A Survey of Clustring Techniques In Data Mining

Manjeet
M.Tech(CSE) Scholar
Dept. of CSE
R.P.S.G.O.I
Mohindergargh-12309 (India)
sheoranmanjeet71@gmail.com

MeghaYadav
Assistant Professor
Deptt. of CSE
R.P.S.G.O.I.
Mohindergargh-12309
meghayadavas@gmail.com

Abstract:Clustering is one of the imperativewatercourses in data mining useful for discovering groups and ascertaining interesting distributions in the underlying data. A clustering process partitions a data set into several groupsbased on the principle of maximizing the intra-class similarity and minimizing the inter-class similarity.
The clustering methods are broadly divided into hierarchical and partitioning clustering methods. In splitting clustering methods various panels are constructed and then evaluations of these partitions are performed by some criterion. Pure categorized clustering method suffers from its incapacity to perform regulation, once a merge or split decision has been executed, so some other procedures are merge with this method so as to produce better result of clustering.

Keyword: Clustering inData Mining, Hierarchical,CURE, BIRCH.

I.INTRODUCRION

Data Mining is the process of extracting hidden knowledge from large volumes of raw data. The importance of collecting data that reflect our business or scientific activities to achieve competitive advantage is widely recognized now. Powerful systems for collecting data and managing it in large databases are in place in all large and mid-range companies. However, the bottleneck of turning this data into our success is the difficulty of extracting knowledge about the system westudy from the collected data .
Modern computer data mining systems self learn from the previous history of the investigated system, formulating and testing hypotheses about the rules which this system obeys. When concise and valuable knowledge about the system of interest had been discovered, it can and should be incorporated into some decision support system which helps the manager to make wise and informed business decisions.

There are mainly two primary goals of data mining prediction and description. Prediction involves using some variables or fields in the data set to predict unknown or future values of other variables of interest. Description, on the other hand, focuses on finding patterns describing the data that can be interpreted by humans. Therefore it is possible to put data mining activities into one of the two categories:
Predictive data mining, which produces the model of the system described by the given data set, orDescriptive data mining, which produces new nontrivial information based on the available data set.On the predictive end of spectrum, the goal of data mining is to produce a model, expressed as an executable code, which can be used to perform classification, prediction, estimation or other similar tasks. On the other hand, descriptive end of spectrum, the goal is to gain understanding of the analyzed system by uncovering patterns and relationships in large data sets .

II Finding of Review

Recent Reacher state that few important methods and techniques of data mining that emphasis on clustering techniques.Data mining systems can be classified according tothe kinds of databases mined, the kinds of knowledge mined, the techniques used orthe applications..Data mining engine is ideally consists of a set of functional modules for tasks such ascharacterization, association, classification, cluster analysis, and evolution. On theother hand, pattern evaluation module typically employs certain measures andinteracts with the data mining modules so as to focus the search towards unknownpatterns [1].All the functional modules of data mining concepts like classification, prediction, association and clustering are explained by them with example that makes it easily understandable and very informative. Many new data mining methods, systems, and applications have been developed. [2].A briefsummary of recent KDD real-world applicationsis also provided by them. Here definitions of KDD and datamining are provided, and the general multistepKDD process is alsooutlined. This multistepprocess has the application of data-mining algorithmsas one particular step in the process. They discussed data-mining step in more detailin the context of specific data-mining algorithmsand their application. Real-world practical application issues are also outlined by the author [3]. A research paper on comparative study between data mining tools over some classification methods at international journals of advanced computer science and applications. This paper has conducted a comparative study between a number of some of the free available data mining and knowledge discovery tools and software packages. Results have showed that the performance of the tools for the classification task is affected by the kind of dataset used and by the way the classification algorithms were implemented within the toolkits. For the applicability issue, the WEKA toolkit has achieved the highest applicability followed by Orange, Tanagra, and KNIME respectively [6].

III Proposed Work

A Clustering in Data Mining

Clustering, in data mining, is useful for discovering groups and identifying interesting distributions in the underlying data. Clustering is the process of grouping the data into classes or clusters, so that objects within a cluster have high similarity in comparison to one another but are very dissimilar to objects in other clusters. The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters.

Properties of Clustering Methods

Clustering is a challenging field of research in which its potential applications cause their own special requirements. The following are typical requirements of clustering in data mining:

A. Ability to Deal with Different Types of Attributes

There exist various types of attribute like ratio, interval based or simple numeric values. All of these fall in the category of numeric representations. On the other hand, we also have nominal and ordinal attribute. An attribute is nominal if it successfully distinguishes between classes but does not have any inherit ranking and cannot be used for any arithmetic. For e.g. If color is an attribute and it has 3 values namely red, green, blue then we may assign 1-red, 2-green, 3-blue. This does not mean that red is given any priority or preference. Another type of attribute is ordinal and it implies ranking but cannot be used for any arithmetic calculation

B. Scalability

Many clustering algorithms work well on small data sets containing fewer than several hundred data objects; however, a large database may contain millions of objects. Clustering on a sample of a given large data set may lead to biased results. For this highly scalable clustering algorithms are needed.

