International Journal of Emerging Engineering Research and Technology
Volume 3, Issue 1, January 2015, PP 22-27
ISSN 2349-4395 (Print) & ISSN 2349-4409 (Online)

Efficient Determination of Clusters in K-Mean Algorithm Using Neighborhood Distance

Kedar B. Sawant

Shree Rayeshwar Institute of Engineering and Information Technology
IT Department, Shiroda-Goa
Kedarsawant22@gmail.com

Abstract: *Clustering is a well known data mining technique which is used to group together data items based on similarity property. Partitional clustering algorithms obtain a single partition of the data instead of a clustering structure. K-mean clustering is a common approach; however one of its drawbacks is the selection of initial centroid points randomly because of which algorithm has to re-iterate number of times. This paper first reviews existing methods for selecting the number of clusters as well as selecting initial centroid points, followed by a proposed method for selecting the initial centroid points and the modified K-mean algorithm which will reduce the number of iterations and improves the elapsed time.*

Keywords: *Clustering, Data mining, K-means, Neighborhood distance, Partitioning algorithm.*

1. INTRODUCTION

Mergence of modern techniques for scientific data collection has resulted in large scale accumulation of data pertaining to diverse fields. Data mining uses sophisticated statistical analysis and modeling techniques to uncover patterns and relationships hidden in organizational databases — patterns that ordinary methods might miss [1]. Cluster analysis is a primary method for database mining. Clustering is a data mining technique that makes meaningful or useful cluster of objects that have similar characteristic using automatic technique [4]. It is used to identify natural groupings of cases based on a set of attributes. Cases within the same group have more or less similar attribute values. Most clustering algorithms build the model through a number of iterations and stop when the model converges, that is, when the boundaries of these segments are stabilized [1, 4]. Cluster analysis could be divided into hierarchical clustering and non-hierarchical clustering techniques. Hierarchical clustering builds a cluster hierarchy or, in other words, a tree of clusters, also known as a dendrogram. Hierarchical clustering can be further categorized into Agglomerative (bottom-up) and Divisive (top-down).The widely used hierarchical clustering algorithm is CURE (Clustering Using Representatives) and BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) [1, 7]. Non-hierarchical techniques include Partitional clustering algorithms which obtain a single partition of the data instead of a clustering structure, such as the dendrogram produced by a hierarchical technique but the main problem accompanying the use of a partitional algorithm is that the number of desired output clusters is predetermined. In practice the algorithm is typically run multiple times with different starting states, and the best configuration obtained from all of the runs is used as the output clustering. Widely used Algorithms in this category are Squared Error Clustering Algorithms, K-means algorithm and K-medoids [3, 10]. The K-means algorithm is a popular data clustering algorithm. To use it requires the number of clusters in the data to be pre-specified. Finding the appropriate number of clusters for a given data set is generally a trial-and-error process made more difficult by the subjective nature of deciding what constitutes ‘correct’ clustering [1]. Apart from this the selection of initial centroid points is done randomly because of which algorithm has to re-iterate number of times. As a result the final clustered output is going to vary every time the algorithm selects different centroid points then the previous time. So the selection of correct centroid points at the beginning of the execution of algorithm is very much important in terms of saving time and to produce best possible clusters. This paper first reviews existing methods for selecting the number of clusters as well as selecting initial centroid points, followed by a proposed method for selecting the initial centroid points and the modified K-mean algorithm which will reduce the number of iterations and improves the elapsed time. The remainder of the paper consists of five sections. Section 2 reviews the main known methods for selecting K and the initial centroid points. Section 3 gives the overview of the K-means clustering algorithm. Section 4 presents the proposed methodology for selecting the initial centroid points and the proposed modified K-means algorithm. Section 5 concludes the paper.

2. LITERATURE SURVEY

The K-means algorithm requires the number of clusters to be specified by the user. To find a satisfactory clustering result, usually, a number of iterations are needed where the user executes the algorithm with different values of K. The validity of the clustering result is assessed only visually without applying any formal performance measures. With this approach, it is difficult for users to evaluate the clustering result for multi-dimensional data sets. The performance of a clustering algorithm may be affected by the chosen value of K. Therefore, instead of using a single predefined K, a set of values might be adopted. It is important for the number of values considered to be reasonably large, to reflect the specific characteristics of the data sets. At the same time, the selected values have to be significantly smaller than the number of objects in the data sets, which is the main motivation for performing data clustering [8].

