Survey on Clustering Techniques of Data Mining

J.AROCKIA JEYANTHI
Assistant Professor,Department of Computer Science,
A.P.C.Mahalaxmi College For Women,Thoothukudi
Email id:arockiajeyanthi@gmail.com
Mobile no.:7708984425

ABSTRACT - The goal of this survey is to provide a comprehensive review of different clustering techniques in data mining. Data mining refers to extracting useful information from vast amounts of data. It is the process of discovering interesting knowledge from large amounts of data stored either in databases, data warehouses, or other information repositories. An important technique in data analysis and data mining applications is Clustering.Cluster Analysis is an excellent data mining tool for a large and multivariate database. Clustering is a suitable example of unsupervised classification. Unsupervised means that clustering does not depend on predefined classes and training examples during classifying the data objects. Each group called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups. There are different types of clustering algorithms such as hierarchical, partitioning, grid, density based, model based, and constraint based algorithms. Hierarchical clustering is the connectivity based clustering. Partitioning is the centred based clustering; the value of k-mean is set. Density based clusters are defined as area of higher density then the remaining of the data set. Grid based clustering is the fastest processing time that typically depends on the size of the grid instead of the data. Model based clustering hypothesizes for each cluster and find the best fit of data to the given model. Constraint based clustering is performed by incorporation of user or application oriented constraints.In this survey paper, a review of different types of clustering techniques in data mining is done.

KEYWORDS : Data mining,Clustering, Types of Clustering, Requirements of Clustering, Applications of Clustering .

INTRODUCTION

Data mining is a new technology which is developing with database and artificial intelligence. It is a processing procedure of extracting credible, novel, effective and

understandable patterns from database. Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. Data mining consists of extract, transform, and load transaction data onto the data warehouse system, Data mining involves the anomaly detection, association rule learning, classification, regression, summarization and clustering.

Data mining concepts and methods can be applied in various fields like marketing, medicine, real estate, customer relationship management, engineering, web mining, etc. Clustering is the most interesting topics in data mining which aims atfinding intrinsic structures in data and find some meaningful subgroups for further analysis. It is a common technique for statistical data analysis, which is used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics. Thus a cluster could also be defined as the "methodology of organizing objects into groups whose members are similar in some way.

Data clustering is a process of putting similar data into groups. A clustering algorithm partitions a data set into several groups such that the similarity within a group is larger than among groups. Cluster Analysis, an automatic process to find similar objects from a database. It is a fundamental operation in data mining.

The aim of cluster analysis is that the objects in a group should be similar to one another and different from the objects in other groups. Clustering is much better when there is greater similarity within a group and greater the difference between the groups.

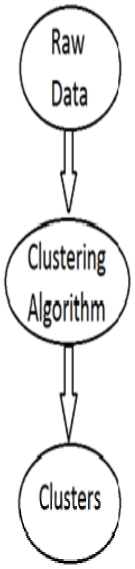


Fig 1: Stages of Clustering

TYPES OF CLUSTERS

1 .Well-Separated Cluster:

A cluster is a set of points such that any point that is in a cluster is closer
(or more similar) to every other point in the cluster than to any point which is not in the cluster .

2.Center-based Cluster:

A cluster is a set of objects such that an object in a clusteris closer (more similar) to the “centre” of a cluster, than to the centre of any other cluster.The centre of a cluster is often a centroid, the average of all the points in the cluster, or a medoid, the “most representative” point of a cluster.

3.Contiguous Cluster (Nearest Neighbour or Transitive Clustering):

A cluster is a set of points such that a point in a cluster is closer (or more similar) to one or more other points in the cluster than to any point not in the cluster.

4. Density-based Cluster:

A cluster is a dense region of points, which is separated by low-density regions, from other regions of high density. This definition is more often used when the clusters are irregular or intertwined, and when noise and outliers are present.

REQUIREMENTS OF CLUSTERING IN DATA MINING

Here are the typical requirementsof clustering in data mining:

- Scalability-We need highly scalable clustering algorithms to deal with large databases.
- Ability to deal with different kind of attributes-Algorithms should be capable to be applied on any kind of data such as interval based (numerical) data, categorical, binary data.
- Discovery of clusters with attribute shape-The clustering algorithm should be capable of detecting cluster of arbitrary shape. Theyshould not be bounded to only distance measures that tend to find spherical cluster of small size.
- High dimensionality-The clustering algorithm should not only be able to handle low-dimensional data but also the high dimensional space.
- Ability to deal with noisy data-Databases contain noisy, missing or erroneous data. Some algorithms are sensitive to such data and may lead to poor quality clusters.
- Interpretability-The clustering results should be interpretable,comprehensible and usable.

