K-Means Clustering - Review of Various Methods for Initial Selection of Centroids

Ms.Beena P, Mr. Sunil Kumar P V, Mr. Balachandran K P

Abstract— Organising data into sensible groups is the most fundamental way of understanding and learning. Clustering helps to organise data based on natural grouping without any category labels to identify the clusters.One of the most popular and simplest partitional clustering algorithms is the K -Means published in 1955. K -Means algorithm is computationally expensive and insists the selection of number of clusters initially. The final clusters depend entirely on the initial selection of centroids. Several modifications have been proposed for the K -Means clustering method. Some such proposals are summarised and reviewed with experimental results.

Index Terms— Algorithms, Cluster analysis, K-means algorithm, Initial centroid, Enhanced, Heuristic, Variation coefficient, Voronoi diagram, Purity.

1 INTRODUCTION

ADVANCES in scientific data collection methods have resulted in the large scale accumulation of scientific data at various data sources. The amount of information available is becoming enormous and tremendous day by day. It is practically difficult to analyze and interpret the data using conventional methods. Effective and efficient data analysis methods are necessary to extract useful information. Cluster analysis is one of the major data mining methods which helps in identifying the natural groupings and interesting patterns from huge data banks.

Data clustering is a process of identifying the natural grouping that exist in a given data-set, such that the patterns in the same cluster are more similar and the patterns in different clusters are less similar. Clustering is a main task of explorative data mining, and a common technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics.

Clustering algorithms are broadly divided into two groups, viz. Hierarchical and partitioned. Hierarchical clustering algorithms find the clusters in agglomerative (bottom-up) mode or in divisive (top-down) mode recursively where as partitional clustering algorithms find all the clusters simulaneously as a partition of the data set. Apart from this, the clustering methods can also be categorized into density-based methods, grid-based methods, model-based methods, etc.

- *Mrs. Beena P currently pursuing masters degree program in Computer Science engineering in University of Calicut E-mail : pbeenagpt@yahoo.co.in*

2 K -MEANS CLUSTERING

The most popular, the simplest, efficient partitional clustering method is the K -Means clustering. The given set of data is grouped into K number of disjoint clusters, where the value of K is fixed in advance. The algorithm consists of two separate phases : the first phase defines K initial centroids, one for each cluster. The next phase associates each point of the given data set to the nearest centroid. Generally Euclidean distance is used as the measure to determine the distance between data points and the centroids. Euclidean distance between two data points X (x_1 , x_2 , . . . , x_n) and Y (y_1 , y_2 , . . . , y_n) is given by (1)

$$\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + ...(x_n - y_n)^2} \qquad (1)$$

When all the points are included in some clusters, the first step is completed and an early grouping is done. Now the centroids are recalculated and clustering is done with these new centroids. This is repeated till the centroids do not change. This is the convergence criterion for the K -Means. The pseudo code for the K -Means clustering is outlined below[2].

Algorithm : K-Means Clustering

Input :

D = d_1, d_2, …..,d_n / set of n data items.

K // Number of desired clusters.

Output :

A set of K clusters.

Steps:

1. Arbitrarily choose K data items from D as initial centroids;

2. Repeat

Assign each item d_i to the cluster which has the closest centroid;

Calculate the new mean for each cluster;

3. Until convergence criterion is met.

Though the algorithm is effective in producing clusters for many practical applications, there are some draw backs. The computational complexity of the original K -Means algorithm is very high, especially for large data sets. The complexity is O(nK l) where n is the number of data points, K the number of clusters and l the number of iterations. Also it different types of clusters based on the selection of initial centroids. Accuracy of the final clusters heavily depends on the initial centroids selected.

3 LITERATURE SURVEY

Anil K. Jain discussed major challenges and key issues in clustering [1]. He provided a brief overview of clustering and summarized well known clustering methods. He also discussed the major challenges and key issues in designing clustering algorithms, and pointed out some of the emerging and useful research directions, including semi-supervised clustering, ensemble clustering, simultaneous feature selection during data clustering, and large scale data clustering.

Researchers are always been conducted to improve the accuracy and efficiency of the K -means algorithm. Some of these innovative approaches to K-Means clustering are discussed in this survey. Though the time complexity is not improved, these proposals could fix the initial centroids and the inconsistency in clustering got reduced. In all these proposals the adjacency between the points are measured by the Euclidean distance between them.

3.1 Efficient Enhanced K-Means

Fahim et al. [3] proposed an Efficient Enhanced K - Means algorithm which refines the second phase of the K -Means suitable for data set with large number of clusters. Fahim's approach makes use of a distance function based on a heuristic to reduce the number of distance calculations. For each data point the distance to the nearest cluster is preserved as an attribute. At the next iteration, the distance to the present nearest cluster is calculated. If this distance is less than or equal to the distance attribute associated with the data point, the point stays in its cluster and is not necessary to compute its distances to the other cluster centers. This saves the time required to compute distances to K − 1 cluster centers. Thus the complexity is reduced. As the initial centroids are determined randomly, there is no guarantee for the accuracy of the final clusters.

