International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

An Optimum Method for Enhancing the Computational Complexity of K-Means Clustering Algorithm with Improved Initial Centers

A. Mallikarjuna Reddy[1], Ramapuram Gautham[2]

[1]Assistant Professor, Department of CSE, Anurag Group of Institutions, India

[2]M.Tech Student, Department of CSE, Anurag Group of Institutions, India

Abstract: ***The amount of data stored in databases continues to grow fast. Intuitively, this large amount of stored data contains valuable hidden patterns, which could be used to improve the decision-making process. Data mining is a process of identifying specific patterns from large amount of data. In data mining, Clustering is one of the major data analysis methods and the K-means clustering algorithm is widely used for many practical applications. Though it is widely used, its generates a local optimal solution based on the randomly chosen initial centroids and the computational complexity is very high O(nkl). In order to improve the performance of the K-means algorithm several methods have been proposed in the literature. The proposed algorithm enhances the performance of K-means clustering algorithm. This algorithm consists of two phases. Phase I algorithm finds the better initial centroids, Phase II algorithm is used for the effective way of assigning data points to suitable clusters. Experiments on a number of real-world data sets show that the proposed approach has produces consistent clusters compared to some well-known methods, reducing the computational complexity O(nlogn) of k-means algorithm.Though the proposed method will improve the accuracy and efficiency of k-means clustering algorithm.***

Keywords: Data Mining, Clustering, Knowledge Discovery in Databases, K-means clustering algorithm, Optimum method.

1. Introduction

Data Mining and Data-warehousing are two branches in the process of Knowledge Discovery in Databases (KDD).Data mining is a process of extracting knowledge from huge databases. Data mining uses sophisticated mathematical algorithms to segment the data and evaluate the probability of future events. Cluster analysis is the one of the major task in data mining. Clustering is unsupervised classification. Clustering is the process of grouping the data into classes or cluster so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters. The objects are clustered or grouped based on the principle of maximizing the intra-class similarity and minimizing the interclass similarity. It is a main task of exploratory data mining, and a common technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis and bioinformatics.

Mainly clustering algorithms are categorized into two types: Hierarchical and partitional algorithms. Hierarchical algorithms build a tree structure from data. Hierarchical algorithms are again categorized into agglomerative and divisive algorithms. In agglomerative algorithms, each data point is considered as a cluster initially and merges the most related cluster pairs in each step [8]. Divisive algorithms start with one cluster and iteratively divide the clusters into sub clusters. Partitioning a large dataset of objects into homogeneous clusters is a basic fundamental operation in data mining. Partitional algorithms partition the data in a single level and hence the clusters do not have a hierarchical structure.

1.1. Literature Survey

Several methods have been proposed to solve the clustering problem. In 1967, Mac Queen developed the simple and easy clustering algorithm is k-means clustering algorithm [6]. It is applied in a various fields like bioinformatics and pattern recognition. The k-means algorithm is the one of the partitional clustering method. It separates the data into k clusters [2] [4] [10]. Efficiency of the k-means algorithm heavily depends upon the initial cluster centers [8]. Various methods have been proposed in the literature to improve the accuracy and efficiency of the k-means algorithm.

K-means algorithm is very sensitive in initial starting points. K-means generates initial cluster centroids randomly. When random initial starting points close to the final solution, K-means has high possibility to find out the cluster centers. Otherwise, it will lead to incorrect clustering results [7].

K. A. Abdul Nazeer and et al. [8] proposed an enhanced method to improve the accuracy and efficiency of the K-means clustering algorithm. In this algorithm the authors proposed two methods, one method for finding the better initial centroids. And another method for an efficient way for assigning data points to appropriate clusters with reduced time complexity. Though this algorithm produced clusters with better accuracy and efficiency compared to k-means, it takes $O(n^2)$ time for finding the initial centroids.

