American International Journal of Research in Science, Technology, Engineering & Mathematics

Available online at http://www.iasir.net

ISSN (Print): 2328-3491, ISSN (Online): 2328-3580, ISSN (CD-ROM): 2328-3629

AIJRSTEM is a refereed, indexed, peer-reviewed, multidisciplinary and open access journal published by International Association of Scientific Innovation and Research (IASIR), USA

(An Association Unifying the Sciences, Engineering, and Applied Research)

Survey on Clustering Techniques of Data Mining

[1]Namrata S Gupta, [2]Bijendra **S.**Agrawal, [3]Rajkumar M. Chauhan

[1] Asst. Prof. Smt. BK Mehta IT Centre (BCA College), Palanpur, *Gujarat, INDIA*

[2] Principal, CCMS, Vadu, *Gujarat, INDIA*

[3]Foreman Instructor I.T.I. Amirgadh Ex. Asst. **P**rofessor BCA **C**ollege Palanpur, Gujarat, INDIA

Abstract*: 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.It divides data into groups of similar objects. Each group, called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups.Data mining has two types of tasks: Predictive and the descriptive. 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, Classification.*

I. Background

Bill Palace defines data mining (sometimes called data or knowledge discovery) as "*the process of analyzing data from different perspectives and summarizing it into useful information--information that can be used to increase revenue, cuts costs, or both. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.*" Data mining is one of the best ways to illustrate the difference between data and information: data mining transforms data into information. Data mining consists of extract, transform, and load transaction data onto the data warehouse system, Store and manage the data in a multidimensional database system, Provide data access to business analysts and information technology professionals, Analyze the data by application software, Present the data in a useful format, such as a graph or table. Data mining is a multi-step process. It requires accessing and preparing data for a data mining algorithm, mining the data, analyzing results and taking appropriate action. The accessed data can be stored in one or more operational databases, a data warehouse or a flat file. Data Mining is a four step: Assemble data, Apply data mining tools on datasets, Interpretation and evaluation of result, Result application.

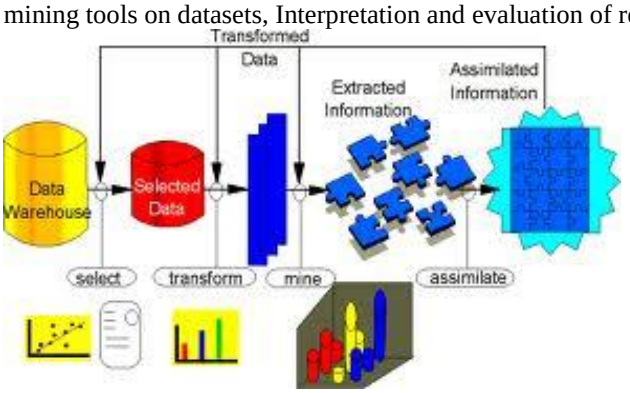


Figure 1: Steps of Data Mining Process

A. Data Mining Approaches

In data mining the data is mined using two learning approaches i.e. supervised learning or unsupervised clustering. Supervised Learning In this training data includes both the input and the desired results. These methods are fast and accurate. The correct results are known and are given in inputs to the model during the

learning process. Supervised models are neural network, Multilayer Perception, Decision trees. Unsupervised Learning The model is not provided with the correct results during the training. It can be used to cluster the input data in classes on the basis of their statistical properties only. Unsupervised models are different types of clustering, distances and normalization, k-means, self organizing maps.

B. Data Mining Tasks

Data mining tasks are generally divided into two major categories:

Predictive task The goal of this task is to predict the value of one particular attribute, based on values of other attributes. The attributes that is used for making the prediction is named as independent variable. The other value which is to be predicted is known as the Target or dependent value.

Descriptive task The purpose of this task is surmise underlying relations in data .In descriptive task of data mining, values are independent in nature and it frequently require post-processing to validate results.

Data mining involves the anomaly detection, association rule learning, classification, regression, summarization and clustering. In this paper, clustering analysis is done.

II. Introduction

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. Clustering algorithms are used extensively not only to organize and categorize data, but are also useful for data compression and model construction. A good clustering algorithm is able to identity clusters irrespective of their shapes. Other requirements of clustering algorithms are scalability, ability to deal with noisy data, insensitivity to the order of input records, etc.

A. Requirements of Clustering in Data Mining

Here are the typical requirements of 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 detect cluster of arbitrary shape. They should 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.

From a machine learning perspective clusters correspond to hidden patterns, the search for clusters is unsupervised learning, and the resulting system represents a data concept. From a practical perspective clustering plays an outstanding role in data mining applications such as scientific data exploration, information retrieval and text mining, spatial database applications, Web analysis, CRM, marketing, medical diagnostics, computational biology, and many others. Presenting data by fewer clusters necessarily loses certain fine details (loss in data compression), but achieves simplification. It represents many data objects by few clusters, and hence, it models data by its clusters. Clustering is often one of the first steps in data mining analysis. It identifies groups of related records that can be used as a starting point for exploring further relationships. Clustering is a data mining (machine learning) technique used to place data elements into related groups without advance knowledge of the group definitions. Clustering techniques fall into a group of undirected data mining tools. The goal of undirected data mining is to discover structure in the data as a whole. In general, there are two types of attributes associated with input data in clustering algorithms, i.e., numerical attributes, and categorical attributes. Numerical attributes are those with a finite or infinite number of ordered values, such as the height of a person or the x-coordinate of a point on a 2D domain. On the other hand, categorical attributes are those with finite unordered values, such as the occupation or the blood type of a person. Many different clustering techniques have been defined in order to solve the problem from different perspective, i.e. partition based clustering, density based clustering, hierarchical methods and grid-based methods etc.

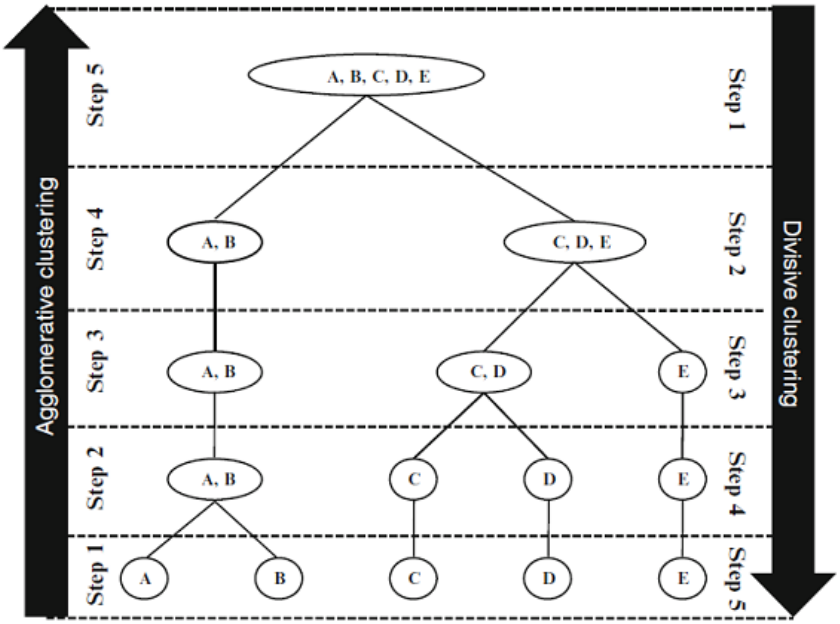
IV. Classification of Clustering

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.

A. 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.



A.1 Advantages of hierarchical clustering

1. Embedded flexibility regarding the level of granularity.
2. Ease of handling any forms of similarity or distance. 3. Applicability to any attributes type.

4.1.2 Disadvantages of hierarchical clustering

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

B. 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 are used in the form of iterative optimization. Specifically, this means different relocation schemes that iteratively reassign points between the k clusters. Relocation algorithms gradually improve clusters.

