International Journal of Contemporary Research in Computer Science and Technology (IJCRCST) ISSN: 2395-5325
Volume 3, Special Issue 3 (September '2017)
Proceeding of - International Conference on Recent Trends in Computational Life Science and Information Technology,
Conference held at Tiruppur Kumaran College for Women,Tirupur,Tamilnadu, India.

AN IMPRESSION OF CLUSTERING ALGORITHMS IN DATA MINING

R. Mallika,
M.Phil Research Scholar (FT),
PG & Research Department of Computer Science,
Vidyasagar College of Arts and Science,
Udumalpet,Taminadu,India.

V. Ramakrishnan,
Head & Assistant Professor,
Department of Computer Application,
Vidyasagar College of Arts and Science,
Udumalpet,Taminadu,India.

Abstract: Data mining is a logical process that is used to search through large amount of data in order to find useful data. The objective of this system 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. Various techniques like Classification, Clustering, Regression etc., are used for knowledge discovery from databases. Clustering analysis is one of the main analytical techniques in data mining. Clustering is the grouping together of similar data items into clusters. This paper discusses the various types of clustering algorithms and also analyse advantages and disadvantages of clustering algorithm based on various environments.

Keywords: Data Mining, Clustering Algorithm, Hierarchical, K-means clustering, Spectral.

I.INTRODUCTION

Data Mining is the impression of all methods and techniques, which permit to analyse large data sets to extract and discover previously unknown structures and relations viasuch huge data sets. These information is filtered, prepared and classified so thatit will be a valuable support for decisions and strategies.There are a number of data mining tasks such as classification, prediction, time-series analysis, association, clustering, summarization etc. All these tasks are either predictive data mining tasks or descriptive data mining tasks. A data mining system can execute one or more of the above specified tasks as part of data mining [1].Descriptive mining tasks characterize the general properties of the data in database. Predictive mining tasks perform inference on the current data in order to make predictions. Data mining consists of five major elements:

- Extract, transform, and load transaction data onto the data warehouse system.
- Store and manage the data in a multidimensional database system.
- Provide data access to business analysts and information technology professionals.
- Analyze the data by application software.
- Present the data in a useful format, such as a graph or table [2].

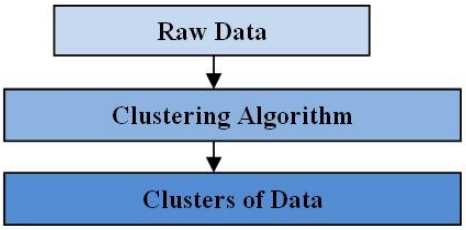

Figure 1: Steps in clusters

This paper includes many data mining and clustering techniques. Clustering is one of the most important research areas in the field of data mining. Clustering means creating groups of objects based on their features in such a way that the objects belonging to the same groups are similar and those belonging in different groups are dissimilar. Clustering is an unsupervised learning technique. This survey explores the behaviour of some of the clustering algorithms and their basic approaches [2].

II.DATA MINING TECHNIQUES

Data Mining (DM) is the process of finding the previously unknown information from the large amount of databases. The other terms carrying a similar meaning to data mining, is knowledge mining from databases, knowledge extraction, data or pattern analysis, data archaeology, and data dredging. It predicts future trends, behaviours and knowledge-driven decision. Data mining is a process of knowledge discovery. The KDD is an automated process of knowledge discovery from the original data. The KDD consists of many steps like data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and knowledge representation [3].