C. Discovery of Clusters with Arbitrary Shape

Many clustering algorithms determine clusters based on Euclidean or Manhattan distance measures. Algorithms based on such distance measures tend to find spherical clusters with similar size and density. However, a cluster could be of any shape. It is important to develop algorithms that can detect clusters of arbitrary shape.

B. CURE

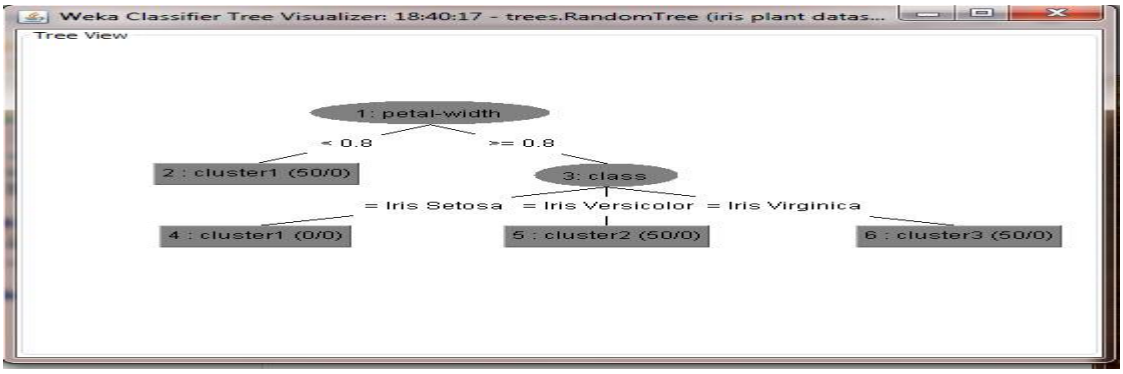
CURE is an agglomerative hierarchical clustering algorithm that adopts a middle ground between centroid-based and representative-object based approaches. Instead of using a single centroid or object to represent a cluster, a fixed number of representative points in space are chosen. The representative points of a cluster are generated by first selecting well-scattered objects for the cluster and then “shrinking” or moving them toward the cluster center by a specified fraction, or shrinking factor. At each step of the algorithm, the two clusters with closest pair of representative point are chosen. Having more than one representative point per cluster allows CURE to adjust well to the geometry of non spherical shapes [3].

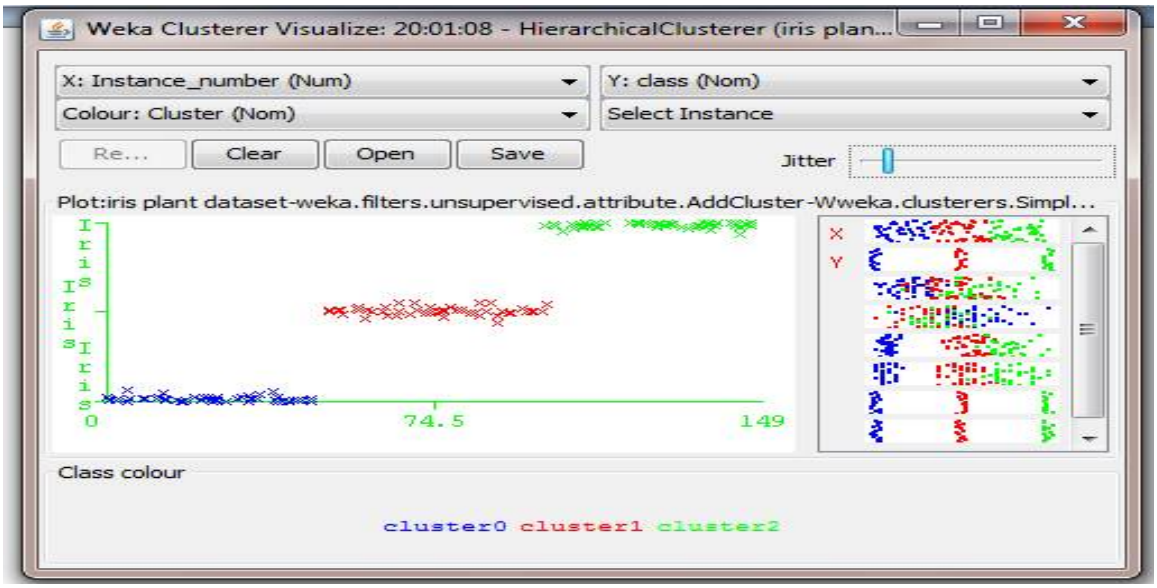
C. BIRCH

BIRCH is an agglomerative hierarchical clustering algorithm proposed by Charikar et al.. It is especially suitable for very large databases. This method has been designed so as tominimize the number of I/O operations. BIRCH incrementally and dynamically clustersincoming multi-dimensional metric data points to try to produce the best quality clustering withthe available

