iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** XII **Month of publication:** December 2017

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor:6.887
Volume 5 Issue XII December 2017- Available at www.ijraset.com

A Review on Clustering Techniques for Data Mining

Gurmeet Kaur[1], Pooja Rana[2]
[1,2]*Assistant Professor, Department of Computer Applications,C T Institute Shahpur Campus, jalandhar, India*

Abstract: Now a day's data is growing in a sense of size and variety. How to fetch information from the databases is a important concern. Decision making out of information is challenging these days. Many techniques have been developed for extracting. One of techniques is data clustering. In this paper, a review of several clustering techniques that are being used in Data Mining is presented. In clustering we use cluster of same type of data and current data mining clustering techniques.
Keywords: database, techniques, clustering, data mining etc.

I. INTRODUCTION

For any organization collection of data for future reference is very important. Along with that having tools that can be used to check coming requirements, trends and problem of existing data is necessary. Data clustering is a technique that can check many data sets and each data set can contain different data types. Each data set is having different size and size of dataset is depends upon the count of objects, dimensions and different data types. In case of data clustering method internal structure of data is not known.

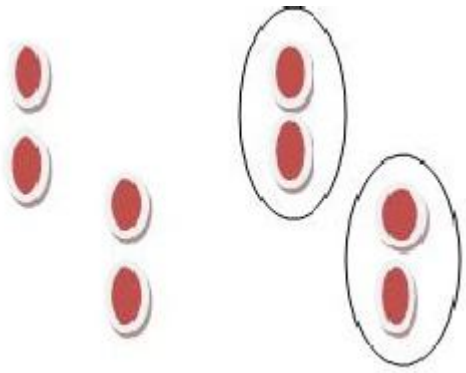
A. Data Mining Techniques

Classification, Predication, Association, neural networks are various techniques of Data Mining.

B. Clustering

Clustering is a fundamental operation in data mining. Clustering can be understood as recognition of alike classes of objects. It can discover overall division pattern and correlations among data attributes. Clustering methods have been projected and they can be mostly classified into four categories like partitioning methods, hierarchical methods, density-based methods and grid-based methods.

C. Partitioning Methods



1) Originalcluster
2) Partitioned cluster

Partitioning method simply moving instances from one cluster to another for relocation. It starts moving from initial Partitioning. In this method user will predetermined number of cluster .

To achieve global optimality in partitioned-based clustering, an complete enumeration process of all possible partitions is required. Because this is not feasible, certain greedy heuristics are used in the form of iterative optimization. Namely, a relocation method iteratively relocates points between the k clusters. The following subsections present various types of partitioning methods.

A partitioning-based clustering algorithms combinatorial optimization algorithm is the most popular class of clustering algorithms. They are also known as iterative relocation algorithms. These algorithms minimize A given clustering criterion is minimized using these algorithms by iteratively relocating data points between clusters until an optimal partition is attained. In a basic iterative algorithm, such as

K-means algorithms which are used as a solution to clustering problem. In this algorithm, a given dataset is classified into a fixed number of clusters (assume k clusters). The main idea is to define the centroids of each cluster. The centroids of each cluster are placed as far as possible from each other. In the next step, each point belonging to a given data set is taken and associated to the nearest centroid. When there exist no point and early grouping is done. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more.

Because the number of data points in any data set is always finite and, thereby, also the number of distinct partitions is finite, the problem of local minima could be avoided by using exhaustive search methods. Graph-Theoretic, Error Minimization Algorithms are used for partitioning method.

D. Hierarchical Agglomerative (divisive) Methods

As the name specifies this method is to create a hierarchy of clusters i.e. like tree structure. Now the point is that we can create a hierarchy by using bottom up or top down approach. So hierarchy clustering method is divided into two parts Agglomerative and Divisive. Agglomerative is a bottom up approach. The process of clustering is continued till certain condition is satisfied.

In a "bottom up" approach: each examination starts in its own cluster, and pairs of clusters are combined as one moves up the hierarchy. Divisive: This is a "top down" approach: all examination start in one cluster and splits are performed recursively as one moves down the hierarchy.

The top down approach works as twice as faster than bottom up approach.

To increase the efficiency of agglomerative clustering algorithm as well as to make it suit for large data, we require the cloud computing virtualized environment]. Virtualization is a key technology used in data centers to optimize resource.

E. Density Based Methods

Density based clustering algorithm is one of the prime methods for clustering in data mining. In this clusters are formed on the bases of density.

Density based clustering method is useful because it can find clusters in random shapes and it can handle noisy data efficiently. It is called as one scan algorithm because raw data is examined only once. In density based clustering clusters are defined as areas of higher density than remainder of the data sets. One cluster is separated from other clusters by lower density regions.

This is easy to understand and it does not limit itself to the shapes of clusters. Some of the existing density based algorithms namely DBSCAN, VDBSCAN, DVBSCAN, ST-DBSCAN and

F. Grid-Based Methods

Grid based method is different from other as performs action on cell rather than data points. This method is having more calculation efficiency as compare to other methods that are traditionally used.