Koheri Arai et al. [7] proposed an algorithm for centroids initialization for K-means. In this algorithm both K-means and hierarchical algorithms are included. First, in this algorithm K-means is applied many times and each maintains centroids in the data set C. Next, the data set C is giving as an input to the hierarchical clustering algorithm.

The hierarchical clustering algorithm runs until it gets the desired number of clusters. After that, calculate mean of each cluster, these means will be the initial centroids. This algorithm gives the better initial centroids. But in this algorithm K-means is applied many times, so it is computationally expensive in the presence of large data sets.

Fahim A.M et al. [2] proposed an enhanced method for assigning data points to clusters. The original K-means algorithm is computationally expensive because each iteration computes the distances between data points and all centroids. Fahim approach makes use of an effective method to reduce the complexity. But this method presumes that the initial centroids are determined randomly, as in the case of the original K-means algorithm. Hence there is no guarantee for the quality of the final clusters which depends solely on the selection of initial centroids.

Fang Yuan et al. [5] proposed a systematic method for finding the initial centroids. The centroids obtained by this method will be consistent with the distribution of data and hence produced better clustering. However, Yuans method does not suggest any improvement to the time complexity of the K-means algorithm.

Bhattacharya et al. [1] proposed a novel clustering algorithm, called Divisive Correlation Clustering Algorithm (DCCA) for grouping of genes. DCCA is able to produce clusters, without taking the initial centroids and the value of k, the number of desired clusters as an input. The time complexity of the algorithm is high and the cost for repairing from any misplacement is also high.

Zhang chen et al. [3] proposed the initial centroids algorithm that avoids the random selection of initial centroids in k-means algorithm. This proposed method for finding the better initial centroids and provide an efficient way of assigning the data points to suitable clusters with reduced time complexity.

2. The standard K-means clustering algorithm

K-means is a numerical, unsupervised, non-deterministic, iterative method. It is simple and very fast, so in many practical applications it is proved to be very effective in producing the good clustering results. The K-means clustering algorithm consists of two phases: the first phase is to define k centroids, one for each cluster. The next phase is to take each point belonging to the given data set and associate it to the nearest centroid. When all the points are integrated in some clusters, the first phase is completed and an early grouping is done. At this point we need to recalculate the new centroids, as the inclusion of new points may lead to a change in the cluster centroids [6]. Once find k new centroids, a new binding is to be created between the same data points and the nearest new centroid, generating a loop. As a result of this loop, the k centroids may change their position in a step by step manner. Eventually, a situation will be reached where the centroids do not move anymore. This signals the convergence of clustering. Pseudo code for the standard K-means clustering algorithm is listed as Algorithm I [4]. The Euclidean is used to calculate the distance between two dimensional data points $X=(x_1,x_2,\ldots\ldots x_n)$ and $Y=(y_1,y_2,\ldots\ldots\ldots y_n)$ formula is described as follows;

$$d(X,Y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}$$

Algorithm I: The K-means clustering algorithm

Require: $D = \{d_1, d_2, \ldots\ldots, d_n\}$ //set of n data items.
k // Number of desired clusters
Ensure: A set of k clusters.
Steps:
1. Arbitrarily choose k data-points from dataset D as initial cluster centroids;
2. **Repeat**
2.1 Calculate the distance between each data point d_i and all k cluster centers c_j and assign each data item d_i to the cluster which has the closest centroid.
2.2 Calculate the new mean of each cluster j.
Until convergence criterion is met.

Shortcomings of K-means algorithm

greatly depends on the randomly selection of the initial centroids. In the original k-means algorithm, the initial centroids are chosen randomly and hence we get different clusters for different runs for the same input data [12]. Moreover, the k-means algorithm is computationally very high. The computational time complexity of the k-means algorithm is $O(nkl)$, where n is the total number of data points in the dataset, k is the required number of clusters and l is the number of iterations [7]. So, the computational complexity of the k-means algorithm is depends upon the number of data elements, number of clusters and number of iterations.