There are several statistical measures available for selecting K. These measures are often applied in combination with probabilistic clustering approaches. They are calculated with certain assumptions about the underlying distribution of the data. The Bayesian information criterion or Akeike's information criterion [3] is calculated on data sets which are constructed by a set of Gaussian distributions. The measures applied by Hardy are based on the assumption that the data set fits the Poisson distribution. Monte Carlo techniques, which are associated with the null hypothesis, are used for assessing the clustering results and also for determining the number of clusters. Although, EM and K-means clustering share some common ideas, they are based on different hypotheses, models and criteria. Probabilistic clustering methods do not take into account the distortion inside a cluster, so that a cluster created by applying such methods may not correspond to a cluster in partitioning clustering, and vice versa. Therefore, statistical measures used in probabilistic methods are not applicable. In addition, the assumptions about the underlying distribution cannot be verified on real data sets and therefore cannot be used to obtain statistical measures [10].

Another method proposed to find the value of K is by equating it to the number of classes. With this method, the number of clusters is equated to the number of classes in the data sets. A data clustering algorithm can be used as a classifier by applying it to data sets from which the class attribute is omitted and then assessing the clustering results using the omitted class information [1, 7]. The outcome of the assessment is fed back to the clustering algorithm to improve its performance. In this way, the clustering can be considered to be supervised. With this method of determining the number of clusters, the assumption is made that the data clustering method could form clusters, each of which would consist of only objects belonging to one class. Unfortunately, most real problems do not satisfy this assumption.

Values of K determined using a neighbourhood measure- A neighbourhood measure could be added to the cost function of the K-means algorithm to determine K [3]. Although this technique has showed promising results for a few data sets, it needs to prove its potential in practical applications. Because the cost function has to be modified, this technique cannot be applied to the original K-means algorithm.

In [8], A data clustering technique by using K-means algorithm is presented, which is based on the initial mean of the cluster, According to this algorithm, whole data space is divided into segments (k*k) and the frequency of data points in each segment is calculated. The segment having the highest frequency will have maximum probability of having centroid. If more than one consecutive segments having the same frequency then that segments are merged. After this, distances of data points and centroids are calculated. In same manner the process is continued.

In [9] the research of k-Means clustering algorithm is presented. K-Means algorithm's is less accurate because of selection of k initial centers is randomly. Therefore, in this paper surveyed different approaches for initial centers selection for k-Means algorithm.

In [11] this paper main aim to reduce the initial centroid for k mean algorithm. This paper proposed Hierarchical K-means algorithm. It uses all the clustering algorithm results of K-means and reaches its local optimal. This algorithm is used for the complex clustering cases with large numbers of data set and many dimensional attributes because Hierarchical algorithm in order to determine the initial centroids for K-means.

A synthetic initial starting point is another popular method for selecting the initial centroid points. There are several different ways of coming up with these synthetic starting point values like scrambled midpoints, unscrambled midpoints, and scrambled medians. Several choices can be made when choosing synthetic initial starting values like: 1) dividing the range of feature values into different numbers of partitions, 2) picking initial starting points for each feature within these partitions, and 3) combining the different starting points of the features to create initial starting points. The number of partitions is k-1, k, or k+1, depending on the specific method. It was found that the number of partitions selected does not significantly affect the results of the k-means clustering algorithm[6]. After dividing the workload into partitions, values for each feature in each partition are chosen as initial starting values. Both the median and the midpoint values for each feature in the partitions are evaluated. After coming up with the initial starting point values for each feature, a choice is made of how to combine the different feature values to construct an initial starting point value. Two methods are considered: scrambled and unscrambled. The unscrambled method selects partition #1 of each feature to construct starting point #1, selects partition #2 of each feature to construct starting point #2, and selects partition #k of each feature to construct starting point #k. The scrambled method combines the different partitions of the feature values randomly to create the initial starting points [2].

In contrast to constructing a synthetic starting point, another way of doing it is using an actual sample starting point. There are several different methods of selecting actual samples and using them as starting point values like: random, breakup, and feature value sums. The random method is a popular way of choosing the initial starting point values. It randomly chooses k of the sample data points and uses these as initial starting values [5]. There are variations in the way that random sample data points are chosen. The second method for selecting actual sample data points as starting points is breakup. This method is reasonable since it seeks to break up the most populous cluster into two smaller clusters. It is hoped that a decreased error value. The third method for selecting actual sample data points is feature value sums [2].