In fact, most of the grid-clustering algorithms achieve a time complexity. All clustering operations are performed in a gridded data space. Grid-based methods are most widely used in comparison to the other conventional models as they have high computational efficiency. The major difference between grid-based and other clustering methods is that all the clustering operations are performed on the segmented data space, instead of the original data objects. In grid-based clustering methods, it has to determine beforehand a proper size of the grid structure which is a major difficulty. Larger grid size can be managed by combining two or more clusters into a single cluster. In case of smaller grid size, a cluster may be divided into several sub-clusters. So, finding the suitable size of grid is a challenging issue in grid clustering methods. Other problem is with the data of clusters having uneven densities and arbitrary shapes in case of which a global density threshold cannot result the clusters with less densities. This problem is known as the locality of cluster. The third problem is how to select a merging condition to form efficient clusters.

G. Model-based methods

These methods attempt to optimize the fit between the given data and some mathematical models. Unlike usual clustering, which identifies groups of objects, model-based clustering methods also find characteristic descriptions for each group, where each group

represents a concept or class. The most frequently used induction methods are decision trees and neural networks. Model-based Clustering MLE (maximum likelihood estimation) is used in model-based clustering method to find the parameter inside the probability model. Since the probability function is a mixture summation of a couple of probability function, it makes the traditional method infeasible to find the maximum value. Latent variable technique is used here, relocation algorithm such as EM and Gibbs sampling are among the most popular.

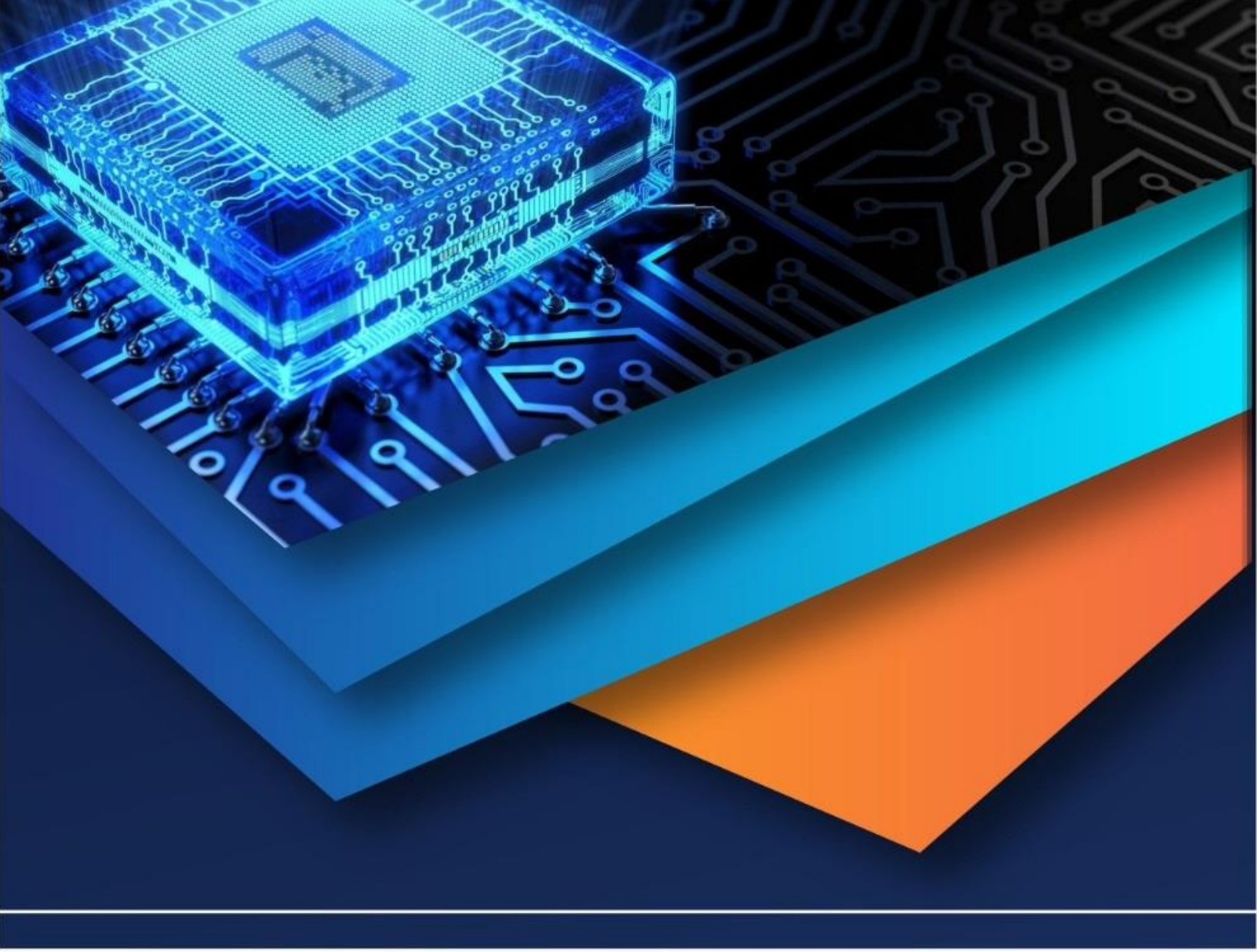
The criterion to split one data set into several data sets is to make the variance between the clusters maximum and inside the clusters minimum.