3.2 Enhanced K -Means

Abdul Nazeer et al. [4], [5] proposed an algorithm addressing the problem with the initial centroid as well as the time complexity. The two phases of the original K -Means were modified. The second phase is a variant of the method proposed in [3]. In the first phase the distance between each data pair is calculated and the closest data points are grouped to form the clusters. A threshold value is fixed for the number of elements in the cluster. The threshold is 0.75 * (n/K), where n is the size of the data set. The mean of these clusters form the initial centroids.

These initial centroids are the input to the second phase, for assigning data point to the appropriate clusters. The second phase is a variation of the method proposed in [3].

3.3 K -Means with Heuristic to fix the initial centroids

K. A. Abdul Nazeer et al. [6] Modified his work in [4] by a heuristic method for finding better initial centroids. He used the second phase of the original K-Means without any modification. The initial centroids of the data points are determined using a heuristic approach. The data points are sorted in ascending order and divided into K number of sets. Mean values of each of these sets are then taken as the initial centroids of the clusters. If the data set is multidimensional each data point may contain multiple attributes. Then the sorting is based on the attribute with maximum range, where range is the difference between the maximum and the minimum values of the attribute. The phase I complexity is reduced to O (n log n).

3.4 Heuristic K -Means

K. A. Abdul Nazeer et al. [7] modified his work in [6] by modifying the second phase. He used the second phase of [3] which is a variant of the method proposed in [3]. The complexity is the same as that of the method in [6].

3.5 Weighted Ranking K -Means

R.Sumathi et al. [8] suggested a weighted ranking algorithm. Weights fixed by the experts are assigned to the attributes of data points. The work is an extension of [4] and produced meaningful clusters. An alternate view of the attributes is presented here. For finding the initial centroid, this

algorithm calculates the sum of Euclidian distance from origin to weighted attributes (Wi Xi) as in equation (2)

$$\underset{j=1}{\overset{n}{U}} = \sum_{i=1}^{m} W_i X_i \qquad (2)$$

Where W is the weight and X is the attribute. The complexi ty is same as that of Enhanced K -Means [4].

3.6 Variation and Correlation Coefficient K -Means

Murat Erisoglu et al. [9] proposed a new method for finding the initial centroids which are well separated. It selects the two attributes that best describe the change in the data set with the help of variation coefficients and correlation coefficients according to two axes. The second phase of the original K -means is used for the clustering. The method produced an improved and consistent cluster structures.

This method initially selects the attribute having the maximum absolute value of the variation coefficient. Variation coefficient is determined by the ratio of standard deviation and the mean of each attribute. The attribute with minimum correlation coefficient is taken as the second axis. The mean of the of data points is selected as the centre of the dataset according to the selected two axes. The data point which is farthest from the mean is fixed as the first centroid C1. The data point which is farthest from the first centroid C1 is fixed as the second centroid C2 . Euclidean distance is taken as the distance measure. To fix the remaining centroids the sum of the distances of each data point from each of the centroids fixed so far is calculated. The largest value of this distance determines the centroid. To reduce the complexity the distance of the data point from the newly fixed centroid is added to the previously accumulated sum. This accumulation scheme helps to determine centroids that are spread out.

3.7 Voronoi K -Means

The method suggested by Damodar Reddy et al. [10] selects initial centroids with the help of voronoi diagram constructed with the data set. The initial centroids are those points that lie on the boundary of higher radius voronoi circles. The centroids thus generated are the input to the second phase of K-Means. The second phase is same as that of the original K-Means.

Given a set of n points $S = p_1 , p_2 , \ldots , p_n$ in a m- dimensional Euclidean space, the Voronoi diagram of S is defined as the subdivision of the space into n cells such that each point belongs to only one cell. If d(a, b) denote the distance between the points a and b in the Voronoi diagram Vor(S) of S then the definition implies that a point u lies in the cell corresponding to the point p_i iff $d(u, p_i) < d(u, p_j)$ for each $p_j \in S$ and $j = i$. For a Voronoi vertex v, the largest empty circle of v with respect to S is defined as the largest circle with v as its center that contains no point of S in its interior denoted by CirS(v). The Voronoi vertices have the property that a point q is a vertex of Vor(S) iff CirS(q) contains three or more points of S on its boundary. There can be a maximum of $2n - 5$ Voronoi vertices in a Voronoi diagram of n points. The initial centers are selected from those points which lie on the boundary of higher radius Voronoi circles. Consistent initial centroids are fixed with the computational complexity as that of the original K -Means.

4 Experiments and Results

The following methods were implemented and tested. Implementations were done in Java.

- K-Means[2].
- Enhanced K-Means with initial centroids by heuristic [6].
- Heuristic K-Means [7].
- K-means with initial centroids by variation and correlation coefficients [9].