APPLICATIONS OF CLUSTERING TECHNIQUES

Clustering techniques are applicable in many fields, such as:

- Libraries:Book ordering.
- Marketing:Finding groups of customers with similar behaviour given a large database of customer datacontaining their properties and past buying records.
- Biology:Classification of plants and animals given their features.
- City-planning:Identifying groups of houses according to their house type, value and geographical location.
- Insurance:Identifying groups of motor insurance policy holders with a high average claim cost and identifying frauds.
- Spatial Data Analysis:Create thematic maps in GIS by clustering feature spaces
- WWW: Document classification, clustering weblog data to discover groups of similar access patterns.
- Earthquake studies:Clustering observed earthquake epicentres to identify dangerous zones.

CLASSIFICATION OF CLUSTERING

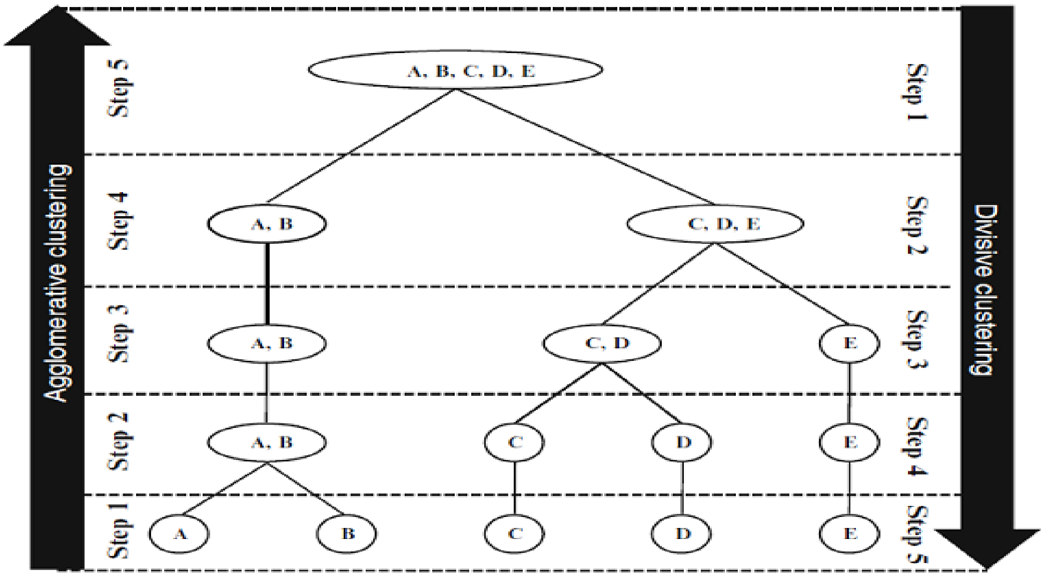
Clustering is the main task of Data Mining and it is done by the number of algorithms. The most commonly used algorithms in Clustering are Hierarchical, Partitioning, Density based, Grid based, Model Based and Constraint based algorithms.

1. Hierarchical Algorithms

Hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters. It is the connectivity based clustering algorithms. The hierarchical algorithms build clusters gradually. Hierarchical clustering generally fall into two types.

In hierarchical clustering, in single step, the data are not partitioned into a particular cluster. It takes a series of partitions, which may run from a single cluster containing all objects to "n" clusters each containing a single object.

Hierarchical Clustering is subdivided into agglomerative methods, which proceed by series of fusions of the "n" objects into groups, and divisive methods, which separate "n" objects successively into finer groupings.



Advantages of hierarchical clustering :

- Embedded flexibility regarding the level of granularity.
- Ease of handling any forms of similarity or distance.
- Applicability to any attributes type.

Disadvantages of hierarchical clustering:

- Vagueness of termination criteria.
- Most hierarchal algorithm do not revisit once constructed clusters with the purpose of improvement.

2. Partitioning Algorithms

Partitioning algorithms divide data into several subsets. The reason of dividing the data into several subsets is that checking all possible subset systems is computationally not feasible; there are certain greedy heuristics schemes that are used in the form of iterative optimization. Specifically, this means that different relocation schemes iteratively reassign points between the k clusters. Relocation algorithms gradually improve clusters.

There are many methods of partitioning clustering.They are K-means, Bisecting K Means Method, Medoids Method, PAM (Partitioning Around Medoids), CLARA (Clustering Large Applications) and the Probabilistic Clustering.

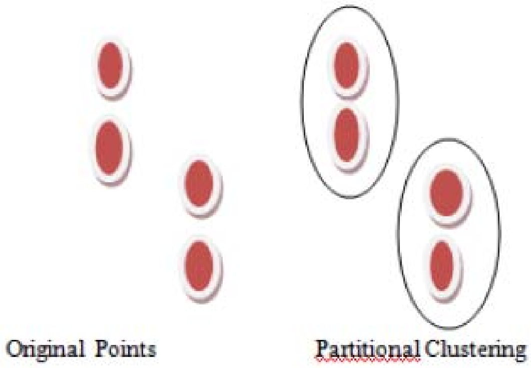


Figure 2: Partitioned Clustering

Methods of Primary Clustering

a.K-means

We are discussing the k-mean algorithm as: In k-means algorithm, a cluster is represented by its centroid, which is a mean (average pt.) of points within a cluster. This works efficiently only with numerical attributes. And it can be negatively affected by a single outlier. The k-means algorithm is the most popular clustering tool that is used in scientific and industrial applications. It is a method of cluster analysis which aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean.

K-means was proposed by MacQueen and is one of the most popular partition-based methods. It partitions the dataset into k disjoint subsets, where k is predetermined. The algorithm keeps adjusting the assignment of objects to the closest current cluster mean until no new assignments of objects to clusters can be made.

One Advantage of this algorithm is its simplicity. It also has several drawbacks. It is very difficult to specify number of clusters in advance. Since it works with squared distances, it also sensitive to outliers.

Another drawback is the centriods is not meaningful in most K-means has problems when clusters are of differing Sizes,Densities,Non-globular shapes and K-means has problems when the data contains outliers.

b.Bisecting K Means Method

This is an extension of K-Means method .The basic concept is as follows that to obtain k clusters split the set of all points into two clusters select one of them and split and repeat thisprocess until the K clusters have been produced.

c.Medoids Method

K-medoid is the most appropriate data point within a cluster that represents it. Representation by k-medoids has two advantages.First, it presents no limitations on attributes types,and, Second, the choice of medoids is dictated by the location of a predominant fraction of points inside a cluster and, therefore,it is lesser sensitive to the presence of outliers.

When medoids are selected, clusters are defined as subsets of points close to respective medoids, and the objective function is defined as the averaged distance or another dissimilarity measure between a point and its medoid.

d.PAM (Partitioning AroundMedoids)

PAM is iterative optimization that combines relocation of points between perspective clusters with re-nominating the points as potential medoids. The guiding principle for the process is the effect on an objective function.

e.CLARA (Clustering Large Applications)

CLARA uses several (five) samples, each with 40+2k points, which are each subjected to PAM. The whole dataset is assigned to resulting medoids, the objective function is computed, and the best system of medoids is retained.

3.Probabilistic Clustering

In the probabilistic approach, data is considered to be a sample independently drawn from a mixture model of several probability distributions.Themain assumption is that data points are generated by,

- first, randomly picking a model j with probability г
- second, by drawing a point x from a corresponding distribution. The area around the mean of each (unimodal) distribution constitutes a natural cluster.So we associate the cluster with the correspondingdistributionsparameters such as mean, variance, etc.Each data point carries not only its(observable) attributes, but also a (hidden) cluster ID (class in pattern recognition). Eachpoint x is assumed to belong to one and only one cluster, and we can estimate theprobabilities of the assignment.

Probabilistic clustering has some important features:

1. Itcan be modified to handle records of complex structure.

2. Itcan be stopped and resumed with consecutive batches of data, since clusters have representation totallydifferent fromsets of points.

3. Atany stage of iterative process the intermediate mixture model can be used to assign cases (on-line property).

4. Itresults in easily interpretable cluster system.

4.DensityBased Algorithms

Density based algorithms are capable of discovering clusters of arbitrary shapes. Also this provides a natural protection against outliers. These algorithms group objects according to specific density objective functions.Density is usually defined as the number of objects in a particular neighbourhood of a data objects. In these approaches a given cluster continues growing as long as the number of objects in the neighbourhood exceeds some parameter.

This type of clustering can be of two types

a.Density -Based Connectivity Clustering

In this clustering technique, density and connectivity both measured in terms of local distribution of nearest neighbours.So defined density-connectivity is a symmetric relation and all the points reachable from core objects can be factorized into maximal connected components serving as clusters.The points that are not connected to any core point are declared to be outliers (they are not covered by any cluster). The non-core points inside a cluster represent its boundary.Finally, core objects are internal points. Processing is independent of data ordering. So far, nothing requires any limitations on the dimension or attribute types.

b.Density Functions Clustering

In this density function is used to compute the density.Overall density is modeled as the sum of the density functions of all objects. Clusters are determined by density attractors,where density attractors are local maxima of the overall density function.The influence function can be an arbitrary one.