resources (i. e., available memory and time constraints). BIRCH can typically find agood clustering with a single scan of the data, and improve the quality further with a few additionalscans. BIRCH is also the first clustering algorithm proposed in the database area to handle "noise"(data points that are not part of the underlying pattern) effectively

V. Result:





VI Analysis

S.No.	Properties	BIRCH	CURE
1	Complexity	O (n2)	O (n2)
2	Input Parameter	Dataset is loaded in memory by building CF-tree	Random Sampling is done on the Dataset
3	Large Dataset Handling	Yes	Yes
4	Geometry	Shapes of cluster produced is almost spherical or convex of uniform size	Cluster produced can be of any arbitrary shape and in wide variance in sizes
5	Noise	Handle noise effectively	Comparatively less sensitive towards noise handling
6	Outlier Handling	Filtering of Outliers contained in dataset is done less effectively	Filtering of Outliers contained in dataset is done more effectively due to shrinking factor
7	Type of Data values	Numerical only	Numerical and Nominal both
8	Type of Model	Dynamic or incremental model	Static Model
9	Pre-clustering Phase	Cluster- Feature tree is formed	Partitioning is to be done on sample
10	Running time	Comparatively more	Partitioning improves running time by 50%
11	Data Input Order Sensitivity	Yes	No

Comparative Table of BIRCH and CURE Clustering Algorithm

VII. Conclusion

Hierarchical huddling is a method of gathering analysis which findsto build a hierarchy of clusters. The quality of a pure classified clustering method suffers from its inability to perform adjustment, once a amalgamate or split decision has been accomplished. This merge or split decision, if not well chosen at some step, may lead to some-what low quality clusters. One promising direction for civilising the clustering quality of hierarchical methods is to participate hierarchical clustering with other techniques for multiple phase clustering. These types of modified algorithm have been discussed in our paper in detail

References

[1] Arun K Pujari, "Data mining techniques", Universities Press (India) Pvt. Ltd, 2001.
[2] Bharat Chaudhari, Manan Parikh, "A Comparative Study Of Clustering Algorithms Using Weka Tools" International Journal Of Application Or Innovation In Engineering & Management (IJAIEM)) Volume 1, Issue 2, October 2012.
[3] Chris ding and Xiaofeng He, "Cluster Merging And Splitting In Hierarchical Clustering Algorithms" 2002.
[4] Daniel Fasulo, "An Analysis of Recent Work on Clustering Algorithm", June 1999.
[5] D.Pramodh Krishna, A.Senguttuvan&T.SwarnaLatha, "Clustering on Large Numeric Data Sets Using Hierarchical Approach: Birch" Global Journal of Computer Scienceand Technology, Volume 12, Issue 12, Version 1.0, Year 2012.
[6] Usama Fayyad, Gregory Piatetsky-Shapiro and Padhraic Smyth, "From Data Mining to Knowledge Discovery in Database" American Association for Artificial Intelligence, 0738-4602 1996.
[7] Fazli can. andEsen A. Ozkaharan, "Two Partitioning Type Clustering Algorithm" Journal of the American Society for Information Science, Volume 35, Issue 5, 268-276, 1984.
[8] G.Karypis, E.H.Han and V.Kumar, "CHAMELEON: Hierarchical Clustering Using Dynamic Modeling", IEEE Computer, Volume 32, 68-75, 1999.
[9] I.K. RavichandraRao, Professor of Indian Statistical Institute, Bangalore, "Data Mining And Clustering Techniques" DRTC Workshop on Semantic Web, Bangalore, December, 2003.
[10] Inderjit S. Dhillon, SubramanyamMallela, Rahul Kumar, "A Divisive Information-Theoretic Feature Clustering Algorithm for Text Classification", Journal of Machine Learning Research, 1265-1287, 2003.

[11] Israel Spiegler, "Representation and interpretation of Clustering Techniques" Tel Aviv University, July 2006.
[12] J.A.S. Almeida, L.M.S. Barbosa, A.A.C.C. Pais and S.J. Formosinho, "Improving Hierarchical Cluster Analysis: A New Method with Outlier Detection and Automatic Clustering", Chemometrics and Intelligent Laboratory Systems, Volume 87, Page no. 208-217, 2007.
.
[13]http://www.wikipedia.org.