II. CONCLUSION

It has been concluded in this that data mining has importance regarding finding the patterns, forecasting, discovery of knowledge etc., in different business domains. Data mining techniques and algorithms such as classification, clustering etc., helps in finding the patterns to decide upon the future trends in businesses to grow. Data mining has variety of applications almost in every domain of industry where the data is generated that's why data mining is considered one of the most important frontiers in database and information systems and one of the most promising interdisciplinary developments in Information Technology.

REFERENCES

[1] Crisp-DM 1.0 Step by step Data Mining guide from http://www.crisp-dm.org/CRISPWP-0800.pdf.
[2] Dr. Gary Parker, Data Mining: Modules in emerging fields, CD-ROM, Vol 7, 2004.
[3] "Hierarchical clustering" in http://hierarchicalclustering .co.tv\
[4] Data Mining and Analytics Resources. [Online]. Available: http://www.kdnuggets.com/gpspubs/aimagkdd-overview-1996-Fayya d.pdf
[5] DamodarReddy Edla and Prasanta K. Jana"A Grid Clustering Algorithm Using Cluster Boundaries" IEEE World Congress on Information and Communication Technologies 2012.
[6] D. Jixue; , "Data Mining of Time Series Based on Wave Cluster," Information Technology and Applications, 2009. IFITA '09. International Forum on, vol.1, no., pp.697-699, 15-17 May 2009.
[7] E. W. M. Ma and T. W. S. Chow, "A new shifting grid clustering algorithm," Pattern Recognition, vol. 37, pp. 503-514, 2004
[8] H. Darong and W. Peng, "Grid-based DBSCAN Algorithm with Referential Parameters," Proc. International Conference on Applied Physics and Industrial Engineering (ICAPIE-2012), Physics Procedia, vol. 24(B), pp. 1166-1170, 2012.
[9] Y. Zhao and J. Song, GDILC: A Grid-based Density-Isoline Clustering Algorithm," Proc. International Conferences on Info-tech and Info-net (ICII-2001), vol. 3, pp. 140-145, October 29-November 1, 2001.
[10] J. Han, M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, San Francisco, 2001. M. S. ALDENDERFER AND R. K. BLASHFIELD, Cluster analysis, Sage Publications, London, England, 1984.
[11] H.Ding, " Querying and Mining of Time Series Data: experimental comparison of representations and distance measures". Proceedings of the VLDB Endowment VLDB Endowment Hompage archive Volume 1 Issue 2, August 2008, pp 1542-1551.
[12] H. Kremer; S. Gunnemann; T. Seidl,; , "Detecting Climate Change in Multivariate Time Series Data by Novel Clustering and Cluster Tracing Techniques," Data Mining Workshops (ICDMW), 2010 IEEE International Conference on , vol., no., pp.96-97, 13-13 Dec. 2010
[13] T.W. Liao, Clustering of time series data—survey, Pattern Recognition 38 (2005), pp. 1857–1874.
[14] Y. Yang and K. Chen," Time-Series Clustering via RPCL Network ensemble with different representations", IEEE Transactions On Systems, Man, And Cyzbernetics—Part C: Applications And Reviews, Vol. No. 2, March 2011, pp. 190-199.
[15] V. Niennattrakul; C.A. Ratanamahatana , "On Clustering Multimedia Time Series Data Using K-Means andDynamic Time Warping," Multimedia and Ubiquitous Engineering, 2007. MUE '07. International Conference on , vol., no., pp.733-738, 26-28 April 2007.
[16] P. SobheBidari ; R. Manshaei ; T. Lohrasebi; A. Feizi; M.A. Malboobi,; J. Alirezaie; , "Time series gene expression data clustering and pattern extraction in Arabidopsis thaliana phosphatase-encoding genes,"BioInformatics and BioEngineering, 2008. BIBE 2008. 8th IEEE International Conference on , vol., no., pp.1-6, 8-10 Oct. 2008.
[17] J. Yin; D. Zhou; Q.-Q. Xie; , "A Clustering Algorithm for Time Series Data," Parallel and Distributed Computing, Applications and Technologies, 2006.
[18] Lefait, G. and Kechadi, T, (2010) "Customer Segmentation Architecture Based on ClusteringTechniques" Digital Society, ICDS'10, Fourth International Conference, 10-02-2010.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY