A Comparative Study of Clustering Data Mining: Techniques and Research Challenges

Dhara Patel[1], Ruchi Modi[2], Ketan Sarvakar[3]

M.Tech Student, U.V Patel College of Engineering, Ganpat University, Kherva, Mehsana, Gujarat, India[1, 2]
Assistant Professor, U.V Patel College of Engineering, Ganpat University, Kherva,Mehsana,Gujarat, India[3]

Abstract- **Clustering data mining is the process of putting together meaning-full or use-full similar object into one group. It is a common technique for statistical data, machine learning, and computer science analysis. Clustering is a kind of unsupervised data mining technique which describes general working behavior, pattern extraction and extracts useful information from electricity price time series. In this paper we have studied the various clustering techniques. A tabular comparison of work done by various authors is presented. This paper reviews five types of clustering data mining techniques- Partitioning Clustering, Hierarchical Clustering, Grid based clustering, Model based clustering, and Density based clustering.**

Keywords: Clustering, Data mining

I. INTRODUCTION

Data mining refers to extracting or "mining" knowledge from large amounts of data [1]. Data mining is the process of discovering attractive information from large amounts of data stored either in databases, data warehouses, or other information repositories.

By performing data mining, interesting knowledge, regularities, or high-level information can be extracted from database and viewed or browse from different angles. The discovered knowledge can be applied to decision making, process control, information management, and query processing. The types of data mining modeling are show in fig1.

II. CLUSTERING DATA MINING

Clustering is the process of grouping a collection of objects into classes of similar objects. *Clustering* is the process of grouping the data into *clusters*, so that objects within a cluster have high similarity in comparison to one another but are very different to objects in other clusters [1]. Clustering has its roots in any areas, with data mining, statistics, biology, and machine learning.

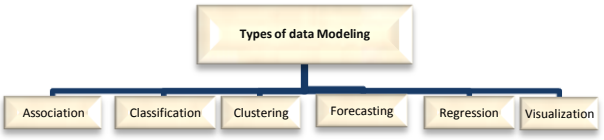

Fig1. Types of data modeling

Now a day, people come across a huge amount of the information. Then store or represent it as data [2]. One of the imperative means in dealing with these data is to group them into a set of clusters. Clustering involves creating groups of objects that are same, and those that are not same. The clustering problem lies in finding groups of same objects in the data.

The similarity between the objects is calculated by the use of a similarity function. It is mainly useful for organizing documents, to improve recovery and support browsing. Data Clustering is a method in which, information that is sensibly similar and physically stored together. To increase the efficiency in the database system, numbers of disk accesses are to be minimized. Clustering algorithms can be used in marketing, insurance, libraries, city-planning, biology, earthquakes, and www document classification.

III. RESEARCH CHALANGES

A. Telecommunication and Network Area

Process discovery is the learning task that works the construction of process models from event logs of information systems [14]. These event logs are large data sets that contain the process executions by registering what activity has taken place at a certain moment in time. The most difficult challenge for process discovery algorithms consists of tackling the problem of accurate and comprehensible knowledge discovery from highly flexible environments. Event logs from such flexible systems often contain a large variety of process executions which makes the application of process mining most interesting. Because of their inaccuracy and complexity, simply applying existing process discovery techniques will highly incomprehensible process models. So the challengeable task for real-world.

B. Stock Exchange

One of the decision problems in the financial domain is assortment management and asset selection [15]. Under the extremely competitive business environment, in order to face the complex market competitions, financial institutions try their best to make an ultimate policy for portfolio selection to optimize the investor returns. To maintain this data and policy, stock data clustering is needed for better efficient portfolio.

C. Pharmaceuticals

In history, information flow in the pharmaceutical industry was relatively simple and the application of technology was limited [16]. Now a day, progress into a more integrated world where technology has become an integral part of the business processes, the process of transfer of information has become more complicated. So using clustering, merging the drug usage and cost of medicines with the patient care records of doctors and hospitals helping firms to conduct universally leads for its new drugs.

IV. CLUSTERING TECHNIQUES

In general, the major clustering methods can be classified into the following categories.

A. Partitioning Method:-

Assume there are *n* objects in the original data set, partitioning methods split the original data set into *k* partitions.

1. Assign each object to the cluster associated with the closest centroid;
2. Compute the new position of each centroid by the mean value of the objects in a cluster.
3. Repeat Steps 2 and 3 until the means are fixed [3].

