A Survey: Different Improvements and Integrated Approaches of K-means

Manisha Goyal[1], Mr. M.B. Chaudhary[2], Ms. Pinal Patel[3]
[1]*Research Scholar, Computer and science engineering, Government Engineering College, Sector 28, Gandhinagar-382028, India*
[1,2]*Professor, Computer and science engineering, Government Engineering College, Sector 28, Gandhinagar-382028, India*

***Abstract*- Cluster analysis is decades' old concept of data mining which performs division of data into groups of similar objects. It is used in various applications in the real world such as data/text mining, voice mining, image processing, web mining, medical data mining and many others. The clustering algorithms are categorized based on different research strategies. One of the widely used clustering methods is the K-Means clustering method which is increasing in popularity day by day because of its simplicity and linear time complexity. However, it has two main disadvantages: - 1) Its highly sensitive to outlier and 2) Its highly dependent on initialization parameters (random choice of k clusters and position of initial cluster centroids). Due to this parametric nature of obtaining inputs from the user, the performance of this algorithm is highly dependent on the nature of inputs. Many improved variants of K-means method is detailed in literature but still it is an open field of research because of its extensive application in the field of Medical, Business & Marketing, Social Media – Sentiment Analysis etc. The aim of this survey paper is to discuss different integrated approaches of k-means which are developed to overcome the aforementioned limitations of the algorithm. This paper also discusses the overlapping version of the K-means algorithm which gives the concept of overlapped clusters.**

***Index Terms*- K-means clustering, Cluster Analysis, Overlapping Clustering.**

INTRODUCTION

Clustering- It is an unsupervised learning problem which deals with finding a structure in a collection of unlabelled data. The main objective of clustering is to find natural groupings among objects. It organizes data in such a way that itracluster similarity is minimized and intercluster similarity is maximized. Clustering method has many real time application in the fields like medical domain for disease prediction, medical image segmentation, biological sequence analysis; Market basket analysis, Pattern Recognition, Text segmentation and Social Media Analysis.

Clustering methods can be categorized according to the following criteria[1]:

1. Type of input data: To deal with different types of input data such as numerical, categorical and mixed, different clustering methods are used.
2. Type of proximity measures: Different types of similarity measures are defined to deal with different type of input data, some of them are Euclidean distance, Manhattan distance etc.
3. Type of generated cluster: In this category two types of clustering methods are defined one is Exclusive (Non-Overlapping) another one is Overlapping.
4. Type of clustering strategy used: In term of cluster strategy, clustering methods are divided into six groups of partitioning, hierarchical, density-based, model-based, graph-based and grid-based.

K-means Clustering- It is the partitioning method for clustering. The aim of the algorithm is to take the input parameter (1) number of clusters (k) and (2) Initial k centroids randomly then partitions n objects into k clusters. Cluster mean is used to update the centroid of that cluster.

The objective function of K-means is, defined as: $E = \sum_{i=1}^{k} \sum_{p \in c_i} ||p - m_i||^2$

Where E is the sum of the square error for all objects in the data set; p represent a given object; and m_i is the mean of cluster C_i.

Algorithm works based on the following steps:

1. Initialization: This is the initial step, in which initialization parameters like number of groups or clusters (k), and initial centroids of the regarding groups and defined.
2. Cluster Assignment: Similarity measures like Euclidean distance of each datapoint from each of the centroid is calculated and according to the closest distance, datapoints are assigned to their respective clusters.
3. Centroid Repositioning: Mean of the each cluster is taken and this mean becomes the new centroid for that cluster.
4. Optimization and Convergence: Repeat previous 2 steps iteratively till the cluster centroid stop changing their position. At some point cluster does not change for further computation that is the point when algorithm converges.

Advantages of the algorithm:

- Easy to implement and robust.
- More efficient and scalable in processing large data sets.
- Having linear time complexity.
- Produce tighter clusters than hierarchical clustering.

Limitations of algorithm:

- Cannot be applied on categorical attributes.
- Sensitive to the selection of number of a clusters k and initial cluster center.
- Not suitable for discovering clusters with different shapes or size.
- More sensitive to noise and outliers.

DIFFERENT VARIANTS OF K-MEANS

Many extensions or variants of K-means have been proposed by researchers to overcome the above limitations and to make it more accurate. There is a brief summary of these variants[2].

