ISSN 2394-739X

International Journal for Research in Science Engineering and Technology

ANALYSIS OF CLUSTERING IN K-MEANS AND K-MEDOIDS

[1] **Dr Antony Selvadoss Thanamani,**
[1] Associate Professor and Head,
[1] NGM College,
[1] Pollachi.

[2] **M. MohanaPriya,**
[2] Folio-B1/371/2014,
[2] M.Phil Scholar, NGM College,
[2] Bharathiar University Pollachi.

Abstract:-

Cluster analysis has long been used in a wide variety of fields, psychology and other social sciences, biology, statistics, pattern recognition, information retrieval, machine learning, and data mining. Cluster analysis divides data into meaningful or useful groups. Data mining adds to clustering the complications of very large datasets with very many attributes of different types. There are various types of algorithmsin data mining process. Clustering has taken its roots from algorithms like k-means and k medoids. From these algorithm k-means algorithmis evolved. K-means is very popular because it is conceptually simple and is computationally fast and memory efficient but thereare various types of limitations in k means algorithm that makes extraction somewhat difficult. K-medoids clustering algorithm suffers from many limitations. In this paper we are discussing these merits and demerits in k-means and k-medoidshow to overcome

Keywords: - K-Means, patterns, K-Medoids

1. INTRODUCTION

Clustering can be said as identification of similar classes of objects. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes. Classification approach can also be used for effective means of distinguishing groups or classes of object but it becomes costly so clustering can be used as preprocessing approach for attribute subset selection and classification. For example, to form group of customers based on purchasing patterns, to categories genes with similar functionality.

2. DATA MINING AND CLUSTERING METHODS

Data mining - also known as knowledge-discovery in databases (KDD) is process of extracting potentially useful information from raw data. A software engine can scan large amounts of data and automatically report interesting patterns without requiring human intervention. Cluster analysis is an iterative process of clustering and cluster verification by the user facilitated with clustering algorithms, cluster validation methods, visualization and domain knowledge to databases. Other knowledge discovery technologies are Statistical Analysis, OLAP, Data Visualization, and Ad hoc queries. Unlike these technologies, data mining does not require a human to ask specific questions. In general, Data mining has four major relationships. They are:

i. Classes
ii. Clusters
iii. Associations

iv. Sequential patterns.

(i) Classes: Stored data is used to locate data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials.

(ii) Clusters: Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities.

(iii) Associations: Data can be mined to identify associations. The beer-diaper example is an example of associative mining.

(iv) Sequential patterns: Data is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

3. TYPES OF CLUSTERING METHODS

Clustering algorithms can be categorized based on their cluster model, as listed,

A. Partitioning Methods
B. Hierarchical Agglomerative methods
C. Density based methods
D. Grid-based methods
E. Model-based methods

A. Partitioning methods

Partitioning clustering algorithms, such as K-means, K-medoids PAM, CLARA and CLARANS assign objects into k clusters, and iteratively reallocate objects to improve the quality of clustering results. K-means is the most popular and easy-to understand clustering algorithm. K-means algorithm is very sensitive to the selection of the initial centroids, in other words, the different centroids may produce significant differences of clustering results. Another drawback of K-means is that, there is no general theoretical solution to find the optimal number of clusters for any given data set. A simple solution would be to compare the results of multiple runs with different k numbers and choose the best one according to a given criterion, but when the data size is large, it would be very time consuming to have multiple runs of K-means and the comparison of clustering results after each run. Instead of using the mean value of data objects in a cluster as the center of the cluster, a variation of K-means, K-medoids calculates the medoid of the objects in each cluster.

B. Hierarchical methods

Hierarchical clustering algorithms assign objects in tree structured clusters, i.e., a cluster can have data points or representatives of low level clusters. Hierarchical clustering algorithms can be classified into categories according their clustering process: agglomerative and divisive.

Agglomerative: One starts with each of the units in a separate cluster and ends up with a single cluster that contains all units. Agglomerative Nesting arranges each object as a cluster at the beginning, then merges them as upper level clusters by given agglomerative criteria step-by-step until all objects form a cluster

Divisive: To start with a single cluster of all units and then form new clusters by dividing those that had been determined at previous stages until one ends up with clusters containing individual units. Divisive Analysis adopts an opposite merging strategy; it initially puts all objects in one cluster, and then splits them into several level clusters until each cluster contains only one object.

C. Density based methods

This method is based on the notion of density. The primary idea of density-based methods is that for each point of a cluster the neighborhood of a given unit distance contains at least a minimum number of points, i.e. the density in the neighborhood should reach some threshold. However, this idea is based on the assumption of that the clusters are in the spherical or regular shapes. DBSCAN algorithm is an important part of clustering technique which is mainly used in scientific literature. Density is measured by the number of objects which are nearest the cluster.

D. Grid-based methods

The idea of grid-based clustering methods is based on the clustering oriented query answering in multilevel grid structures. The upper level stores the summary of the information of its next level, thus the grids make cells between the connected levels. The two examples of Grid-based methods are STING and CLIQUE.

STING (Statistical Information Grid): - It is used mainly with numerical values. It is a grid-based multi resolution clustering technique which is computed the numerical attribute and store in a rectangular cell. The quality of clustering produced by this method is directly related to the granularity of the bottom most layers, approaching the result of DBSCAN as granularity reaches zero.

CLIQUE (Clustering in Quest): - It was the first algorithm proposed for dimension growth subspace clustering in high dimensional space. CLIQUE is a subspace partitioning algorithm introduced in 1998.

E. Model-based methods

Model-based clustering methods are based on the assumption that data are generated by a mixture of underlying probability distributions, and they optimize the fit between the data and some mathematical model. the application of clustering algorithms to detect grouping information in real world applications in data mining is still a challenge, primarily due to the inefficiency of most existing clustering algorithms on coping with arbitrarily shaped distribution of data of extremely large and high dimensional datasets

4. COMPARISON OF K-MEANS & K-MEDOIDS

A partitioning method creates k partitions, called clusters, from given set of n data objects. Initially, each data objects are assigned to some of the partitions. An iterative relocation technique is used to improve the partitioning by moving objects from one group to anotherA distancemeasure is one of the feature space used to identify similarity or dissimilarity of patterns between data objects. Some of the well known methods are k-means, k-medoids; Partitioning around Medoids (PAM), Clustering Large Applications (CLARA) and Clustering Large Applications based upon Randomized Search (CLARANS).Out of these methods k-means and k-medoids are reviewed here and also similarity measure for both algorithms is carried out by distance measure. It is most common to calculate the dissimilarity between two patterns using a distance measure defined on the feature space.

A. Analysis of K-means clustering

K-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. K-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. All the data objects are placed in a cluster having centroid nearest (or most similar) to that data object. After processing all data objects, k-means, or centroids, are recalculated, and the entire process is repeated. All data objects are bound to the clusters based on the new centroids. In each iteration centroids change their location step by step. In other words, centroids move in each iteration. This process is continued until no any centroid move. As a result, k clusters are found representing a set of n data objects.

An algorithm for k-means method

1. The algorithm arbitrarily selects k points as the initial cluster centers ("means").
2. Each point in the dataset is assigned to the closed cluster, based upon the Euclidean distance between each point and each cluster center.
3. Each cluster center is recomputed as the average of the points in that cluster.
4. Steps 2 and 3 repeat until the clusters converge.

