AN OVERVIEW OF DIFFERENT CLUSTERING METHODS IN DATA MINING

MANDAN NARESH
M. Tech Asst. Prof (Cse)
Holy Mary Institute of
Technology and Science
mandan.naresh3@gmail.com

RAVIKUMAR BANOTH
M. Tech (Ph.D) Asst. Prof (Cse)
Holy Mary Institute of
Technology and Science
brk.1820@gmail.com

Dr. A.P.SIVA KUMAR
M. Tech, Ph.D Asst. Prof (Cse)
JNTU, Anathapur
sivakumar.ap@gmail.com

ABSTRACT

Clustering is defined as partition of data into groups of similar objects and it is an unsupervised learning approach. Clustering is one of the Data Mining task for grouping data objects based on distance between two objects or points, such that points within a particular group have similar distinctiveness, while points in different groups are dissimilar. Conventional clustering algorithms that use distances between points for clustering are not appropriate for Boolean and categorical attributes. Clustering is said to be a collection of objects. It is used in various applications in the real world. Such as text mining, voice mining, image processing, web mining and so on. It is important in real world in certain fields. So this paper goal is to provide review of various clustering methods in data mining.

Keywords: *Data Mining, Clustering, Types of Clustering*

INTRODUCTION:

The improvement of Information tools has generated huge amount of databases and enormous data in various areas. The study in databases and information technology has known rise to an approach to store and operate this precious data for further decision making. Data mining is a procedure of extraction of useful information and patterns from huge data. It is also called as knowledge discovery process, knowledge mining from data, knowledge extraction or data pattern analysis.

Information mining additionally called learning revelation in databases is the way toward finding helpful examples and connections in expansive volumes of information. Information mining is worried about the examination of information and the utilization of programming strategies for finding covered up and sudden examples and connections in sets of information. The concentration of information mining is to discover the data that is covered up and startling. Information mining is a multi-step process. It requires getting to and planning information for an information mining calculation, mining the information, breaking down outcomes and making fitting move. The information can be put away in at least one operational database, an information distribution center or a level record. In information mining the information is mined utilizing two learning approaches i.e. directed learning or

unsupervised learning?

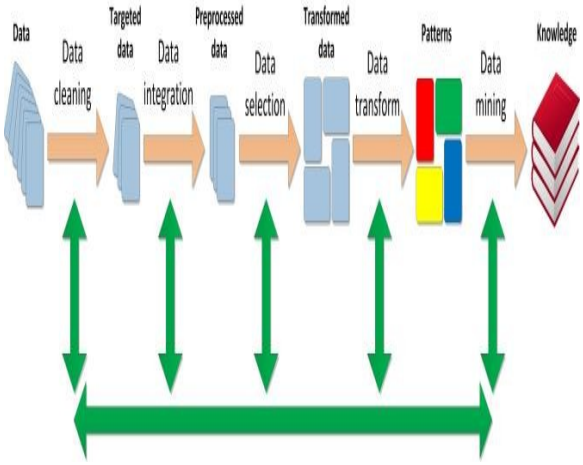

Fig: Knowledge discovery process

Data mining is a logical process that is used to search through large amount of data in order to find useful data. The goal of this technique is to find patterns that were previously unknown. Once these patterns are found they can further be used to make certain decisions for development of their businesses.

Data Mining Techniques

Different algorithms and method like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor method etc., are used for knowledge discovery from databases.

A cluster is a group of data objects or points that are alike to one another within the same cluster and are unrelated to the objects in other clusters. 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. Clustering is 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 technique used to place data elements into related groups without advance knowledge of the group definitions.

General Types of Clusters

Well-separated Clusters

A cluster is a set of points so that any point in a cluster is nearest (or more similar) to every other point in the cluster as compared to any other point that is not in the cluster.

Center-based Clusters

A cluster is a set of objects such that an object in a cluster is nearest (more similar) to the "center" of a cluster, than to the center of any other cluster. The center of a cluster is often a centroid.

Contiguous Clusters

A cluster is a set of points so that a point in a cluster is nearest (or more similar) to one or more other points in the cluster as compared to any point that is not in the cluster.

Density-based Clusters

A cluster is a dense region of points, which is separated by according to the low-density regions, from other regions that is of high density.

Shared Property or Conceptual Clusters

Finds clusters that share some common property or represent a particular concept.

Methods of Clustering

Partitioning methods:

Partitioning methods 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. The partitioning methods generally result in a set of M clusters, each object belonging to one cluster. Each cluster may be represented by a centroid or a cluster representative; this is some sort of summary description of all the objects contained in a cluster.