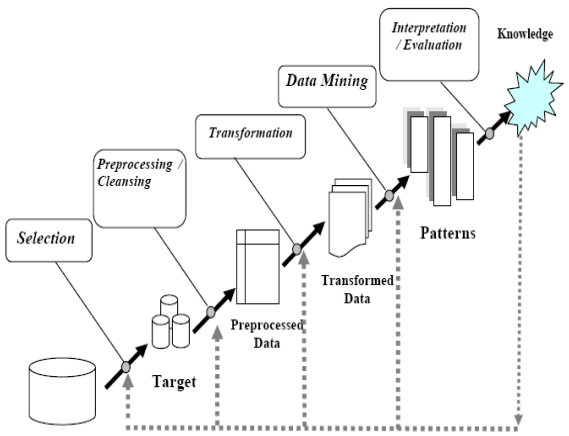

Figure 2: Stages in Data mining

Typically, a descriptive model is found through undirected data mining; i.e. a bottom-up approach where the data "speaks for itself". Undirected data mining finds patterns in the data set but leaves the interpretation of the patterns to the data miner. The purpose of a predictive model is to allow the data miner to predict an unknown (often future) value of a specific variable; the target variable. If the target value is one of a predefined number of discrete (class) labels, the data mining task is called classification. If the target variable is a real number, the task is regression [4].

The predictive model is thus created from given known values of variables, possibly including previous values of the target variable. The training data consists of pairs of measurements, each consisting of an input vector x (i) with a corresponding target value y(i). The predictive model is an estimation of the function y=f(x; q) able to predict a value y, given an input vector of measured values x and a set of estimated parameters q for the model f. The process of finding the best q values is the core of the data mining technique [4].

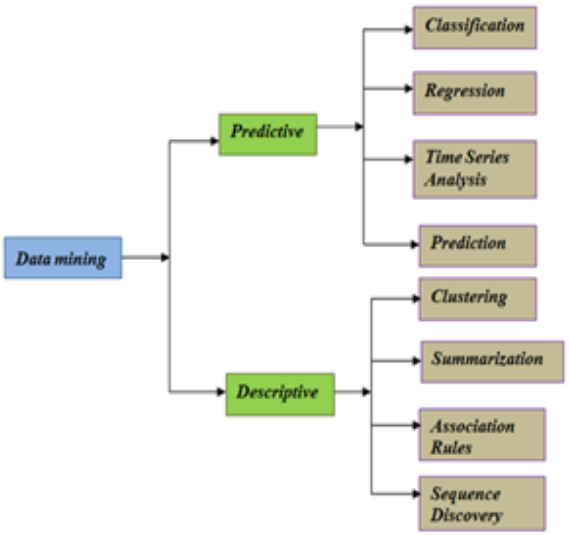

Figure 3: Data mining techniques

In data mining Clustering is a technique that's aims to single out the data elements into different clusters based on useful features. In this technique data elements that are similar to one another are placed within the same cluster and those which are dissimilar are placed in different clusters. Clustering technique is useful to identify different information by considering various examples and one can see where the similarities and ranges agree. By examining one or more attributes or classes, you can group individual pieces of data together to form a structure opinion. At a simple level, clustering is using one or more attributes as your basis for identifying a cluster of correlating results. Clustering can work both ways. You can assume that there is a cluster at certain point and then use our identification criteria to see if you are correct [5].

III.CLUSTERING ALGORITHMS

The goal of this paper is to provide a comprehensive review of different clustering techniques in data mining. Clustering is a division of data into groups of similar objects. Each group, called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups. Representing data by fewer clusters necessarily loses certain fine details but achieves simplification. It represents many data objects by few clusters, and hence, it models data by its clusters. [6] Clustering algorithm can be divided into the following categories:

- Hierarchical clustering algorithm
- K-means clustering algorithm
- Density Based Clustering algorithm
- Partition clustering algorithm
- Spectral clustering algorithm
- Grid based clustering algorithm

Hierarchical clustering algorithm

Hierarchical clustering algorithm groups data objects to form a tree shaped structure. It can be broadly classified into agglomerative hierarchical clustering and divisive hierarchical clustering. In agglomerative approach which is also called as bottom up approach, each data points are considered to be a separate cluster and on each iteration clusters are merged based on a criteria. The merging can be done by using single link, complete link, centroid or wards method. In divisive approach all data points are considered as a single cluster and they are splitted into number of clusters based on certain criteria, and this is called as top down approach[7]. Examples for this algorithms are LEGCLUST [8], BRICH [9] (Balance Iterative Reducing and Clustering using Hierarchies), CURE (Cluster Using Representatives) [10], and Chemeleon [11].

K-means clustering algorithm

K-means clustering is a partitioning method. K-means clustering is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean [12]. The k-means algorithm has the following important properties:

- ❖ It is efficient in processing large data sets.
- ❖ It often terminates at a local optimum
- ❖ It works only on numeric values.
- ❖ The clusters have convex shapes

Density Based Clustering Algorithm

Density based algorithm continue to grow the given cluster as long as the density in the neighbourhood exceeds certain threshold [11]. This algorithm is suitable for handling noise in the dataset. The following points are enumerated as the features of this algorithm.

- ❖ Handles clusters of arbitrary shape
- ❖ Handle noise
- ❖ Needs only one scan of the input dataset.
- ❖ Needs density parameters to be initialized.

DBSCAN, DENCLUE and OPTICS [11] are examples for this algorithm.

Partition Clustering Algorithm

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 a summary description of all the objects contained in a cluster. The precise form of this description will depend on the type of the object which is being clustered. In cases where real-valued data is available, the arithmetic mean of the attribute vectors for all objects within a cluster provides an appropriate representative; alternative types of centroid may be required in other cases; e.g., a cluster of documents can be represented by a list of those keywords that occur in some minimum number of documents within a cluster. If the number of clusters is large, the centroids can be further clustered to produce a hierarchy within a dataset [13].

Spectral Clustering Algorithm

Spectral clustering refers to a class of techniques, which relies on the Eigen structure of a similarity matrix. Clusters are formed by partitioning data points using the similarity matrix. Any spectral clustering algorithm will have three main stages [15]. They are preprocessing, spectral mapping and post mapping. Preprocessing deals with the construction of the similarity matrix. Spectral Mapping deals with the construction of Eigen vectors for the similarity matrix. Post Processing deals with the grouping of data points. The advantages of the spectral clustering algorithm are: strong assumptions on the cluster shape are not made; it is simple to implement and objective; it does not consider local optima; it is statistically consistent and works faster. The major drawback of this approach is that it exhibits high computational complexity. For large data set it requires O (n3), where n is the number of data points [16]. Examples of this algorithm are, SM(Shi and Malik) algorithm, KVV (Kannan,VempalaandVetta) algorithm, and NJW (Ng, Jordan and Weiss)algorithm [14].

Grid based Clustering Algorithm

Grid based Clustering Algorithm The grid based algorithm quant sizes the object space into a finite number of cells, that forms a grid structure [1].Operations are done on these grids. The advantage of this method is its lower processing time. Clustering complexity is based on the number of populated grid cells, and does not depend on the number of objects in the dataset. The major features of this algorithm are, no distance computations, Clustering is performed on summarized data points, Shapes are limited to the union of grid-cells, and the complexity of the algorithm is usually O(Number of populated grid-cells). STING [11] is an example of this algorithm.

IV. LITERATURE REVIEW

S. Anupama Kumar and M. N. Vijayalakshmi [17] illustrate that various data mining techniques like classification; clustering are apply on the student's data base. This study can be used to enable the learner and teaching community increase the performance. These techniques can also be combined with other specific discovery model to increase the capacity of the model. In this paper explain the many techniques of data mining according to Educational data to design a new environment Result of this paper is that education system can enhanced their performance by using data mining techniques. In this paper shows that every method has its own key area in which it performs accurate.

Bharat Chaudhari, Manan Parikh [18] represents comparative study of clustering algorithms using weka tools. Clustering is a process in which data is divided into different clusters according their functionality. Data of one cluster is different to another cluster but within that cluster data is homogenous. In this paper they compare the performance of clustering algorithm in term of class wise cluster building ability of algorithm. The outcomes of this paper is that k mean is better than other clustering algorithm(Hierarchical Clustering algorithm, Density based clustering algorithm) but is produce quality when we use large amount of data.

Sharaf Ansari, Sailendra Chetlur, Srikanth Prabhu, N. [19] management system. Student performance in university courses provide an overview of clustering algorithms used in data mining. They represent an important role in our life because we need much information (data) and we know that data mining is a process to extracting data and recognize the patterns. In this paper they provide an overview of some clustering analysis techniques such as DBSCAN, OPTICS, STING and CLIQUE.