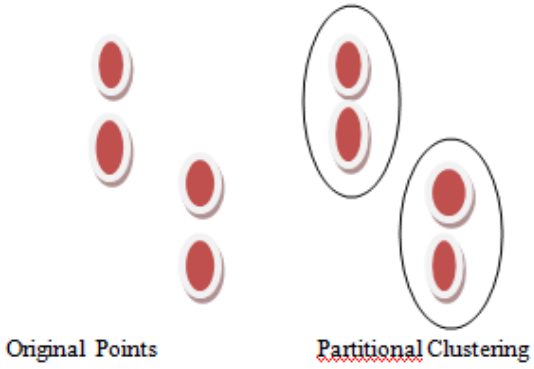


Figure 2: Partitioned Clustering

There are many methods of partitioning clustering; they are k-mean, Bisecting K Means Method, Medoids Method, PAM (Partitioning Around Medoids), CLARA (Clustering LARge Applications) and the Probabilistic Clustering. 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. The basic algorithm is very simple

1. Select K points as initial centroids.
2. Repeat.
3. Form K clusters by assigning each point to its closest centroid.
4. Re compute the centroid of each cluster until centroid does not change.

The *k*-means algorithm has the following important properties:

1. It is efficient in processing large data sets.
2. It often terminates at a local optimum.
3. It works only on numeric values.
4. The clusters have convex shapes.

C. Density-Based Clustering

In density-based clustering, clusters are defined as areas of higher density than the remaining of the data set. Objects in these sparse areas - that are required to separate clusters - are usually considered to be noise and border points. There are two major approaches for density-based methods. The first approach pins density to a training data point and is reviewed in the sub-section Density-Based Connectivity. 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. Representative algorithms include DBSCAN, GDBSCAN, OPTICS, and DBCLASD. The second approach pins density to a point in the attribute space and is explained in the sub-section Density Functions. In this, density function is used to compute the density. Overall density is modelled 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. It includes the algorithm DENCLUE. Density Based Spatial Clustering of Applications Noise DBSCAN (Density Based Spatial Clustering of Applications with Noise) algorithm is a density based clustering algorithm. In this algorithm the regions grow with sufficiently high density are known as clusters. The Eps and the Minpts are the two parameters of the DBSCAN. The basic idea of DBSCAN algorithm is that for each object of a cluster, the neighbourhood of a given radius (*Eps*) has to contain at least a minimum number of objects (*MinPts*).The clustering quality of DBSCAN algorithm strongly depend on the parameters does not depend upon the database. The parameters are set by users which will consider in the computational of clusters. The users have to select the parameters properly to get the better results the reason is that the same database with different parameters; the algorithm can produce different results. However, DBSCAN algorithm uses global parameters, which are not suitable for discovering clusters with different densities, without considering different possible density, only using a given possible density of any clusters, when the densities of clusters are totally separated.

D. Grid Based Algorithms

Grid-based clustering where the data space is quantized into finite number of cells which form the grid structure and perform clustering on the grids. Grid based clustering maps the infinite number of data records in data streams to finite numbers of grids. Grid based clustering is the fastest processing time that typically depends on the size of the grid instead of the data. The grid based methods use the single uniform grid mesh to partition the entire problem domain into cells and the data objects located within a cell are represented by the cell using a set of statistical attributes from the objects. These algorithms have a fast processing time, because they go through the data set once to compute the statistical values for the grids and the performance of clustering depends only on the size of the grids which is usually much less than the data objects. The grid-based clustering algorithms are STING, Wave Cluster, and CLIQUE. All these methods use a uniform grid mesh to cover the whole problem. For the problems with highly irregular data distributions, the resolution of the grid mesh must be too fine to obtain a good clustering quality. A finer mesh can result in the mesh size close to or even exceed the size of the data objects, which can significant increase the computation load for clustering. Adaptive Mesh Refinement Adaptive Mesh Refinement (AMR) is a type of multi resolution algorithm. This algorithm achieves high resolution in localized regions of dynamic, multidimensional numerical simulations. This is successfully applied to model large scale scientific applications in a range of disciplines, such as computational fluid dynamics, astrophysics, meteorological simulations, structural dynamics, magnetic, and thermal dynamics. Basically, it can place very high resolution grids precisely where the high computational cost requires. The adaptability of the algorithm allows simulating multi resolution that are out of reach with methods using a global uniform fine grid. The AMR clustering algorithm firstly creates different resolution grids based on the density. After that grids comprise a hierarchy tree that represents the problem domain as nested structured grids of increasing resolution. The algorithm considers each leaf as the center of an individual cluster and recursively assigns the membership for the data objects located in the parent nodes until the root node is reached. The AMR clustering algorithm can detect the nested clusters at different levels of resolutions by using the hierarchical tree. As the AMR algorithm is grid density based algorithm so it also shares the common characteristics of all grid-based methods. AMR algorithm has a fast processing time. It has the ability to separate from the noise. The

order of input data is insensitive. AMR is a technique that starts with a coarse uniform grid covering the entire computational volume and automatically refines certain regions by adding finer sub grids. From the connected parent grid cells, the new child grids are created whose attributes, density for instance, exceed given thresholds.

