ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 2, February 2015

An Overview of Clustering Techniques in Data Mining

N. Thinaharan, P.Vetriselvi

Asst. Professor, Dept. of Computer Science, Thanthai Hans Roever College, Perambalur, India

M.Phil Research Scholar, Dept. of Computer Science, Thanthai Hans Roever College, Perambalur, India

ABSTRACT: Data Mining refers to the analysis of observational datasets to find relationships and to summarize the data in ways that are both understandable and useful. It refers to extracting or mining the knowledge from large amount of data. The term data mining is appropriately named as "Knowledge mining". Data and Information or Knowledge has a significant role on human activities. Data mining involves the tasks like anomaly detection, classification, regression, association rule learning and clustering. 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 behavior of some of the clustering algorithms and their basic approaches.

KEYWORDS: Data Mining Techniques, Descriptive, Predictive, Clustering Techniques.

I. INTRODUCTION

The initiation of information technology in various fields of human life has direct to the large volumes of data storage in various formats like records, documents, images, sound recordings, videos, scientific data, and many new data formats. The data collected from different applications require proper mechanism of extracting knowledge /information from large repositories for better decision making. Knowledge discovery in databases (KDD), often called data mining, aims at the discovery of useful information from large collections of data [1]. Data mining is also stated as essential process where intelligent methods are applied in order to extract the data patterns.

Data mining consists of five major elements:

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

Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks. Data mining tasks can be classified in two categories descriptive and predictive. 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 [2].

II. DATA MINING TECHNIQUES

The various data mining techniques that allow extracting unknown relationships among the data items from large data collection that are useful for decision making. The purpose of a data mining effort is normally either to create a descriptive model or a predictive model. A descriptive model presents, in concise form, the main characteristics of the data set. It is essentially a summary of the data points, making it possible to study important aspects of the data set. 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

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 2, February 2015

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 [3].

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 [3].