Narendra Sharma , Aman Bajpai and Mr. Ratnesh Litoriya [20] represents the comparison between various clustering algorithm using weka tool .There are various tools in data mining which are used to analysis the data. They allow the users to analysis the data in different dimension or angles, categorize it, and summarize the relationships identified.

Weka is also a data mining tool which is used for analysis the data. The main objective is to show the comparison of the different- different clustering algorithms of weka and to find out which algorithm will be most suitable for the users. Every algorithm has their own importance and we use them on the behavior of the data, but on the basis of this research we found that k-means clustering algorithm is simplest algorithm as compared to other algorithms

.

Johns, S., Santos, M.V [21] represents a paper .on the Evolution of Neural Networks for Pair wise Classification Using Gene Expression Programming. Neural networks are a common choice for solving classification problems, but require experimental to adjustments of the topology are effective.

Connectivity based clustering, also known as hierarchical clustering, is based on the core idea of Objects being more related to nearby objects than to objects farther away. As such, these algorithms connect "objects" to form "clusters" based on their distance. A cluster can be described largely by the maximum distance needed to connect parts of the cluster.

At different distances, different clusters will form, which can be represented using a dendrogram, which explains where the common name "hierarchical clustering" comes from: these algorithms do not provide a single partitioning

of the data set, but instead provide an extensive hierarchy of clusters that merge with each other at certain distances. In a dendrogram, the y-axis marks the distance at which the clusters merge, while the objects are placed along the x-axis such that the clusters don't mix [22].

Clustering Method (RCM), which uses Subtractive Clustering combined with Fuzzy CMeans clustering along with a histogram sampling technique to provide quick and effective results for large sized datasets. Rapid Clustering Method can be used to cluster the dataset and analyze the characteristics in a social network. It can also be used to enhance the cross-selling practices using quantitative association rule mining [23].

V.ADVANTAGES AND DISADVANTAGES OF CLUSTERING ALGORITHM

The advantages and disadvantages of test cases reduction techniques in data mining are given in Table I.

Algorithms	*Advantages*	*Disadvantages*
Hierarchical clustering algorithm	No apriori information about the number of clusters required. Easy to implement and gives best result in some cases.	Algorithm can never undo what was done previously. No objective function is directly minimized
K-means clustering algorithm	If variables are huge, then K-Means most of the times computationally faster than hierarchical clustering, if we keep k smalls. K-Means produce tighter clusters than hierarchical clustering, especially if the clusters are globular.	Difficult to predict K-Value. Different initial partitions can result in different final clusters. It does not work well with clusters (in the original data) of Different size and Different density
Density Based Clustering algorithm	It detects erroneous data very well. It allows a brief description of non-spherical shaped clusters in high-dimensional data sets.	It needs many constants. It is less sensitive to outliers.
Partition clustering algorithm	Relatively scalable and simple. Suitable for datasets with compact spherical clusters that are well-separated	Poor cluster descriptors Reliance on the user to specify the number of clusters in advance High sensitivity to initialization phase, noise and outliers
Spectral clustering algorithm	Easy to implement and Good clustering results. Reasonably fast for sparse data sets of several thousand elements.	May be sensitive to choice of parameters Computationally expensive for large datasets
Grid based clustering algorithm	High computational efficiency and ability to identify clusters with complex structure. significant reduction of the computational complexity, especially for clustering very large data sets.	The quality of clustering depends on thegranularity of the lowest level of the grid structure. It results into inaccurate clusters despite the fastprocessing time of the technique

Table I. Advantages and disadvantages of clustering algorithm

VI. CONCLUSION

Data mining is a broad area that integrates with several fields including machine learning, statistics, pattern recognition, artificial intelligence, and to analysis of large volumes of data etc. This paper examines overview of techniques of data mining and clustering techniques.Cluster analysis is the task of assigning a set of objects into groups called clusters. Main task of clustering areexplorative data mining. Cluster analysis itself is not onespecific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in theirnotion of what constitutes a cluster and how to efficiently find them. During the survey, we also points advanced concepts of clustering techniques.