Ex: K-means, K-medoids, CLARA(Clustering Large Applications), CLARANS(Clustering Large Applications based upon Randomized Search) and PAM(Partitioning Around Medoids)

B. Hierarchical Method:-

This technique provides the tree relationship between clusters and produces a dendogram representing the nested grouping relationship among objects [3]. If the clustering hierarchy is formed from bottom up, at the start each data object is a cluster by itself, then small clusters are merged into bigger clusters at every level of the hierarchy. This type of hierarchical method is called agglomerative. The opposite process is called divisive [4].

Ex: BIRCH (balanced iterative reducing and clustering using hierarchies), CURE (Clustering Using REpresentatives), ROCK (Robust Clustering), AGNES(AGlomerative NESting)

C. Grid Based Method:-

In this technique measures the object space into a finite number of cells that form a grid structure on which all of the operations for clustering are performed [5]. It is based on clustering oriented query answering in multilevel grid structures. In upper level stores analysis of the information of it's after that level, therefore the grids make cells between the connected levels.

Ex: WaveCluster, STING (Statistical Information Grid), CLIQUE (Clustering in Quest)

D. Model Based Method:-

Model-based clustering technique is based on the best guess that data are generated by a blend of underlying probability distributions, and they optimize the fit among the data and some mathematical model.

Ex: EM(Expectation Maximization), SOM (Self Organizing feature Map)

E. Density Based Method:-

The density-based methods follow the expanding the cluster until a density threshold is reached. [6] For these methods a "neighborhood" has to be defined and the density must be calculated according to number of substance in the neighborhood.

Ex: DBSCAN (Density Based Spatial Clustering of Applications with Noise), DENCLUE (Density-based Clustering), OPTICS (Ordering Point to Identify Clustering Structure)

V. COMPARISON BETWEEN DIFFERENT TECHNIQUES

There are a many different techniques available in the field of data mining clustering. Algorithms are classified on the bases of performance, efficiency, complexity, scalability etc. We show a table representation of work; done by different authors. We have take parameters technique used for clustering its approaches and results.

Sr No	Author name	Year	Technique	Approach	Result
1	F. Mart´ınez-Alvarez, A. Troncoso1, J.C. Riquelme, and J.M. Riquelme	2007	K-means, EM	Extracting useful information of the prices time series by using clustering techniques to discover behavior's pattern to improve forecasting techniques.	K-means has a great degree of accuracy compare to EM so more suitable for daily prices classification [10].
2	Osama Abu Abbas	2008	Soft Clustering, Hierarchical clustering	Every algorithm is compared for performance, quality, and accuracy.	Hierarchical clustering is better than K-means and EM algorithm. In K-means and EM algorithm are very sensitive for noise [12].
3	T. Soni Madhulatha	2012	Hierarchical and partition based techniques	Determined number of cluster, size and type of datasets, types of software.	When applying a cluster analysis we are hypothesizing that the groups exist. But this assumption may be false or weak [8].
4	Gonzalo E. Paredes,Luis S. Vargas	2012	Partitioning methods	CirCle method is applied to numerical data used with both static data and time-series with same length.	The proposed method obtains an average of 81% of well-classified samples in all datasets. Also, as compared to other algorithms, CirCle made a better classification in 98.7% of the datasets as compared to the Model-Base [6].
5	Leonardo N. Ferreira, A. R. Pinto and Liang Zhao	2012	QK-mean	A hybrid clustering technique based on community detection in complex networks and traditional K-Means applied to detect better cluster and allow deploying the cluster head noad in large network.	QK-Means detect communities and sub-communities so lost message rate is decrease. WSL coverage is increased [9].
6	Osmar R . Za¨ıane, Andrew Foss, Chi-Hoon Lee, and Weinan	2012	Partitioning methods, hierarchical methods, density-based methods and grid-based methods	External and internal approaches use with the data can be validate and scalable.	For large dataset various cluster shapes including clusters within clusters and a great deal of noise [4].
7	S.R.Pande, Ms. S.S.Sambare, V.M.Thakre	2012	Hierarchical clustering, k-means, density-based, grid-based	Internal cluster validation excludes any information, and focuses on assessing clusters' quality based on the clustering data themselves.	A survey of several clustering techniques that are being used in Data Mining is presented [11].
8	Aastha Joshi,Rajneet Kaur	2013	K-mean, DBSCAN, STING, OPTICS	Different techniques are applied on numeric and categorical datasets to define distance function between data points.	K- mean and DBSCAN algorithm is better than Hierarchical Clustering Algorithm for categorical data [7].
9	P. IndiraPriya,Dr. D.K.Ghosh	2013	soft clustering, Hierarchical clustering	System proposed a different clustering algorithm for large and unlabeled datasets.	For large datasets clustering efficiency is ruined so need to improve time and scalability values [2].
10	Rahumath Beevi A, Remya R.	2014	Hard and soft clustering	Here to analyze huge dataset and their relationships require hard and fuzzy representations of different clustering.	Fuzzy based clustering approach provides significant performance [13].