3. K-Mean Algorithm

3.1. Overview

According to the basic K-mean clustering algorithm, clusters are fully dependent on the selection of the initial clusters centroids. K data elements are selected as initial centers; then distances of all data elements are calculated by Euclidean distance formula. Data elements having less distance to centroids are moved to the appropriate cluster. The process is continued until no more changes occur in clusters. The following figure shows steps of the basic K-mean clustering algorithm [3, 8].

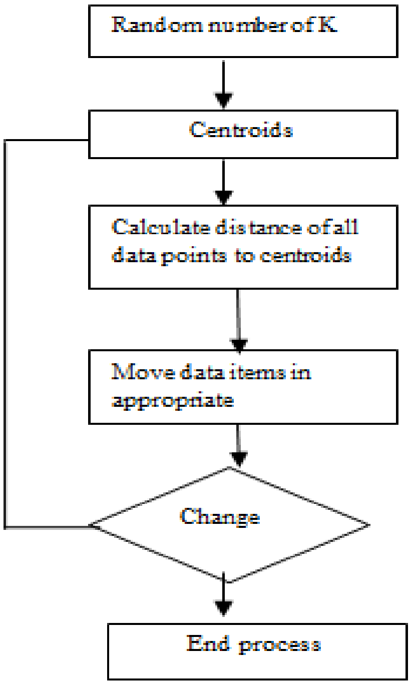
Random number of K
Centroids
Calculate distance of all data points to centroids
Move data items in appropriate
Change
End process

Fig1. *K-Mean Clustering process*

Following are the algorithmic steps for basic K-mean algorithm [10].

1. Choose a number of desired clusters, *k*.
2. Choose *k* starting points to be used as initial estimates of the cluster centroids. These are the initial starting values.
3. Examine each point in the given dataset and assign it to the cluster whose centroid is nearest to it.
4. When each point is assigned to a cluster, recalculate the new *k* centroids.
5. Repeat steps 3 and 4 until no point changes its cluster assignment, or until a maximum number of passes through the data set is performed.

Like other clustering algorithms, k-means requires that a distance metric between points be defined. This distance metric is used in step 3 of the algorithm given above. A common distance metric is the Euclidean distance. Euclidian distance is used as the similarity measure and it is defined as the square root of the sum of squares of differences between feature components of two patterns [1], if $X_i = X_{i1}$; X_{i2}; ……X_{id} and $X_j = X_{j1}$; X_{j2};…..X_{jd} are two pattern vectors then euclidian distance is -

$$d_{ij} = \sqrt{\sum_{k=1}^{d} (x_{ik} - x_{jk})^2}$$

Time complexity of K-mean Clustering is represented by O(*nkt*). Where *n* is the number of objects, *k* is the number of clusters and *t* is the number of iterations [8].

3.2. K-mean Clustering Problems

K-means clustering algorithm works on the assumption that the initial centers are provided. The search for the final clusters or centers starts from these initial centers. Without a proper initialization the algorithm may generate a set of poor final centers and this problem can become serious if the data are clustered using an on-line k-means clustering algorithm. In general, there are three basic problems that normally arise during clustering namely dead centers, local minima and centre redundancy [3]. Dead centers are centers that have no members or associated data. These centers are normally located between two active centers or outside the data range. The problem may arise due to bad initial centers, possibly because the centers have been initialized too far away from the data. Therefore, it is a good idea to select the initial centers randomly from the training data or to set them to some random values within the data range. However, this does not guarantee that all the centers are equally active. Some centers may have too many members and be frequently updated during the clustering process whereas some other centers may have only a few members and are hardly ever updated [6].

4. PROPOSED METHOD AND ALGORITHM

4.1. Centroid Determination Method

Given the value of K which is the number of clusters to be formed by the user as an input to the K-Means algorithm, next important work is to select K numbers of cluster centroid points from the given dataset which the normal K-Means algorithm does randomly. But the random selection leads to more number of iterations also choosing different centroid points every time gives different clusters thus leading to almost wrong output. To overcome these problems centroid determination method has been proposed which uses neighborhood distance between the points to determine the best possible cluster centroid points which will at least reduce the number of times K-Means algorithm need to re-iterate.

Considering a dataset with n points, take the very first point of the given dataset and find out its distance to all the other (n-1) points in the dataset using:

$$d(Pi) = \sum_{i=1}^{n} dis\tan ce(Pi, Xi)$$

Here,

$d(P_i)$ = Stores the distance of first point in the dataset to all others.

Next step is to sort all the points and arrange it based on this sorted distance. Assuming that the user has entered the value of K i.e. the number of clusters to be formed, we will divide the entire dataset into k numbers of equal proportion and select the very first point of each proportion as a cluster centroid. Once this selection is done, next step is to apply the normal k-Means algorithm.

4.2. Proposed K-Means algorithm

Step 1: Accept the number of clusters K to group data into and the dataset to cluster as input values.

Step 2a: calculate the distance of first point to all other points in the dataset using:

$$d(Pi)=\sum_{i=1}^{n} dis\tan ce(Pi, Xi)$$

Step 2b: Arrange all the points in the dataset according to the above sorted distance using Sort {d(P_i)}

Step 3a: Divide the entire dataset into K equal proportion.

Step 3b: Choose the first point of every proportion as the K different initial cluster centroid points.

Step 4: Examine each point in the given dataset and assign it to the cluster whose centroid is nearest to it based on Euclidean distance.

Step 5: Calculate the arithmetic means of each cluster formed in the dataset and recalculate the new K centroids.

Step 6: repeat steps 4 and 5 until no point changes its cluster assignment, or until a maximum number of passes through the data set is performed.

5. CONCLUSION

Existing methods of selecting the number of clusters and the initial centroid points for K-means clustering algorithm have a number of drawbacks. Simple examples show that the initial starting point selection may have a significant effect on the results of the algorithm, both in the number of clusters found and their centroids. An overview of the existing methods of choosing the value of K i.e. the number of clusters along with new method to select the initial centroid points for the K-means algorithm has been proposed in the paper along with the modified K-Means algorithm to overcome the deficiency of the classical K-means clustering algorithm. The new method is closely related to the approach of K-means clustering because it takes into account information reflecting the performance of the algorithm. The improved version of the algorithm uses a systematic way to find initial centroid points which reduces the number of dataset scans and will produce better accuracy in less number of iteration with the traditional algorithm. The method could be computationally expensive if used with large data sets because it requires calculating the distance of every point with the first point of the given dataset as a very first step of the algorithm and sort it based on this distance. However this drawback could be taken care by using multi threading technique while implementing it within the program. However further research is required to verify the capability of this method when applied to data sets with more complex object distributions.

REFERENCES

[1] Al-Daoud, M. B., Venkateswarlu, N. B., and Roberts, S. A. New methods for the initialisation of clusters. Pattern Recognition Lett., 1996, 17, 451–455.

[2] Bradley, S. and Fayyad, U. M. Refining initial points for K-means clustering. In Proceedings of the Fifteenth International Conference on Machine Learning (ICML ‘98) (Ed. J. Shavlik), Madison, Wisconsin, 1998, pp. 91–99 (Morgan Kaufmann, San Francisco, California).

[3] Fahim A. M., Salem A. M., Torkey F. A. and Ramadan M. A., “An efficient enhanced k-means clustering algorithm,” Journal of Zhejiang University Science A., pp. 1626–1633, 2006

[4] Frank Robinson, Amy Apon, “Initial Starting Point Analysis for K-Means Clustering: A Case Study”.

[5] Han, J. and Kamber, M. Data Mining: Concepts and Techniques, 2000 (Morgan Kaufmann, San Francisco, California).

[6] Jain, A., Murty, M., Flynn, P., “Data Clustering: A Review,” http://scgwiki.iam.unibe.ch:8080/SCG/uploads/596/p264-jain.pdf.

[7] Jieming Zhou, J.G. and X. Chen, "An Enhancement of K-means Clustering Algorithm," in Business Intelligence and Financial Engineering, BIFE '09. International Conference on, Beijing, 2009.

[8] K. Jain and R. C. Dubes, Algorithms for Clustering Data. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[9] M.P.S Bhatia, Deepika Khurana Analysis of Initial Centers for k-Means Clustering Algorithm International Journal of Computer Applications (0975 – 8887) Volume 71– No.5, May 2013

[10] Master, C.P. and X.G. Professor, 2011. "A Brief Study on Clustering Methods Based on the K-means algorithm," in 2011 International Conference on E-Business and E-Government (ICEE), Shanghai, China.

[11] P.S. Bradley and U. Fayyad, ªRefining Initial Points for K-means Clustering,º Proc. 15th Int'l Conf. Machine Learning, pp. 91-99, 1998.