Convergence may be defined differently depending upon the implementation, but it normally means that either no observations change clusters when steps 2 and 3 are repeated or that the changes do not make a material difference in the definition of the clusters.

a. Description

Given a set of observations (x1, x2, …, xn), where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into k sets (k ≤ n) S = {S1, S2, …, Sk} so as to minimize the within cluster sum of squares (WCSS):

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

Is an indicator of the distance of the n data points from their respective cluster centers. The generalized version of k-means algorithm has been presented, which produces ellipse-shaped as well as ball-shaped clusters. It also gives correct clustering results without specifying the exact number of clusters.

b. Merits in K-means

Relatively scalable and efficient in processing large data sets; complexity is O (i k n), where i is the total number of iterations, k is the total number of clusters, and n is the total number of objects. Normally, k<<n and i<<n.

i. Easy to understand and implement.

ii. It takes less time to execute as compared to other techniques.

iii. It can handle with only categorical values.

a) Demerits of K-means

i. K-means has problems when clusters are of differing, sizes, densities, non-globular shapes.
ii. K-means has problems when the data contains outliers.
iii. Applicable only when the mean of a cluster is defined; not applicable to categorical data
iv. Unable to handle noisy data.
v. May terminate at local optimum.
vi. Result and total run time depends upon initial partition

b. Analysis of K-medoids method

K-medoids method overcomes this problem by using medoids to represent the cluster rather thancentroid. A medoid is the most centrally located data object in a cluster., k data objects are selected randomly as medoids to represent k cluster and remaining all data objects are placed in a cluster having medoid nearest (or most similar) to that data object. After processing all data objects, new medoid is determined which canrepresent cluster in a better way and the entire process is repeated. Again all data objects are bound to the clusters based on the new medoids. In each iteration, medoids change their location step by step. Or in other words, medoids move in each iteration. This process is continued until no any medoid move. As a result, k clusters are found representing a set of n data objects.

An algorithm for k- medoids method

1. Arbitrarily choose 'k' objects as the initial medoids;
2. Assign each remaining object to the cluster with the nearest medoid;
3. Randomly select a non-medoid object;
4. Compute the total cost of swapping old medoid object with newly selected non-medoid object.
5. If the total cost of swapping is less than zero, then perform that swap operation to form the new set of k- medoids.
6. Until no change

K-mode algorithm, which clusters the categorical data by replacing the means of clusters with modes, using new dissimilarity measures produces only spherical clusters. The problem is K-Medoids does not generate the same result with each run, because the resulting clusters depend on the initial random assignments. It is more robust than k-medoids in the presence of noise and outliers; however it's processing is more costly than the k-medoid method.

a) Merits in k- medoids

i. More robust than k-means in the presence of noise andoutliers; because a medoid is less influenced by outliersor other extreme values than a mean.

b) Demerits in k- medoids

i. Relatively more costly; complexity is O(i k (n-k)2), where i is the total number of iterations, is the total number of clusters, and n is the total number of objects.

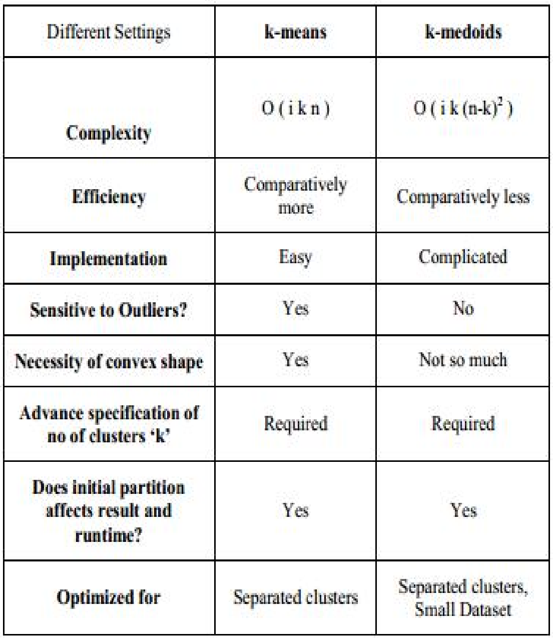
ii. Relatively not so much efficient.

iii. Need to specify k, the total number of clusters in advance.

iv. Result and total run time depends upon initial partition.

5. COMPARISON BETWEEN K-MEANS AND K- MEDOIDS

Comparison of both the clustering algorithm which highlights the realistic approach as well as desirable features of the algorithm which is useful in spatial database for different required clusters.