3. Proposed algorithm

3.1 An optimum method for enhances the computational complexity K-means algorithm

The proposed algorithm enhances the performance of K-means clustering algorithm. The paper [2] authors proposed an enhanced method to improve the efficiency of the k-means clustering algorithm. But in this method the initial centroids are selected randomly. So this method is very sensitive to the initial starting points and it does not produce the unique clustering results. This algorithm consists of two phases. Phase I algorithm explains for finding the initial centroids, Phase II algorithm explains the effective way of assigning data points to suitable clusters. Pseudo code for the proposed algorithm organized is as follows.

Algorithm II: An optimum method for enhances the computational complexity K-means algorithm

Require: $D = \{d_1, d_2, \ldots d_n\}$ // Set of n data points.
k // Number of desired clusters
Ensure: A set of k clusters.
Steps:
1: Phase I: Finding the initial cluster centroids by using Algorithm 3.
2: Phase II: Efficient way of assigning each data points to its suitable cluster based on initial centroids by using Algorithm 4.

3.2 Phase I Algorithm

This method follows a new approach for finding the better initial centroids with reduced time complexity and improving the accuracy. The proposed algorithm is outlined as Algorithm III. More often, we may have to deal with multidimensional data values. Each data point *di* may contain multiple attributes such as d_{i1}, d_{i2}, ... d_{in}, where *m* is the number of attributes or columns in each data value. In this algorithm, first check whether the data set contains any negative value attribute. If the dataset contains the negative value attributes then transform the all data points in the data set to the positive space by subtracting the each data point attribute with the minimum attribute value in the given data set.

At this point, the transformation is required, because in the proposed algorithm, calculate the distance from origin to each data point in the data set. So, for the different data points as showed in *Fig.1*, it will get the same Euclidean distance from the origin. This will result in incorrect selection of the initial centroids. To overcome this problem all the data points are transformed to the positive space. Then for all the data points as showed in *Fig.1*, it will get the unique distances from origin. Then store all the distance into D_o set. if data set contains the all positive value attributes then the transformation is not required. In the next step, for each data point, calculate the distance from origin. Then, the original data points are sorted accordance with the sorted distances. After sorting and partition the sorted data points into *k* equal sets. In each set take the middle points as the initial centroids. Record all the centroids into centroid set as *C*. These initial centroids lead to the better unique clustering results.

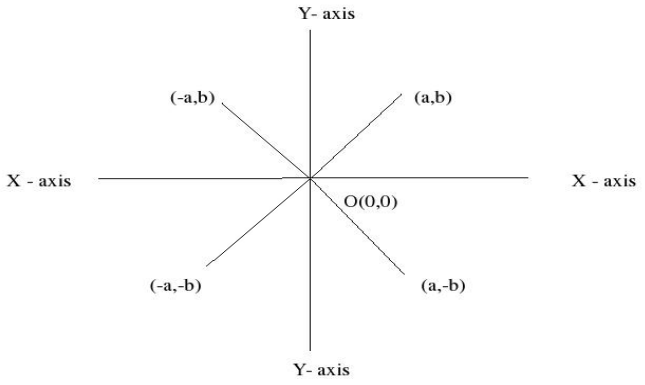

Figure 1: Data points in Two Dimensional Space

Algorithm III: Finding the Initial centroids algorithm

Require: $D = \{d_1, d_2, d_3 \ldots d_n\}$ // Set of n data points.
k // Number of desired clusters.
Ensure: Initial centroids of *C* clusters.
Steps
1:In the given data set *D*, if the data points contains the both positive and negative attribute values then go to step 2, otherwise go to step 4.
2:Find the minimum attribute value in the given data set *D*.
3:For each data point attribute, subtract with the minimum attribute value.
4:For each data point calculate the distance from origin.
5:Then record all distances as $D_O = \{D_{Oi} / i = 1, 2, 3 \ldots \ldots, k\}$.
6:Sort the *Do* in non decreasing order obtained in step 4. Sort the data point's accordance with the distances.
7:Partition the sorted data points into *k* equal sets.
8:In each set, take the middle point as the initial centroid.
9:Record the initial centroids as $C = \{C_i / i = 1, 2, \ldots k\}$.

3.3 Phase II algorithm

After finding the initial centroids from algorithm III, the data points are assigned to different clusters by using the phase II algorithm. Each data point d_i is assigned to the cluster having the closest centroid. Euclidean distance is used as the measure for determining the distance between the data points and the centroids.

Algorithm IV: Efficient way of Assigning data points to suitable clusters based on initial centroids

Require: $D = \{d_1, d_2, d_3 \ldots d_n\}$ // Set of n data points.
$C = \{c_1, c_2, \ldots c_k\}$ // set of k centroids.
Ensure: A set of k clusters.
Steps:
1. Choose initial centroids from *C*;
2. Repeat
2.1 Assign each data item d_i $(1 <= i <= n)$ to the cluster which has the closest centroid C_j $(1 <= j <= n)$;
2.2 Calculate the new mean of each clusters *j*.
Until convergence criterion is met.

In the original K-means algorithm in which the initial centroids are chosen randomly, the heuristic algorithm finding the initial centroids in a more meaningful way, in accordance with the distribution of data. Consequently, the algorithm converges much faster than the original K-means algorithm. Moreover, since the method for determining the initial centroids is based on the technique of sorting, this phase requires less time compared to other similar approaches available in the literature [8,5].

4. Time complexity

The time complexity of proposed algorithm for finding the initial centroids is *O(nlogn)* in both average and worst case, where *'n'* is the number of data points. In the proposed algorithm, the second step finding the minimum value in the given data set requires *O(n)* time where *n* is the number of data items. The time required to subtract with the minimum value for each data point requires *O(m)*.The next step for sorting the distances takes *O(nlogn)* time using heap sort.

Next step for partitioning the n data points into k equal sets takes *O(n)*. Thus the overall time complexity of phase I is *O(nlogn)* . The second phase algorithm consisting of the assignment of data-points to the nearest clusters and the subsequent recalculation of centroids is executed repetitively until the convergence criterion is reached.

This procedure takes time *O(nkl)* where *'n'* is the number of data-points, *'k'* is the number of clusters and *'l'* is the number of iterations. Nevertheless, the algorithm converges in much less number of iterations as the initial centroids are computed in a strategic manner in tune with the data distribution. Thus the overall time complexity of the proposed algorithm is *O(nkl+nlogn)*. Here the convergence will met in less number of iterations. Hence, *'k'* can be neglected. Therefore the overall time complexity of the proposed algorithm becomes *O(nlogn)*. So the proposed algorthim has less time complexity compared to the original K-means clustering algorithm

5. Experimental Results

The original K-means, and the Enhanced K-means algorithms require the values of the initial centroids also as input, apart from the input data values and the value of *k*. The proposed algorithm is applied to Multi-dimensional data taken from the UCI (university of California Irvine) repository [13] .
The input data are the iris data [11], the e-coli data [9], and the new thyroid data [12] obtained from the web site of disabled-world [13]. For testing the accuracy and efficiency of the proposed algorithm, a set of data values with known clustering is used.

The same set of data inputs are used for the standard K-means algorithm, Enhanced K-means as well as proposed algorithm. Other inputs required for all the algorithms are the value of k (number of clusters). The standard K-means algorithm [8], Enhanced K-means and proposed algorithm is executed noted accuracy and time taken for each run. The results are compared with that of the standard K-means algorithm [8], Enhanced K-means as well as proposed method. The percentage of accuracy and the time taken for each experiment are computed and tabulated as follows in Table1.

Table 1: Performance comparison of the algorithms for different data sets.