The data sets available in the UCI data repository were used for testing. The data sets available in the UCI data repository were used for testing. Iris is Iris plants database. Spambase is a spam e-mail database. BCW is the breast cancer Wisconsin (original) data set. Details of the data sets used are summarized in Table 1.

TABLE 1
DATA SETS

Dataset	No. of Samples	No. of Attributes	No. of Clusters
Iris	150	4	3
Spambase	4601	57	2
BCW	699	9	2

Purity[14] of the clusters Cj is used as the measure for testing. It's a measure of correctly classified data points. Purity of a cluster is defined by equation (3).

$$Purity(C_j) = \frac{1}{|C_j|} \max |C_j|_{class=i} \qquad (3)$$

where $|C_j|_{class=i}$ denotes the number of items of class i assigned to cluster j. Overall cluster purity is given by equation(4)

$$Purity(C_j) = \sum_{j=1}^{k} \frac{|C_j|}{|D|} Purity(C_j) \qquad (4)$$

The results vary with data sets and are tabulated in table 2. For the original K -Means results of three executions are averaged and tabulated.

TABLE 2
COMPARISON

Algorithm	Purity		
	Iris	Spambase	BCW
K –Means[2]	81.56	63.6	96.1
K-Means with Heuristic[6]	88.67	63.6	96.2
Heuristic K -Means[7]	92	67.88	97.5
Variation and Correlation Coeft K-Means[9]	88.67	63.6	96.05

5 CONCLUSION AND FUTURE WORKS

These algorithms eliminated the inconsistency with initial centroid selection and produced better result. The following problems still persist with these variations also.

- K the number of clusters should be fixed beforehand.
- The computational complexity is very high.
- Mode of selection of initial centroids determines the cluster purity.

Researches in the field of improving the performance of K -Means could not develop a widely accepted version. A modification can be tried out to improve the accuracy and complexity of the K -Means incorporating standard deviation and skewness.

References

[1] Anil K. Jain. Data Clustering : 50 years beyond K- means. Pattern Recogni tion Letters Elsevier,31(8):651–666, 2010.W.-K. Chen, Linear Networks and Systems. Belmont, Calif.: Wadsworth, pp. 123-135, 1993. (Book style).

[2] Margaret H. Dunham, Data Mining- Introductory and Advanced Concepts, Pearson Education, 2012.

[3] Fahim A.M, Salem A.M, Torkey F.A and Ramadan M.A. An efficient enhanced k-means clustering algorithm.Journal of Zhejiang Univer- sity SCIENCE A ISSN 1009-3095 (Print); ISSN 1862-1775 (Online),www.springerlink.com, 7(10), 2006.K.

[4] K.A.A. Nazeer and M.P. Sebastian. Clustering Biological Data Using enhanced k-Means Algorithm. Springer Netherlands, First edition, 2010.C. J. Kaufman, Rocky Mountain Research Laboratories, Boulder, Colo., personal communication, 1992. (Personal communication)

[5] M.P.Sebastian and K.A. Abdul Nazeer. Improving the accuracy and efficiency of the k-means clustering algorithm. In Proceedings of the World Congress on Engineering 2009 , Vol I July 2009.

[6] M.P.Sebastian, K.A Abdul Nazeer and S.D.Madhu Kumar. Enhancing the k-means clustering algorithm by using a O(n logn) heuristic method for finding better initial centroids In Second International Conference on Emerging Applications of Information Technology IEEE , Feb 2011 pp. 261 –264.

[7] M.P.Sebastian, K.A. Abdul Nazeer and S.D.Madhu Kumar. A Heuristic k-Means Algorithm with Better Accuracy and Efficiency for Clustering Health Informatics Data American Scientific Publishers Journal of Medical Imaging and Health Informatics , 1:66–71, 2011.

[8] R.Sumathi and E.Kirubakaran. Enhanced Weighted K-Means Clustering Based Risk Level Prediction for Coronary Heart Disease European Journal of Scientific Research ,ISSN 1450-216X

[9] Murat Erisoglu , Nazif Calis and Sadullah Sakalli- ogl. A new algorithm for initial cluster centers in k-means algorithm. Pattern Recognition Letters Elsevier, 32(14):1701–1705, October 2011.

[10] Damodar Reddya and Prasanta K. Janaa. Initialization for K-means clustering using Voronoi diagram. Procedia Technology Elsevier, 4:395–400, October 2012.

[11] Bryan Bergeron, Bioinformatics Computing " in PHI.

[12] The UCI Repository website. [Online]. Available: http://archive.ics.uci.edu/

[13] O. L. Mangasarian and W. H. Wolberg: "Cancer diagnosis via linear programming", SIAM News, Volume 23, Number 5, September 1990, pp 1 & 18.

[14] Christopher D. Manning, Prabhakar Raghavan & Hinrich Schütze; : Introduction to Information Retrieval ; Website:http: //informationretrieval.org/ Cambridge University Press © 2008 Cambridge University Press.