K-Means The Classical Clustering Method

BEENA P
11MCS3302
Guided by: Mr.Sunil Kumar P V(Asst.Professor CSE Dept)

January 28, 2013

Keywords: Clustering algorithms, Cluster analysis, K-means algorithm, Initial centroid, Voronoi diagram, Hybrid clustering

Abstract

The amount of information available is becoming enormous and tremendous day by day. It is practically dicult 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 and are mainly categorised into- Hierarchical and Partitioned. There are some other classifications also - Density-based methods, Model-based clustering and Grid-based method. The K-Means clustering belongs to the partitonal method and is the most popular method due to its simplicity. This method is not an efective method and faces some challenges.

- Inconsistency of the clusters with varying initial centroids.
- High time complexity.

A lot of researches are being conducted to improve the performance of K-Means Algorithm. Some such innovations are analysed here. The methods are

- Enhanced K-Means[3]
- Enhanced K-Means with initial centroids by heuristic[5]
- Heuristic K-Means[6] Weighted Ranking K-Means[7]
- K-means with initial centroids by new method[8]
- K-means with initial centroids by Voronoi diagram[9]
- Hybrid LK-Means[10]

1 Introduction

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 algorithms can be broadly divided into two groups, viz., hierarchical and partitional.There are some other classifications also - Density-based methods, Model-based clustering and Grid-based method.

1.1 K-Means Clustering

The most popular, the simplest, efficient partitional clustering method is the K-Means clustering method. 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 is to define k initial 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. Generally Euclidean distance is used as the measure to determine the distance between data points and the centroids. 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.

2 Literature Survey

Researches are always being conducted to improve the accuracy and efficiency of the K-means algorithm. Some of the innovative approaches to K-Means clustering are described here. K-Means algorithm can be divided into two phases. The first phase determines the initial centroids and the second associates data points to the initial centroids and refines the centroids.

The paper Data Clustering: 50 years beyond K-Means by Anil .K.Jain discusses the major challenges and key issues in clustering.

Fahim et al. [2] proposed an efficient enhanced k-means algorithm which refines the second phase of the algorithm. Fahims approach makes use of a distance function based on a heuristics to reduce the number of distance calculations. As the initial centroids are determined randomly there is no guarantee for the accuracy of the final clusters.

K. A. Abdul Nazeer et al. [3] [4] proposed an enhanced algorithm by considering relative distance of each point. The first phase, the initial centroids are determined systematically to produce clusters with better accuracy]. The second phase makes use of a variant of the clustering method discussed in [2] that suits for spherical shaped clusters. Though this algorithm produces clusters with better accuracy and efficiency compared to k-means, it is also computationally expensive with a time complexity of $O(n^2)$.

K. A. Abdul Nazeer et al. [5] modified his work in [3] by a heuristic method for finding better initial centroids. He used the second phase of the original K-Means without any modification. This method selects the attribute based on which the data points are to be clustered.

K. A. Abdul Nazeer et al. [6] refined work in [5] with a variant of the method in [2] for the second phase. This definitely improved the accuracy and efficiency.

R.Sumathi et al. [7] suggested a weighted ranking algorithm. Weights to the attributes of data points are assigned by experts. The work is an extension of [6] and produced meaningful clusters.

Murat Erisoglu et al. [8] proposed a new method for finding the initial centroids which are well separated. It selects the two attributes that best describe the data set with the help of variation coefficients and correlation coefficients. The second phase of the original K-means was used for the clustering. The method produced an improved and consistent cluster structures.

The method suggested by Damodar Reddy et al. [9] 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 literature survey also included a hybrid approach suggested by T.Hithendra Sarma et al. [10]. This paper presented a prototype based hybrid approach to speed up the clustering. The data set is partitioned into small clusters each represented by a prototype. These prototypes become the candidate for clustering. A correction is also proposed for the final clusters generated. The method is suitable for high dimensional large data sets.

3 Conclusion

The original K-means algorithm is computationally expensive and the quality of the resulting clusters heavily depends upon the selection of the initial centroids. Though much researches are going on in this field a widely accepted version of K-Means is yet to come. The hybrid algorithm discussed in this paper could reduce the time complexity, but the inherent problem of inconsistency of the clusters with varying initial centroids still remians. The initial algorithms could fix this problem to an extent. An hybrid version with weighted attributes may address the problems.

References

[1] Anil K. Jain. Data Clustering : 50 years beyond K-means. *Pattern Recognition Letters Elsevier*, 31(8):651–666, 2010.

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

[3] K.A.A. Nazeer and M.P. Sebastian. *Clustering Biological Data Using enhanced k-Means Algorithm.* Springer Netherlands, First edition, 2010.

[4] 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.

[5] 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.

[6] 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.

[7] 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 71(4):490–500 2012 .

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

[9] Damodar Reddya . Prasanta K. Janaa Initialization for K-means clustering using Voronoi diagram. *Procedia Technology Elsevier*, 4(0):395–400, October 2012.

[10] T.H Sarma . P. Viswanath . B. Reddy A hybrid approach to speed-up the k-means clustering method. *nternational Journal of Machine Learning and Cybernetics Springer-Verlag*, pp.1–11, January 2012.