5.GridBased Clustering

These focus onspatial data i.e the data that model the geometric structure of objects in the space, their relationships, properties and operations. This technique quantize the data set

with a no of cells and then work with objects belonging to these cells. They do notrelocate points but ratter build several hierarchical levels of groups of objects. The merging ofgrids

and consequently clusters, does not depend on a distancemeasure .It is determined by a predefined parameter.

Advantages

•The major advantage of this method is fast processing time.

•It is dependent only on the number of cells in each dimension in the quantized space.

6.Model -Based Method

In this method a model is hypothesized for each cluster and find the best fit of data to the given model. Thismethod locatesthe clusters by clustering the density function. This reflects spatial distribution of the data points.This method also serve a way of automatically determining number of clusters based on standardstatisticstaking outlier or noise into account. It therefore yields robust clustering methods.

7.Constraint -Based Method

In this method ,the clustering is performed by incorporation of user or application oriented constraints. The constraint refers to the user expectation or the properties of desired clustering results. The constraint givesus the interactive way of communication with the clustering process. The constraint can be specified by the user or the application requirement.

CONCLUSIONS

Data Mining is a growing technology that combines techniques including statistical analysis, visualization, decision trees and neural network to explore large amount of data and discover relationship and patterns that shed light on business problems. Among other data mining techniques, clustering technique is of great use. It is an unsupervised learning method that attempts to find collections/groups of objects that are close to each other.

Clustering lies at the heart of data analysis and data mining applications.

Clustering is that technique of data mining which is used to extract the useful information from raw data. We can say that raw data is meaningless without the different clustering techniques.

REFERENCES

[1] PavelBerkhin, “A Survey of Clustering Data Mining Techniques”, pp.25-71, 2002.

[2] Wei-keng Liao, Ying Liu, AlokChoudhary, “A Grid-based Clustering Algorithm using Adaptive Mesh Refinement”, Appears inthe 7th Workshop on Mining Scientific and Engineering Datasets, pp.1-9, 2004.

[3] Cheng-Ru Lin, Chen, Ming-SyanSyan , “Combining Partitional and Hierarchical Algorithms for Robust and Efficient Data Clustering with Cohesion Self-Merging” IEEE Transactions On Knowledge And Data Engineering, Vol. 17, No. 2, pp.145-159,2005.

[4] OdedMaimon, LiorRokach, “DATA MINING AND KNOWLEDGE DISCOVERY HANDBOOK”, SpringerScience+BusinessMedia.Inc, pp.321-352, 2005.

[5] PradeepRai, Shubha Singh” A Survey of Clustering Techniques” International Journal of Computer Applications, October 2010.

[6] ZhengHua, Wang Zhenxing, Zhang Liancheng, WangQian, “Clustering Algorithm Based on Characteristics of Density Distribution” Advanced Computer Control (ICACC), 2010 2nd International Conference on National Digital Switching System Engineering & Technological R&D Center, vol2”, pp.431-435, 2010.

[7] MR ILANGO, Dr V MOHAN, “A Survey of Grid Based Clustering Algorithms”, International Journal of Engineering Science and Technology, pp.3441-3446, 2010.

[8] AminehAmini, Teh Ying Wah,, Mahmoud Reza Saybani, Saeed Reza AghabozorgiSahafYazdi, “A Study of Density –Gridbased Clustering Algorithms on Data Streams”,IEEE 8th International Conference on Fuzzy Systems and Knowledge Discovery, vol.3, pp.1652-1656, 2011.

[9] Guohua Lei, Xiang Yu, et.all, “An Incremental Clustering Algorithm Based on Grid”, IEEE 8th InternationalConference on Fuzzy Systems and Knowledge Discovery (FSKD), pp.1099-1103, 2011.

[10] Anoop Kumar Jain, Prof. Satyam Maheswari “Survey of Recent Clustering Techniques in Data Mining”, International Journal of Computer Science and Management Research, pp.72-78, 2012.

[11] M.Vijayalakshmi, M.Renuka Devi, “A Survey of Different Issue of Different clustering Algorithms Used in Large Data sets” , International Journal of Advanced Research in Computer Science and Software Engineering, pp.305-307, 2012.

[12] Ritu Sharma, M. AfsharAlam, Anita Rani , “K-Means Clustering in Spatial Data Mining using Weka Interface” , International Conference on Advances in Communication and Computing Technologies, pp. 26-30, 2012.

[13] PragatiShrivastava, Hitesh Gupta. “A Review of Density-Based clustering in Spatial Data”, International Journal of Advanced Computer Research (ISSN (print), pp.2249-7277, September-2012.