Data set	Number of clusters	Algorithm	Run	Accuracy (%)	Time Taken (msec)
E-coil	*K=3*	Original K-means	10	77.14	0.119
		Enhanced K-means	1	81.5	0.13
		Proposed Algorithm	1	90.27	0.115
New Thyroid	*K=3*	Original K-means	10	73.15	0.127
		Enhanced K-means algorithm	1	82.3	0.12
		Proposed Algorithm	1	84.58	0.115
Iris	*K=3*	Original K-means	10	68.93	0.119
		Enhanced K-means algorithm	1	86.58	0.11
		Proposed Algorithm	1	88.33	0.107

6. Conclusion

The original k-means and the enhanced k-means algorithm requires the values of the initial centroids also as input, apart from the input data values and the value k. For the proposed Heuristic K-means algorithm, the data values and the value of *k* are the only inputs required. The proposed algorithm enhances the computational complexity of the k-means algorithm *O(nlogn)* with improved initial centers. Experimental results shown that the proposed method produces consistent clusters in less time compared to the original k-means algorithm.

7. Future Work

The value of k (the number of clusters) is still required to be given as an input. It can be determined automatically based on the distribution of data,by using some statistical approaches. Automatic determination of the value of k is suggested as a future work.

References

[1] Bhattacharya and R. K. De, "Divisive Correlation Clustering Algorithm (DCCA) for grouping of genes: detecting varying patterns in expression profiles, *bioinformatics*, Vol. 24, pp. 13591366, 2008
[2] M. Fahim, A. M. Salem, F. A. Torkey and M. A.Ramadan, "An Efficient enhanced K-means clustering algorithm," journal of Zhejiang University, 10(7): 162 61633, 2006.
[3] Chen Zhang and Shixiong Xia, " Kmeans Clustering Algorithm with Improved Initial center," *in Second International Workshop on Knowledge Discovery and Data Mining* (WKDD), pp. 790792, 2009.
[4] Margaret H. Dunham ,Data Mining Introductory and Advanced Concepts. Person Education , 2006.
[5] F. Yuan, Z. H. Meng, H. X. Zhangz, C. R. Dong, " A New Algorithm to Get the Initial Centroids," Proceedings of the 3rd International Conference on Machine Learning and Cybernetics, pp. 2629, August 2004.
[6] Mc Queen J, "Some methods for classification and analysis of multivariateObservations," *Proc. 5th Berkeley Symp. Math.Statist.Prob.*,(1):281–297, 1967.
[7] Koheri Arai and Ali Ridho Barak bah, "Hierarchical k-means: an algorithm for Centroids initialization for k-means," department of information science and Electrical Engineering Politechnique in Surabaya, Faculty of Science and Engineering, Saga University, Vol. 36, No.1, 2007.
[8] K.A.Abdul Nazeer and M. P. Sebastian, "Improving the accuracy and efficiency of the K –means clustering algorithm", in International Conference on Data Mining and Knowledge Engineering (ICDMKE), Proceedings of the World Congress on Engineering (WCE2009), Vol 1, July 2009, London, UK.
[9] Ecoli Data available at,http://archieve.ics.uci.edu/Ml/machine-learningdatabases/ecoli/ecoli.data

[10] Elmsari , navathe, Somayajula, Gupta Fundamentals of Database Systems, Pearson Education, First edition, 2006.
[11] Irisdata available at, www.iris.washington.edu/data.
[12] Thyroid data available at, http://archive.ics.uci.edu/ml/machine-learning-databases/thyroid-disease/new- thyroid.data.
[13] The UCI Repository website. [Online]. Available: http://archive.ics.uci.edu/.

Author Profile

A. Mallikarjuna Reddy received the B.E. degree in computer science and Engineering from Anna University in 2007 and M.Tech. degree in Computer science and Engineering from NIT Calicut University in 2010, respectively. He now currently is working as assistant professor in CSE department at Anurag Group of Institutions.

Ramapuram Gautham received the B.E. degree in information technology from JNTU Anantapur in 2012 and pursuing M.Tech. degree in Computer science and Engineering from JNTU Hyderabad.