 <= j<=k)
- Recalculate the cluster center until no change in the center of clusters.

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 client 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 inlarge amount of data.
- On using a large amount of data set the storage space increases in K-mean algorithm

4. THE GLOBAL K - MEANS ALGORITHM

The Global K-Means (GKM) algorithm minimized clustering error using deterministic valuable global clustering algorithm. The k-means algorithm as a local search procedure is the algorithm follows incremental approach to solve M cluster clustering trouble, all intermediate troubles with 1, 2… M −1 clusters are sequentially solved [14].

The GKM scheme provides optimal solution for clustering trouble with M clusters through a series of local searches (using K-means algorithm). For each local search, M-1 cluster centres must be initially placed in their optimal positions corresponding to the clustering trouble with M-1 clusters. Then placed the remaining Mth cluster center at several positions within the data space. If for M=1 the optimal solution is known, we could iteratively apply the above procedure to get the 2nd optimal solutions for all k-clustering troubles K=1, 2…, M. This effective scheme 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, kth iteration of this algorithm utilizes k-1 clusters centres from the previous iteration. We could 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 x1 of the set A:

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

And set k = 1.

Step 2: Set k = k + 1 and consider the centres $x^1, x^2, \ldots, x^{k-1}$ from the previous iteration.

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

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.

5. FAST GLOBAL K-MEANS CLUSTERING ALGORITHM [15]

Global k-Means clustering algorithm requires m executions of the k-Means algorithm for each data point xi in X. Therefore computational time complexity of Global k-means algorithm is rather higher. The fast global k-Means algorithm accelerates the global k- Means algorithm [14]. Given the solution of the (k−1)- clustering trouble {v1 (k-1), …, vk-1(k-1)} and corresponding value of the objective function J(k-1) representing the sum of the squared error (SSE) as described in [14], the algorithm does not execute the k-Means algorithm for each data point repeatedly to find the solution of the k –clustering trouble. Instead, the algorithm computes the upper bound J (k) ≤ J (k -1) - bi on the resulting error J(k) for each possible data point xi ε X , where J(k-1) is the error value of (k-1)-clustering trouble and bi is defined as:

$$b_i = \sum_{j=1}^{m} \max\left[d^j_{k-1} - \left\|x_i - x_j\right\|^2, 0\right], \forall\, i = 1 \ldots m$$

Here,

$$d^j_{k-1} = \min\left[\left\|x_j - v_1^{k-1}\right\|^2, \ldots, \left\|x_j - v_{k-1}^{k-1}\right\|^2\right]$$

d^J_{k-1} is the closest distance between the squared distance between x_j and the closet centre among (k-1)-cluster centres{v1 (k-1), …, vk-1(k-1)} , that is, the squared distance between xj and the centre of a cluster it belongs to m is the size of the data set. A data point xi ε X with the maximum value of bi is chosen as an initial centre for the kth cluster centre. The quantity bi measures the guaranteed reduction in the error measure obtained by inserting a new cluster centre at position xi.

6. RELATED WORK

Researchers are continuously working on Fast Global K-Means (FGKM) 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 was developed for clustering in gene expression data sets which is effective for solving clustering troubles in gene expression data sets [16]. This algorithm computes clusters incrementally and to compute k-partition of a data set it utilizes $k - 1$ cluster centres from the previous iteration. Computation of the starting point for the kth cluster centre is the key point. Starting point is calculated by minimizing so-called auxiliary cluster functions.

For Autonomous Cluster Initialization of Probabilistic Neural Network an approach was demonstrated [17]. In this approach statistical based Probabilistic Neural Network (PNN) was utilized for pattern classification troubles with Expectation – Maximization (EM) chosen as the training algorithm. Global K-means algorithm solves the trouble of random initialization. Initially, client needs to predefine the number of clusters using trial and error scheme. 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 troubles. Currently new version of the k-mean algorithm has been developed, which is known as global k-means algorithm [18]. Global k-means is the incremental algorithm that allows us to add one cluster centre at a time and utilizes each data point as a candidate for the k-th cluster centre. 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 utilizes minimizing an auxiliary cluster function to compute the starting point for the kth cluster centre. Numerical results of these experiments (i.e. 14 data sets) demonstrate the superiority of the new algorithm, nevertheless it required more computational time than global k-mean algorithm.

Based on colon dataset, global k-means and x-means algorithms were analysed [19]. 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 scheme of how the next new cluster centres is created by using some idea of K-mediods clustering algorithm suggested by Park and Jun [20]. 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 centre could avoid the impact of noisy data on the clustering result.

A New version of GKM algorithm utilizes auxiliary cluster function to compute the starting point for the kth cluster centre by minimizing it [21]. The difference between the FGKM algorithm and new version is in the way of starting point for the kth cluster centre is obtained. A local minimizer utilized as starting point for the kth cluster centre. This algorithm not only utilized to compute the cluster incrementally but also to compute k-partition of a data set, it utilizes $k-1$ cluster centres 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 centre at a time. Numerical experiments show that these algorithms considerably improve the k-means algorithm.