Table1. Comparison between different techniques

CONCLUSION

Clustering is concern to cluster or categories the "similar" or "dissimilar" dataset into different groups Our survey in this paper focuses on the existing literature in the field of data mining clustering. From our analysis we have found that there is no single technique is applicable/dependable in all domains. All methods perform different role depending on type of data assign and type of application. But still from analysis, we have conclude that K-means method perform better than other method in many domain.

REFERENCES

[1] Data Mining Concepts and Techniques (Jiawei Han and Micheline Kamber)

[2] P. IndiraPriya,Dr. D.K.Ghosh " A Survey on Different Clustering Algorithms in Data Mining Technique" of *International Journal of Modern Engineering Research (IJMER) ,Vol.3, Issue.1, Jan-Feb. 2013 pp-267-274 ISSN: 2249-6645*

[3] S.R.Pande, Ms. S.S.Sambare , V.M.Thakre "Data Clustering Using Data Mining Techniques" of 2012

[4] Osmar R. Za¨iane, Andrew Foss, Chi-Hoon Lee, and Weinan "On Data Clustering Analysis: Scalability, Constraints and Validation" 2012

[5] Suman and Mrs.Pooja Mittal "Comparison and Analysis of Various Clustering Methods in Data mining On Education data set Using the weak tool" *of International Journal of Emerging Trends & Technology in Computer Science (IJETTCS) Volume 3, Issue 2, March – April 2014*

[6] Gonzalo E. Paredes,Luis S. Vargas "Circle-Clustering: A New Heuristic Partitioning Method for the Clustering Problem" *of WCCI 2012 IEEE World Congress on Computational Intelligence*

[7] Aastha Joshi, Rajneet Kaur " A Review: Comparative Study of Various Clustering Techniques in Data Mining" *of International Journal of Advanced Research in Computer Science and Software Engineering Volume 3, Issue 3, March 2013 ISSN: 2277 128X*

[8] T. Soni Madhulatha " An overview of clustering methods" *of IOSR Journal of Engineering Apr. 2012, Vol. 2(4) pp: 719-725*

[9] Leonardo N. Ferreira, A. R. Pinto and Liang Zhao "QK-Means: A Clustering Technique Based on Community Detection and K-Means for Deployment of Cluster Head Nodes" *of WCCI 2012 IEEE World Congress on Computational Intelligence June, 10-15, 2012*

[10] F. Mart´ınez-Alvarez, A. Troncoso1, J.C. Riquelme, and J.M. Riquelme "Partitioning-Clustering Techniques Applied to the Electricity Price Time Series" *H. Yin et al. (Eds.): IDEAL 2007, LNCS 4881, pp. 990–999, 2007. _c Springer-Verlag Berlin Heidelberg 2007*

[11] S.R.Pande, Ms. S.S.Sambare, V.M.Thakre "Data Clustering Using Data Mining Techniques" *of International Journal of Advanced Research in Computer and Communication Engineering Vol. 1, Issue 8, October 2012*

[12] Osama Abu Abbas" Comparisons between data clusterin algorithms" *of the international Arab Journal of Information Technology, Vol. 5, No. 3, July 2008*

[13] Rahumath Beevi A, Remya R. "A Comparison of Clustering Techniques in Data Mining"*of IJCAT International Journal of Computing and Technology, Volume 1, Issue 4, May 2014 ISSN : 2348 – 6090*

[14] Jochen De Weerdt, Seppe vanden Broucke, Jan Vanthienen, and Bart Baesens "Active Trace Clustering for Improved Process Discovery" *of ieee transactions on knowledge and data engineering, vol. 25, no. 12, december 2013*

[15] S.R. Nanda, B. Mahanty, M.K. Tiwari "Clustering Indian stock market data for portfolio management" *of Expert Systems with Applications @ 2010*

[16] Jayanthi Ranjan "Applications of data mining techniques in pharmaceutical industry" *of Journal of Theoretical and Applied Information Technology © 2005 - 2007 JATIT.*