VII.REFERENCE

[1]. http://www.wideskills.com/data-mining-tutorial/05-data-mining-tasks

[2]. Joseph, Zernik, "Data Mining as a Civic Duty – Online Public Prisoners Registration Systems", International Journal on Social Media: Monitoring, Measurement, Mining, vol. - 1, no.-1, pp. 84-96, September2010.

[3]. Heikki, Mannila. 1996. Data mining: machine learning,statistics, and databases, IEEE
[4]. Zhao, Kaidi and Liu, Bing, Tirpark, Thomas M. and Weimin, Xiao,"A Visual Data Mining Framework for Convenient Identification of Useful Knowledge", ICDM '05 Proceedings of the Fifth IEEE International Conference on Data Mining, vol.-1, no.-1, pp. - 530- 537, Dec 2005.
[5]. BABU, G.P. and MARTY, M.N. 1994. Clustering with evolution strategies Pattern Recognition, 27, 2, 321-329.
[6]. P. Berkhin, 2002. Survey of Clustering Data Mining Techniques. Technical report, AccrueSoftware, San Jose, Cailf.
[7]. S.Anitha Elavarasi and Dr. J. Akilandeswari and Dr. B. Sathiyabhama, January 2011, A Survey On Partition Clustering Algorithms.
[8]. Santos, J.M, de Sa, J.M, Alexandre, L.A , 2008. LEGClust- A Clustering Algorithm based on Layered Entropic subgraph. Pattern Analysis and Machine Intelligence, IEEE Transactions : 62-75.
[9]. M. Livny, R.Ramakrishnan, T. Zhang, 1996. BIRCH: An Efficient Clustering Method for Very Large Databases. Proceeding ACMSIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery :103-114.
[10]. S. Guha, R. Rastogi, and K. Shim, 1998. CURE: An Efficient Clustering Algorithm for Large Databases. Proc. ACM Int'l Conf. Management of Data : 73-84.
[11]. Jiawei Han, Micheline Kamber, "Data Mining Concepts and Techniques" Elsevier Publication.
[12]. U. Boryczka, "Finding groups in data: Cluster analysis with ants," Applied Soft Computing Journal, vol. 9, pp. 61-70,2009.
[13]. P. Berkhin, 2002. Survey of Clustering Data Mining Techniques. Ttechnical report, AccrueSoftware, San Jose, Cailf.
[14]. Santos, J.M, de SA, J.M, Alexandre, L.A, 2008. LEGClust- A Clustering Algorithm based on Layered Entropic subgraph. Pattern Analysis and Machine Intelligence, IEEE Transactions: 62-75.
[15]. M Meila, D Verma, 2001. Comparison of spectral clustering algorithm. University of Washington, Technical report.
[16]. A. Geva, "Hierarchical unsupervised fuzzy clustering," IEEE Trans. Fuzzy Syst., vol. 7, no. 6, pp. 723–733, Dec. 1999.
[17]. S. Anupama Kumar and M. N. Vijayalakshmi" Relevance of Data Mining Techniques in Edification Sector" International Journal of Machine Learning and Computing, Vol. 3, No. 1, February 2013.
[18]. Bharat Chaudhari1, Manan Parikh2" 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
[19]. Sharaf Ansari1, Sailendra Chetlur2, Srikanth Prabhu3, N. Gopalakrishna Kini4, Govardhan Hegde5, Yusuf Hyder6" An overview of clustering algorithms used in data mining" International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 12, December 2013).
[20]. Narendra Sharma , Aman Bajpai and Mr. Ratnesh Litoriya" Comparison the various clustering algorithms of weka tools" International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, Volume 2, Issue 5, May 2012)
[21]. Johns, S., Santos, M. V.: On the Evolution of Neural Networks for Pairwise Classification Using Gene Expression Programming. In: Proceedings of the Annual Conference on Genetic and Evolutionary Computation, pp. 1903-1904, 2009
[22]. Lan Yu, "Applying Clustering to Data Analysis of Physical Healthy Standard", 2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2010), pp. 2766-2768.
[23]. J. Prabhu and M. Sudharshan and M. Saravanan and G.Prasad,(2010). Augmenting Rapid Clustering Method for Social Network Analysis", International Conference on Advances in Social Networks Analysis and Mining, pp. 407- 408.