A Fast global k-means clustering algorithm is introduced by making utilize of the cluster membership and geometrical information about a data point [22]. This algorithm is referred to as MFGKM. The algorithm utilizes a set of inequalities developed to determine a starting point for the jth cluster centre of global k-means clustering. Adopting multiple cluster centres 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 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 is developed in this paper [23]. In this development, firstly apply local geometrical information to describe approximately the set of objects represented by a candidate cluster centre. 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 neighbourhoods are induced by multi probe functions with many category features [24]. In this paper, by combining global k-means clustering algorithms and topology neighbourhoods, two k-means clustering algorithms are proposed, in which AFS (Axiomatic Fuzzy Sets) topology neighbourhoods 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 neighbourhood is employed in the step of determining initial cluster centres.
The algorithms are independent of the initial conditions, which allow working with many category features.

DOI 10.17148/IJARCCE.2016.5912

7. COMPARISION OF K-MEANS CLUSTERING SCHEMES

Scheme	Advantage	Disadvantage
K-Means	K-means scheme is relatively efficient for small data sets, which is classify the data objects in k clusters depends on selection of initial k means.	K-means are sensitive to the choice of starting points and inefficient for solving outliers and large data sets troubles.
Global K-means	GKM is incremental approach to clustering that dynamically adds one cluster center at a time through local search procedure. It is not depends on initial conditions for cluster center.	GKM is a local search procedure which takes large computational time and storage space.
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 could effectively determine a global or near global minimizer of the cost function.	FGKM algorithm needs a large amount of storage space and computational time when handling larger data sets.

8. CONCLUSION

Clustering is a technique of data mining. It is an unsupervised learning technique, which does not rely on predefined model and output classes. Cluster analysis is not a one-shot process. In many circumstances, it needs a series of trials and repetitions. Moreover, there are no universal and effective criteria to guide the selection of features and clustering schemes. In this article we have offered detailed survey on different K-means, Global K-means and Fast Global K means clustering algorithms for large dataset. GKM clustering algorithm is a deterministic clustering scheme, autonomous of any initial circumstances and grants outstanding outcomes in terms of the sum of the squared error standard. This scheme executes greatly better than the k- Means algorithm with numerous random restarts. The Fast Global k-Means algorithm extensively diminish the requisite computational attempt, while at the similar time providing solutions of approximately the equal clustering error superiority. Thus Fast Global k-Means algorithm is a lot more scalable than other K Means algorithm.

REFERENCES

1. S. Theodoridis and K. Koutroumbas. Pattern Recognition. Elsevier Science, 2003.
2. Kleinberg, Jon. "An impossibility theorem for clustering." Advances in neural information processing systems (2003): 463-470
3. J. Han and M. Kamber. Data Mining: Concepts and Techniques, Morgan Kaufmann, 2001.
4. L. Kaufman and P. Rousseuw. Finding groups in data - An introduction to cluster analysis. Wiley series in probability and mathematical statistics, 1990.
5. R. Ng and J. Han. Efficient and effective clustering methods for spatial data mining. pages 144–155, 09 1994.
6. Ester, Martin, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. "A density-based algorithm for discovering clusters in large spatial databases with noise." In Kdd, vol. 96, no. 34, pp. 226-231. 1996.
7. Ankerst, Mihael, Markus M. Breunig, Hans-Peter Kriegel, and Jörg Sander. "OPTICS: ordering points to identify the clustering structure." In ACM Sigmod Record, vol. 28, no. 2, pp. 49-60. ACM, 1999.
8. Hinneburg, Alexander, and Daniel A. Keim. "An efficient approach to clustering in large multimedia databases with noise." In KDD, vol. 98, pp. 58-65. 1998.
9. STING, Wang W ˈYang Jˈ Muntz R."A Statistical Information Grid Approach to Spatial Data Mining." In Athens, Proceedings of the 23rd Conference on VLDBˈ1997, pp. 186-195.
10. Sheikholeslami, Gholamhosein, Surojit Chatterjee, and Aidong Zhang. "Wavecluster: A multi-resolution clustering approach for very large spatial databases." In VLDB, vol. 98, pp. 428-439. 1998.
11. Agrawal, R., Gehrke, J., Gunopulos, D., & Raghavan, P. (1998). Automatic subspace clustering of high dimensional data for data mining applications (Vol. 27, No. 2, pp. 94-105). ACM.
12. Bradley, Paul S., Usama Fayyad, and Cory Reina. Scaling EM (expectation-maximization) clustering to large databases. Redmond: Technical Report MSR-TR-98-35, Microsoft Research, 1998.
13. Na, Shi, Liu Xumin, and Guan Yong. "Research on k-means clustering algorithm: An improved k-means clustering algorithm." In Intelligent Information Technology and Security Informatics (IITSI), 2010 Third International Symposium on, pp. 63-67. IEEE, 2010.
14. A. Likas, M. Vlassis, and J. Verbeek, "The global k-means clustering algorithm," Pattern Recognition, vol. 36, pp. 451–461, 2003.
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. 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.
17. 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.
18. Bagirov, Adil M. "Modified global k-means algorithm for minimum sum-of-squares clustering problems." Pattern Recognition 41, no. 10 (2008): 3192-3199.
19. 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.
20. 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.
21. 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.
22. 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.
23. Bai, Liang, Jiye Liang, Chao Sui, and Chuangyin Dang. "Fast global k-means clustering based on local geometrical information." Information Sciences (2013).
24. 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.2016.5912