Different Settings	k-means	k-medoids
Complexity	O (i k n)	O (i k (n-k)^2)
Efficiency	Comparatively more	Comparatively less
Implementation	Easy	Complicated
Sensitive to Outliers?	Yes	No
Necessity of convex shape	Yes	Not so much
Advance specification of no of clusters 'k'	Required	Required
Does initial partition affects result and runtime?	Yes	Yes
Optimized for	Separated clusters	Separated clusters, Small Dataset

CONCLUSION

By the survey of cluster analysis aboveK means an k-medoids both data mining algorithm is most popularand efficient algorithm because it is very simpler in their operations, it is clear that there are two major drawbacks that influence the feasibility of cluster analysisin real world applications in data mining. The first one is the weakness of most existing automated clustering algorithms on dealing with arbitrarily shaped data distribution of the datasets. The second issue is that, the evaluation of the quality of clustering results by statistics-based methods is time consuming when the database is large, primarily due to the drawback of very high computational cost of statistics-based methods for assessing the consistency of cluster structure between the sampling subsets, some limitations in this algorithm makes it somewhat difficult. By removing these limitations we have used this algorithm in various field of our life such that in financial analysis, image segmentation and various fieldsin which we have extracting the information.

REFERENCES

[1]Jiawei Han and MichelineKamber (2006), Data Mining Concepts and Techniques, published by Morgan Kauffman,2nd ed.

[2] Lefait, G. and Kechadi, T, (2010) “Customer Segmentation Architecture Based on Clustering Techniques” Digital Society, ICDS’10, Fourth International Conference, 10-02-2010.

[3]J. McQueen, “Some methods for classification and analysis of multivariate observations”, Proc. of 5th Berkeley Symposium on Mathematics, Statistics and Probability, Volume 1, 1967, pp. 281-298.

[4] J. Han and M. Kamber, “Data Mining: Concepts and Techniques,” Morgan Kaufmann Publishers, 2001.

[5] Manish Verma, MaulySrivastava, NehaChack, AtulKumar Diswar, Nidhi Gupta,” A Comparative Studyof Various Clustering Algorithms in Data Mining, “International Journal of Engineering Research and Applications (IJERA), Vol. 2, Issue 3, pp. 1379-1384,2012.

[6] Jiawei Han and MichelineKamber, Jian Pei, B Data Mining: Concepts and Techniques, 3rd Edition,2007.

[7] P. Berkhin, “A Survey of Clustering Data Mining Techniques” Kogan, Jacob; Nicholas, Charles; Teboulle, Marc (Eds) Grouping

Multidimensional Data, Springer Press (2006) 25-72.

[8] A. K. Jain and R. C. Dubes, “Algorithms for Clustering Data”, Prentice Hall, 1988.

[9] A. Jain, M. N. Murty and P. J. Flynn, “Data Clustering: A Review”,ACM Computing Surveys, Volume 31(3), 1999, pp. 264-323.

[10] Hinneburg and D.A. Keim, “A General Approach to Clustering in Large Databases with Noise”, Knowledge and Information Systems (KAIS), Vol. 5, No. 4,pp. 387- 415, 2003.

[11] Yiu-Ming Cheung, “K*-means : A new generalized k-means clustering algorithm” Pattern Recognition Letters 24, pp. 2883-2893, 2003.

[12] Z. Huang, “Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values”, Data Mining and Knowledge Discovery 2, pp. 283–304,1998.

[13] Rob Short, Rod Gamache, John Vert and Mike Massa "Windows NT Clusters for Availability and Scalability" Microsoft Online Research Papers, Microsoft Corporation.

[14] Jim Gray "QqJim Gray’s NT Clusters Research Agenda" Microsoft Online Research Papers, Microsoft Corporation.

[15] Bruce Moxon "Defining Data Mining, The Hows and Whys of Data Mining, and How It Differs From Other Analytical Techniques" Online Addition of DBMS Data Warehouse Supplement, August 1996.

[16] Willet, Peter "Parallel Database Processing, Text Retrieval and Cluster Analyses" Pitman Publishing, London, 1990.