Hierarchical methods

Hierarchical clustering (also called hierarchical cluster analysis or HCA) is a technique of cluster analysis which seeks to build a hierarchy of clusters also known as dendrogram. Hierarchical clustering involves creating clusters that have a predetermined ordering from top to bottom. It is based on the middle suggestion of objects individual more related to nearby objects than to objects beyond away. So it can be concluded, these algorithms connect "objects" to form "clusters" on the basis of distance based clustering. And this hierarchical method dividing into two types one is Agglomerative Method and Divisive method

Density-based methods

In density-based clustering, clusters are defined as areas of higher density than the enduring of the data set. Objects in these sparse areas - that are required to separate clusters - are usually considered to be noise and border points. Density-based algorithms are proficient of discovering clusters of arbitrary shapes. Likewise this gives a characteristic insurance against exceptions. These calculations bunch objects as indicated by particular thickness target functions.Density is normally characterized as the quantity of items in a specific neighborhood of an information objects. In these methodologies a given bunch keeps developing as long as the quantity of items in the area surpasses some parameter.

Grid-based methods

Grid based bunching where the information space is quantized into limited number of cells which shape the lattice structure and perform grouping on the networks. Matrix based grouping maps the endless number of information records in information streams to limited quantities of networks. Matrix based grouping is the quickest preparing time that normally relies upon the measure of the lattice rather than the information.

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.

K-Means clustering

K-means partition the data points into K groups or cluster, where parameter k is a positive integer specified by the user. K-means starts with K centroids, and it iteratively performs the following steps

a) Assign each data instance to the cluster whose centroid is nearest to it.

b) Compute the new centroids of each cluster.

The K-Means Algorithm Process

The dataset is partitioned into K clusters and the data points are randomly assigned to the clusters resulting in clusters that have roughly the same number of data points.

For each data point:

1. Calculate the distance from the data point to each cluster.
2. If the data point is closest to its own cluster, leave it where it is. If the data point is not closest to its own cluster, move it into the closest cluster.
3. Repeat the above step until a complete pass through all the data points results in no data point moving from one cluster to another. At this point the clusters are stable and the clustering process ends.
4. The choice of initial partition can greatly affect the final clusters that result, in terms of inter-cluster and intracluster distances and cohesion.

Application Areas of Clustering

Clustering algorithms can be applied in many fields, for instance:

• Marketing: finding groups of customers with similar interests and behaviour given a large database of customer data containing their properties and past buying records.

• Medicine: IMRT segmentation, Analysis of antimicrobial activity, Medical imaging.

• Financial task: Forecasting stock market, currency exchange rate, bank bankruptcies, understanding and managing financial risk, trading futures, credit rating.

• Computer Science: Software evolution, Image segmentation, Anomaly detection.

• Biology: classification of plants and animals given their features, human genetic clustering, transcriptomics.

• Insurance: identifying groups of motor insurance policy holders with a high average claim cost; identifying frauds.

• City-planning: identifying groups of houses according to their house type, value and geographical location.

• Earthquake studies: clustering observed earthquake epicentres to identify dangerous zones.

• WWW: document classification; clustering web log data t

CONCLUSION

This paper deals with review of various kind of clustering algorithms. It first describes the data mining process which is the method of finding predictive information from an enormous amount of databases. Then it defines the clustering method which is the procedure of collection of the objects in groups whose members contain some kind of similarity. After that a detailed study of clustering algorithms and their comparison in different perceptions are examined.

REFERENCES

[1] Han, J., Kamber, M. 2012. Data Mining: Concepts and Techniques, 3rd ed, 443-491

[2] 2.Arockiam, L., S. S. Baskar, and L. Jeyasimman. 2012. Clustering Techniques in Data Mining.

[3] RenJingbiao and Yin Shaohong "Research and Improvement of Clustering Algorithm in Data Mining", 2010 2nd International Conference on Signal Processing Systems (ICSPS)

[4] M. Srinivas and C. Krishna Mohan, "Efficient Clustering Approach using Incremental and Hierarchical Clustering Methods", 2010 IEEE

[5] OdedMaimon, LiorRokach, "Data Mining and Knowledge Discovery Handbook", Springer Science+BusinessMedia.Inc, pp.321-352, 2005.

[6] AsmitaYadav, "A Survey Of Issues And Challenges Associated With Clustering Algorithms", International Journal for Science and Emerging Technologies with Latest Trends, July 2013.

[7] Amandeep Kaur Mann, "Survey Paper on Clustering Techniques", International Journal of Science, Engineering and Technology Research (IJSETR) Volume 2, Issue 4, April 2013.

[8] Aastha Joshi, "A Review: Comparative Study of Various Clustering Techniques in Data Mining", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 3, March 2013.

[9] Pradeep Rai, "A Survey of Clustering Techniques", International Journal of Computer Applications Volume 7– No.12, October 2010.

[10] Manish Verma, Mauly Srivastava, NehaChack, Atul Kumar Diswar, Nidhi Gupta," A Comparative Study of Various Clustering Algorithms in Data Mining," International Journal of Engineering Reserch and Applications (IJERA), Vol. 2, Issue 3, pp.1379-1384, 2012.

[11] 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.

[12]. S. Chakrabarti, B. Dom, R. Agrawal, and P. Raghavan. Using taxonomy, discriminants, and signatures for navigating in text databases. In Proceedings of the 23rd VLDB Conference, Athens, Greece, 1997.

[13]T. M. Cover and J. A. Thomas. Elements of Information Theory. John Wiley & Sons, New York, USA, 1991.