International Journal of
Medicinal Chemistry & Analysis
www.ijmca.com
e ISSN 2249 - 7587
Print ISSN 2249 - 7595

A SURVEY ON EFFICIENT CLUSTERING TECHNIQUES IN DATA MINING

V.R. Geetha[1*] and N. Jayaveeran[2]

[1]Research Scholar, Department of Computer Science, A.V.C.College of Arts & Science, Mannampandal, Tamilnadu, India.
[2]Head, P.G and Research Department of Computer Science, Khadir Mohideen College, Adirampattinam, Tamilnadu, India.

ABSTRACT

Data Mining, popularly known as Knowledge Discovery in databases (KDD) is the automated extraction of patterns representing knowledge implicitly stored in large databases. Data sets grow in size in part because they are increasingly being gathered by cheap and numerous information-sensing mobile devices, aerial (remote sensing), software logs, cameras, microphones, radio-frequency identification (RFID) readers, and wireless sensor networks. The World Wide Web as a Global Information System has flooded us with a tremendous amount of data and information. This explosive growth in stored data has generated an urgent need for new technologies and automated tools to assist us in transforming the data into useful information and knowledge. Data mining process helps to extract information from a data set and transform it into an understandable structure. Two common data mining techniques for finding hidden patterns in data are clustering and classification analyses. Clustering is an automated process to group related records together. Classification is similar to clustering in that it also segments customer records into distinct segments called classes. This paper provides survey on various clustering techniques and their role in Data mining.

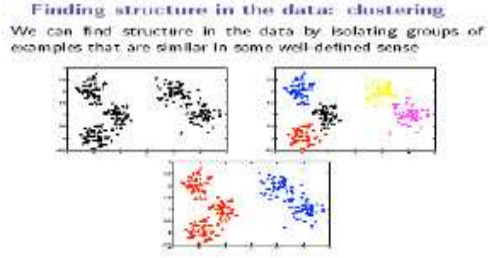
Keywords: clustering, Data Mining, Classification and Partition.

INTRODUCTION

Clustering is the process of assigning data sets into different groups so that data sets in same group having similar behavior as compared to data sets in other groups. Clustering plays an important role in data mining process and it's an unsupervised learning, where the class label of data sets is not previously known. A common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics. It can therefore be formulated as a multi-objective optimization problem. The appropriate clustering algorithm and parameter settings depend on the individual data set and intended use of the results. The most compact cluster means greater similarity within group and between groups gives best clustering result for data mining [1-4].

Clustering methods

There are many clustering methods available, and each of them may give a different grouping of a dataset. A good clustering method will produce high quality clusters with high intra-class similarity - Similar to one another within the same cluster low inter-class similarity - Dissimilar to the objects in other clusters. The quality of a clustering result depends on the similarity measure used by the method, its implementation and also measured by its ability to discover some or all of the hidden patterns.



Corresponding Author: - **V.R.Geetha** Email: geetharammesh@gmail.com

Traditionally clustering techniques are broadly divided in hierarchical and partitioning. For reader's convenience we provide a classification of clustering algorithms closely followed by this survey [5-7].

Clustering methods

There are many clustering methods available, and each of them may give a different grouping of a dataset. A good clustering method will produce high quality clusters with high intra-class similarity - Similar to one another within the same cluster low inter-class similarity - Dissimilar to the objects in other clusters. The quality of a clustering result depends on the similarity measure used by the method, its implementation and also measured by its ability to discover some or all of the hidden patterns. Traditionally clustering techniques are broadly divided in hierarchical and partitioning. For reader's convenience we provide a classification of clustering algorithms closely followed by this survey [8]

• Hierarchical Methods
- Agglomerative Algorithms
- Divisive Algorithms

• Partitioning Methods
- Probabilistic Clustering
- K-medoids Methods
- K-means Methods

• Density-Based Partitioning Methods
- Density-Based Connectivity Clustering
- Density Functions Clustering

• Grid-Based Methods

Hierarchical methods

Hierarchical algorithms produce a nested sequence (dendrogram) of clusters, with a single all-inclusive cluster at the top and singleton clusters of individual points at the bottom. It is also known as connectivity based clustering. This method merges smaller clusters to larger one and larger clusters to smaller ones.
The hierarchy can be formed in bottom-up (agglomerative) or top-down (divisive) fashion and need not necessarily be extended to the extremes. Once the desired number of clusters has been formed the merging or splitting stops.

The hierarchical agglomerative clustering methods are most commonly used. The construction of an hierarchical agglomerative classification can be achieved by the following general algorithm.

1. Find the two closest objects and merge them into a cluster
2. Find and merge the next two closest points, where a point is either a cluster of objects or an individual object.
3. If more than one cluster remains , return to step 2

A divisive clustering starts with a single cluster containing all data points and recursively splits the most appropriate cluster. The process continues until a stopping criterion (frequently, the requested number k of clusters) is achieved.

Advantages of hierarchical clustering include

• Ease of handling any form of similarity or distance
• Flexibility regarding the level of granularity
• Applicability to any attributes types

Disadvantages of hierarchical clustering are related to:

• Vagueness of termination criteria
• Most hierarchical algorithms do not revisit (intermediate) clusters once constructed.

Partitioning methods

Partitioning algorithms divides data into several subsets. They construct a partition of a dataset containing 'n' objects into a set of 'k' clusters, such that 'k' is less than or equal to 'n' and each cluster contains at least one element. This method is effective for small to medium sized datasets.

Probabilistic Clustering

In the probabilistic approach, data is considered to be a sample independently drawn from a mixture model of several probability distributions. Some of the important features of probabilistic clustering are

- It can be modified to handle points that are recodes of complex structure
- It can be stopped and resumed with consecutive batches of data, because clusters have representation totally independent from sets of points
- At any stage of the iterative process the intermediate mixture model can be used to assign points to clusters
- It results in easily interpretable cluster system [9].

K-Medoids Methods

K-medoid is the most appropriate data point within a cluster that represents it. Each cluster is represented by one of the objects in the cluster. K-medoids chooses datapoints as centers (medoids or exemplars) and works with an arbitrary matrix of distances between datapoints

It has two advantages: it presents no limitations on attributes types and the choice of medoids is dictated by the location of a predominant fraction of points inside a cluster and, therefore, it is insensitive
to the presence of outliers.

A medoid can be defined as the object of a cluster whose average dissimilarity to all the objects in the cluster is minimal. i.e. it is a most centrally located point in the cluster. The most common realisation of *k*-medoid clustering is the Partitioning Around Medoids (PAM) algorithm and is as follows

Partitioning Around Medoids (PAM) Algorithm

1. Initialize: randomly select (without replacement) *k* of the *n* data points as the medoids
2. Associate each data point to the closest medoid.
3. While the cost of the configuration decreases:

For each medoid *m*, for each non-medoid data point *o*:

1. Swap *m* and *o*, recompute the cost (sum of distances of points to their medoid) If the total cost of the configuration increased in the previous step, undo the swap [10].

Means clustering Algorithm

In this algorithm data is being divided into pre-specified clusters. Its main purpose is to define k centers ,one for every cluster. The algorithm works as follows:

- Divide the given data set into k partitions .Assign each partition the same number of elements.
- Find the mean of the elements of each partition.
- Check whether each element in the partition is closer distance wisw to the mean of current cluster or to the mean of another cluster. Perform this check for every element in each partition.
- Repeat process 3 until a set of stable clusters has been obtained.

Time complexity of k-means algorithm is O(nkt) where n is the total number of objects , k is the number of clusters and t is number of random numbers.Normally k<<n and t<<n. This method is related ,scalable and efficient in processing large data sets.

Density based partition method

Density based connectivity clustering: Density based methods use a concept of density to find arbitrarily shaped clusters sush as 'S' shaped and oval clusters.Examples of density based methods are DBSCAN AND OPTICS.

DBSCAN(Density based spatial clustering applications with Noise)

DBSCAN is a density based clustering method designed to find clusters of arbituary shape.Given a set D of objects , we can identify all core objects with respect to the given parameters E and Minpts number of points in its neighbourhood, we add those points to the cluster N.

- Initially all the objects in given data set D are marked as unvisited.

- DBSCAN randomly selects an unvisited object x.If x has atleast Minpts number of objects in its neighborhood, then a new cluster N is created otherwise it is marked as a noise point.
- If cluster is created , we iteratively visit each point y in this new cluster , if it is unvisited mark it as visited and if this point has Minpts number of points in its neighborhood, we add those points to the cluster N. If y is not a member of any cluster, it is added to the created cluster N.
- Repeat steps 2 and 3 until all objects are visited.

OPTICS (Ordering points to Identify Clustering Structure)

OPTICS is a variation of DBSCAN .It does not explicitly produce a data set clustering, instead gives us cluster ordering such that objects which are in a denser cluster are closer in a list. OPTICS stores two additional attributes ; Core Distance and Reachability. Distances which are used to derive the ordering such that clusters with higher density will be finished first. It has the same time complexity as the DBSCAN.

CONCLUSION

Choosing a clustering algorithm, however, can be a difficult task most of the algorithms assume some implicit structure in the data set. The problem; however is that usually we have little or no information regarding the structure. The worst case would be the one in which previous information about the data or the clusters is unknown, and the trial and error is the best option however there are many elements that are usually known and helpful in choosing an algorithm. One of the most important elements is the nature of the data and the nature of the desired cluster. Another issue to keep in mind is the kind of input and tools that the algorithm. For example some algorithm numerical inputs, some use categorical, some require a definition of a distance or similarity measures for the data. An additional issue for selecting an algorithm is correctly choosing the initial set of clusters. An adequate choice of clusters can strongly influence both the quality of and the time required to obtain a solution. It has been observed that the K-means clustering technique is most widely used because it produces better cluster results as compared to other clustering techniques and is also said to be computationally faster. Thus with the different type of clustering techniques and its use in many of the applications it has been adopted by many of the researchers in various fields to make the analysis of the data easier.

ACKNOWLEDGEMENT

Nil

CONFLICT OF INTEREST

No interest

REFERENCES

1. Parsons L, Haque P and Liu H. Subspace clustering for high dimensional data: a review. *ACM SIGKDD Explorations*, 6, 2009, 90-103.
2. XU R and Donald I, Wunsch C. clustering, A johnwiley & sons, Pub, 2008.

3. Fern XZ and Brodley CE. Random projection for high-dimensional data clustering. *A cluster ensemble approach*, 2008, 186.
4. Jain A, Murty M and Flynn P. Data clustering: A review. *ACM Computing Surveys*, 31, 2009, 264-323.
5. Jain N, *et al*. Data Mining Techniques: A survey paper. *International Journal of Research in Engineering and Technology*, 02(11), 201, 2321-7308.
6. Guha M, Mishra A, Motwani M. Clustering data streams, Theory and practice. *IEEE,* 15, 2008, 515-528.
7. Optics. Ordering Points To Identify the Clustering Structure, Proceedings of ACM SIGMOD international conference on Management of data, 2009.
8. Clustering by means of medoids, Statiscal Data Analysis Based on the L1-Norm and Related Methods 2000.
9. Saurabh A, Inderveer C. A Survey of Clustering Techniques for Big Data Analysis, *IEEE,* 2014, 59-65.
10. Saurkar A, *et al*. A Review paper on various Data Mining Techniques. *International Journal of Advance Research in Computer Science and Software Engineering*, 4(4), 2009, 98-101.