CLUSTERING TECHNIQUES IN DATA MINING

Mohd Javaid[1],Mrs.Anu Sharma[2], Mrs.Rolly Gupta[3]
1Student, , College of Computing Sciences & Information Technology, TMU, Moradabad

2Assistant Professor, , College of Computing Sciences & Information Technology, TMU, Moradabad

3Assistant professor, , College of Computing Sciences & Information Technology, TMU, Moradabad

[1]Javaidm347@gmail.com
[2]anu.computers@tmu.ac.in
[3]rolly.computers@tmu.ac.in

Abstract- **The main aim of data mining process is to extract meaningful information from large databases and convert it in to an understandable form for further use.Clustering is a process of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups.This paper presents a comprehensive review of major clustering techniques in data mining such as hierarchical clustering and partitioning clustering.**

\
***Keywords-Data* Mining, Clustering, Hierarchical Clustering, Partitioning Clustering.**

[1] INTRODUCTION

Data mining is define as extracting information for from large set of data.In other words,data mining is the procedure of mining knowledge from data.the information or knowledge extracted so can be used for any of the following applications:

A. Market analysis and management.
B. Corporate analysis and risk management.
C. Fraud detection.
D. Customer retaining.
E. Production control.
F. Science exploration.

Major Data Mining Tasks

Data mining consists of four classes of tasks:

1) .Clustering: Clustering is the atomic learning technique in which division of the data elements in to groups of similar objects takes place.

2) Classification:It is the supervised learning technique which is used to map the data in to predefine classes.

3) Regression:It is the statistical technique which is used to develop a mathematical formula that fits the dataset.

4) Association Rule Mining:It is the data mining technique which is used to identify relationship from a set of items in a database.

Cluster is a group of objects that belongs to the same class. In other words, similar objects are grouped in one cluster and dissimilar objects are grouped in another cluster. Clustering is the process of making a group of abstract objects into classes of similar objects. A cluster of data objects can be treated as one group. The main advantage of clustering over classification is that, it is adaptable to changes and helps single out useful features that distinguish different groups.

[2] LITERATURE REVIEW

According to analyst the process of data mining goes through several cycles. Data mining process is getting the paramount result and new important information from a huge searching on databases. There are two main data mining goals, predictive and descriptive. Predictive model is to retrieve information from the known system model and descriptive model is to retrieve the new, remarkable information. Data mining can be done through different methods like statistical method, machine learning, etc. The statistical methods are mathematical models, where as machine learning is an algorithm. Statistical

model gives emphasis to structure of data and machine learning is to analyze and manage collection of data [15]. Data mining provides some major properties:

A. Extract, transform, and load transaction data onto the data warehouse system.
B. Store and manage the data in a multidimensional database system.
A. Provides facility for data access by business analysts and information technology professionals.
B. Analyze the data & Present the data in a useful format, such as a graph or table by application software.

According to Varun and Nisha the integration of clustering techniques gives more accurate and robust result than applying either of them alone. In their study, they have shown that clustering is an unsupervised learning method and it creates classes by partitioning number of clusters according to their instances.

S.NO.	Authors Name	Title Name	Findings
1	Wu Yuntian	Based on Machine Learning of Data Mining to Further Explore	Retrieve the hidden and not apparent information.
2	Neelamadhab Padhy, Dr. Pragnyaban Mishra, Rasmita Panigrahi	The Survey of Data Mining Applications And Feature Scope	Produce the precise information.
3	Areej Shhab, Gongde Guo	A Study on Applications of Machine Learning Techniques in Data Mining	Higher capability for solving data mining problems.
4	Jyothi Bellary, Bhargavi Peyakunta	Hybrid Machine Learning Approach In Data Mining	Proposed a hybrid approach
5	Neagu C.D., Guo G., Trundle P.R., Cronin M.T.D.	A comparative study of machine learning algorithms applied to predictive toxicology data mining	A specific characteristic of a data set would be preferred for improving the models correctness.
6	Yogendra Kumar, Upendra	An efficient intrusion detection based on decision tree classifier using feature reduction	Instructions and compare performances.
7	Yugal kumar, G. Sahoo	Analysis of bayes, neural network and tree classifier of classification technique in data mining using WEKA	In their result they found that the presentation of the J48 technique is reasonably better than other techniques.
8	Aaditya Desai	Analysis of Machine Learning Algorithms using WEKA	Multilayer perceptron provides better outcome.

[3] CLUSTERING TECHNIQUE

Clustering is the most fundamental technique in Data mining. The goal of clustering is to divide the data elements into groups of similar objects, where each group is referred to as

a cluster, consisting of objects that are similar to one another and dissimilar to objects of other groups. Clustering is efficiently used in several exploratory pattern analysis, machine learning, data mining and bioinformatics problems. The basic problem in the context of clustering is to group a given assortment of unlabelled patterns into significant clusters. Cluster Analysis is the automatic process of grouping data into different groups, so that the data in each group share similar trends and pattern. The clusters which are formed are defined as the organization of datasets into homogeneous and/or well separated groups with respect to distance or equivalently similarity measure. The following diagram shows the stages in a clustering process.

Phases of Clustering Process

A. Principles of Clustering

The formed clusters need to follow and satisfy the following principles of clustering.

1) *Homogeneity* elements of the same cluster are maximally closeto each other.
2) *Separation:* data elements in separate clusters are maximally far apart from each other.

A good clustering method will produce high quality clusters with high intra-cluster similarity and low inter-cluster similarity. The quality of a result produced by clustering depends on both the similarity measure used by method and its implementation. The quality of a cluster produced by clustering method is also measured by its ability to discover some or all of the hidden patterns.

B. Types of Clusters

1) *Well- separated Clusters:* A cluster is a set of data points such that any point in the cluster is nearer (or more similar) to every other pint in the cluster than to any point not in the cluster.
2) *Center-based Clusters:* A cluster is a set of objects such that an object in a cluster is closer or more similar to the "center" of a cluster, than to center of any other cluster.
3) *Contiguous Clusters:* A cluster is a set of data 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 Clusters:* A cluster is a dense region of points, which is separated by low density regions from other regions of high density.
5) *Conceptual Clusters:* finds clusters that share some common property or represent a particular concept.
6) *Clusters Described by an Objective Function* finding clusters that minimize or maximize an objective function and enumerating all possible ways of dividing the points into clusters and evaluating goodness of each potential set of clusters.

C. DIFFERENT CLUSTERING TECHNIQUES

Various clustering approaches can be broadly classified into two main groups: hierarchical and partitioning methods. Furthermore, according to Han and Kamber (2001) the clustering methods are categorized into three additional categories which are: density-based methods, model-based and grid based methods [4].

1) Hierarchical Method

Hierarchical clustering is also known as Connectivity based clustering. Hierarchical clustering is a method of cluster analysis that constructs the clusters or groups by recursively partitioning the instances in either a top-down or bottom-up approach. Hierarchical clustering algorithm builds a cluster hierarchy or a tree of clusters.

2) Partitioning Method

In the partitioning methods, the general outcome is a set of N clusters, where each object belongs to one cluster. Each cluster or group may be represented by a centroid or a cluster

representative. Partitional Clustering is also known as iterative relocation algorithm and centroid based clustering.

[4] CONCLUSION

Clustering is a descriptive task in data mining. Clustering is used to divide the data into groups of similar objects. This paper presents a brief study of various clustering techniques such as hierarchical and partitioning clustering. Hierarchical clustering is referred as connectivity based clustering. Partitioning method is referred as centroid based clustering.The clustering technique also plays a significant role in data analysis and data mining applications.

ACKNOWLEDGEMENT

The sincere thanks to our guide for his help and assistance towards the successful completion of this review paper.

REFERENCES

[1] W. Frawley, G. P. Shapiro, and C. Matheus, "Knowledge Discovery in Databases: An Overview," *AI Magazine*, pp. 213-228, 1992.

[2] F. Usama, G. P. Shapiro, and P. Smyth (1996), "From Data Mining to Knowledge Discovery in Databases," Available: http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf.

[3] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data Clustering: A Review," ACM Computing Surveys, vol. 31, No. 3 pp. 264–323, Sept. 1999.

[4] O. Maimon and L. Rokach, *Data Mining and Knowledge Discovery Handbook*, Springer Science +Business Media.Inc, pp. 321–352, 2005.

[5] P. Rai and S. Singh, "A Survey of Clustering Techniques," *International Journal of Computer Applications*, Oct. 2010.

[6] P. Berkhin, "A Survey of Clustering Data Mining Techniques," pp. 25–71, 2002.

[7] A. K. Jain and S. Maheswari, "Survey of Recent Clustering Techniques in Data Mining," *International Journal of Computer Science and Management Research (IJCSMR)*, pp. 72–78, 2012.

[8] N. Rajalingam and K. Ranjini, "Hierarchical Clustering Algorithm-A Comparative Study," *International Journal of Computer Applications (IJCA)*, vol. 19, Issue 3, pp. 42– 46, April. 2011.

[9] C. R. Lin and M. S. Chen, "Combining Partitional and Hierarchical Clustering Algorithms for Robust and Efficient Data Clustering with Cohesion Self-Merging," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, No. 02, Issue 5, pp. 1041–4347, Feb. 2005.

[10] A. K. Mann and N. Kaur, "Survey Paper on Clustering Techniques," *International Journal of Science, Engineering and Technology Research (IJSETR)*, vol. 02, Issue 4, pp. 2278–779, April. 2013.