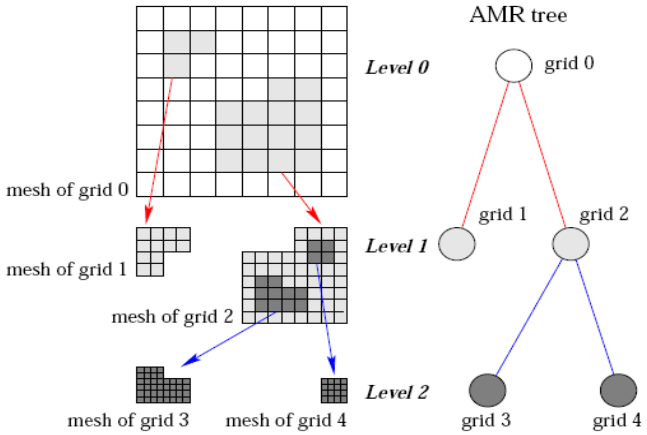


Figure 3: A 2-dimensional AMR example with 2 levels of refinement. A finer resolution mesh is applied each time a sub grid is created.

Advantage

- 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.

E. Model-Based Methods

In this method a model is hypothesize for each cluster and find the best fit of data to the given model. This method locates the 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 standard statistics, taking outlier or noise into account. It therefore yield robust clustering methods.

F. 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 gives us the interactive way of communication with the clustering process. The constraint can be specified by the user or the application requirement.

V. Conclusions

The overall goal of the data mining process is to extract information from a large data set and transform it into an understandable form for further use. Clustering is important in data analysis and data mining applications. It is the task of grouping a set of objects so that objects in the same group are more similar to each other than to those in other groups (clusters). Clustering can be done by the different no. of algorithms such as hierarchical, partitioning, grid and density based algorithms. Hierarchical clustering is the connectivity based clustering. Partitioning is the centroid 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. The grid based methods use the single uniform grid mesh to partition the entire problem domain into cells.

References

[1] Pavel Berkhin, "A Survey of Clustering Data Mining Techniques", pp.25-71, 2002.

[2] Wei-keng Liao, Ying Liu, Alok Choudhary, "A Grid-based Clustering Algorithm using Adaptive Mesh Refinement", Appears in the 7th Workshop on Mining Scientific and Engineering Datasets, pp.1-9, 2004.

[3] Cheng-Ru Lin, Chen, Ming-Syan Syan , "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] Oded Maimon, Lior Rokach, "DATA MINING AND KNOWLEDGE DISCOVERY HANDBOOK", Springer Science+Business Media.Inc, pp.321-352, 2005.

[5] Pradeep Rai, Shubha Singh" A Survey of Clustering Techniques" International Journal of Computer Applications, October 2010.

[6] Zheng Hua, Wang Zhenxing, Zhang Liancheng, Wang Qian, "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] Amineh Amini, Teh Ying Wah,, Mahmoud Reza Saybani, Saeed Reza Aghabozorgi Sahaf Yazdi, "A Study of Density-Grid based 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 International Conference 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. Afshar Alam, Anita Rani , "K-Means Clustering in Spatial Data Mining using Weka Interface" , International Conference on Advances in Communication and Computing Technologies (ICACACT Proceedings published by International Journal of Computer Applications® (IJCA), pp. 26-30, 2012.

13. Pragati Shrivastava, Hitesh Gupta. "A Review of Density-Based clustering in Spatial Data", International Journal of Advanced Computer Research (ISSN (print), pp.2249-7277, September-2012.

14. Gholamreza Esfandani, Mohsen Sayyadi, Amin Namadchian, "GDCLU: a new Grid-Density based CLUstring algorithm", IEEE 13th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, pp.102-107, 2012.