Variants for Initialization Parameters:[1][3][4]

Title	Description
K-means++	This is an extension of k-means which provide initial centroid by using weighted probability distribution. Probability of each object is considered to be proportional to the Euclidean distance of that point from the most adjacent centroid.
Sorted K-means	It determines the initial centroid after sorting the data points on the basis of their distances from the origin. For sorting the objects quick or merge sort can be used according to the application.
KHM Method	K-harmonic means method is less sensitive to the outlier as it compute the centroids of clusters based on the harmonic average instead of arithmetic average. The harmonic mean gives a weight to each data point based on its distance measures to each centroid.
Density K-means(DKM)	It first calculate local density of each data point in the dataset then it compute the initial centroid in incremental fashion. Each data points has assigned an attribute potential according to its measured density. Now, the data point having the highest potential is added to the empty set of initial centroid. Recalculate the potential of each data points by multiplied it with the Euclidean distance between that data point to recently added centroid and choose highest potential for next centroid. Repeat the process until k centroids are not found.

Variants to improve accuracy of the algorithm:[2]

Title	Description
K-medians	It computes medians of clusters instead of mean to update the centroid. Outlier effect is lesser on it as compared to k-means.
K-modes	This method improves the k-means to make it applicable for categorical data by calculating modes of categorical data instead of mean. It uses matching dissimilarity measures rather than Euclidean distance to represent cluster center.
K-medoids	It tries to remove the sensitivity towards outlier which can distort the distribution of data. Instead of taking mean value of the cluster objects as a reference point, it chooses an object per cluster as representative of that cluster. Then dissimilarity is measured between this representative object and rest of the objects and on the basis of this measure partitioning is performed.

Overlapped variants of K-means:[5]

Title	Description
FCM	It is first overlapped version of K-means, also called fuzzy K-means which is based on the concept of fuzziness. Here data points are assigned to a particular cluster with membership degree (between 0 to 1). A data with highest membership for the particular cluster is assigned to that cluster.
OKM	Overlapping K-means uses heuristic approach to assign data points to one or more clusters. Distance from each data points and clusters centroids are calculated and assignment of data points to multiple clusters is done by sorting the clusters from nearest to farthest.
WOKM	It is the extension of OKM and Weighted K-means which include a weighting factor into the objective function of OKM and this weighting factor is used to cluster the data points more appropriately.

Hierarchical variants of K-means:[6][7][8]

Title	Description
HK means	In this integrated approach hierarchical and k-means clustering methods are integrated to take advantages of both technique. It first determine the initial parameters for the k-means by using agglomerative method then k-means is applied on the dataset on the bases of that parameters. It gives more time complexity for huge dataset.
iHK means	It is an improved version of HK method which construct N/2 data cluster of N objects by using K-means method then mean vector of each cluster is used as an input for agglomerative method. Thus, it improve time complexity of algorithm.
BIRCH-kmeans	This method combines BIRCH and K-means to find appropriate clusters. It first generate a B- tree by inserting and splitting data points using BIRCH algorithm which gives a large number of clusters. After that K-means method has been called to construct the clusters from the leaf nodes of the tree thus reduce the clusters and gives the k clusters with increased accuracy.

CONCLUSION

Despite of number of limitations of K-means partitioning algorithm it is widely used in clustering analysis and in prediction analysis. Its scope is not limited in one particular domain. Many improvements have been done to overcome the existing limitations of K-means to improve its performance. This survey show that k-means can be improved further to make it more accurate and applicable for different application.

REFERENCES

[1] S. Khanmohammadi, N. Adibeig, and S. Shanehbandy, "An improved overlapping k-means clustering method for medical applications," Expert Syst. Appl., vol. 67, pp. 12–18, 2017.

[2] A. Choudhary, "Survey on K-Means and Its Variants," Int. J. Innov. Res. Comput. Commun. Eng., vol. 4, no. 1, pp. 949–953, 2016.

[3] P. Arora, D. Virmani, and H. Jindal, "Proceedings of International Conference on Communication and Networks," Springer, Singapore, vol. 508, pp. 479–486, 2017.

[4] N. Nidheesh, K. A. Abdul Nazeer, and P. M. Ameer, "An enhanced deterministic K-Means clustering algorithm for cancer subtype prediction from gene expression data," Comput. Biol. Med., vol. 91, pp. 213–221, 2017.

[5] S. Baadel, F. Thabtah, and J. Lu, "Overlapping clustering: A review," Proc. 2016 SAI Comput. Conf. SAI 2016, pp. 233–237, 2016.

[6] M. E. Celebi and H. A. Kingravi, "Deterministic initialization of the k-means algorithm using hierarchical clustering," Int. J. Pattern Recognit. Arti¯cial Intell., vol. 26, no. 7, pp. 1250018(1–25), 2012.

[7] W. Liu et al., "Improved Hierarchical K-means Clustering Algorithm without Iteration Based on Distance Measurement," Springer, vol. 432, pp. 38–46, 2016.

[8] J. Kaur and H. Singh, "Performance evaluation of a novel hybrid clustering algorithm using birch and K-means," 2015 Annu. IEEE India Conf., pp. 1–6, 2015.