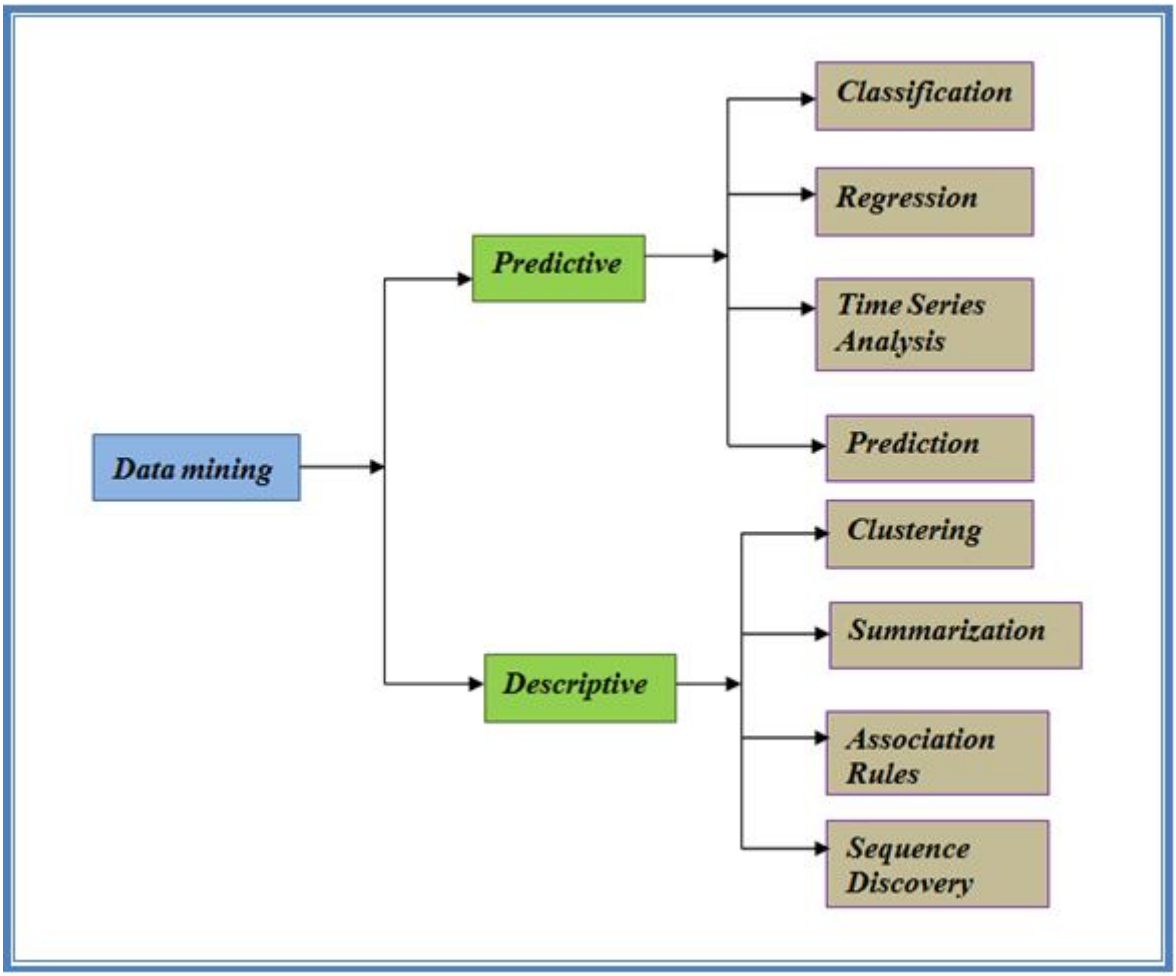

Fig.1. Data mining techniques

A. Classification

Classification consists of examining the features of a newly presented object and assigning to it a predefined class. The classification task is characterized by the well-defined classes, and a training set consisting of reclassified examples. The task is to build a model that can be applied to unclassified data in order to classify it [4]. Examples of classification tasks include:

- Classification of credit applicants as low, medium or high risk
- Classification of mushrooms as edible or poisonous
- Determination of which home telephone lines are used for internet access

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 2, February 2015

B. Prediction

Prediction method in combination with the other data mining techniques, involves analyzing trends, classification, pattern matching, and relation. Prediction is a wide topic and runs from predicting the failure of components or machinery, to identifying fraud and even the prediction of company profits. By analyzing past events or instances, you can make a prediction about an event [5].

C. Summarization

Summarization deals with continuously valued outcomes. Given some input data, we use summarization to come up with a value for some unknown continuous variables such as income, height or credit card balance.

D. Clustering

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 [6].

E. Association Rules

Association (or relation) is probably the better known and most familiar and straight forward data mining technique. Here, you make a simple correlation between two or more items, often of the same type to identify patterns.

F. Sequence Discovery

DUDAR and HART P describe the various uses of sequential patterns for identifying trends, or regular occurrences of similar events. For example, with customer data you can identify that customers buy a particular collection of products together at different times of the year. In a shopping basket application, you can use this information to automatically suggest that certain items be added to a basket based on their frequency and past purchasing history [7].

G. Association Rules

An association rule is a rule which implies certain association relationships among a set of objects (such as “occur together” or “one implies the other”) in a database. Given a set of transactions, where each transaction is a set of literals (called items), an association rule is an expression of the form X Y , where X and Y are sets of items. The intuitive meaning of such a rule is that transactions of the database which contain X tend to contain Y. An example of an association rule is: “30% of farmers that grow wheat also grow pulses; 2% of all farmers grow both of these items”. Here 30% is called the confidence of the rule, and 2% the support of the rule. The problem is to find all association rules that satisfy user-specified minimum support and minimum confidence constraints.

III. CLUSTERING TECHNIQUES IN DATA MINING

The goal of this survey 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. [8] Clustering algorithm can be divided into the following categories:

A. Hierarchical clustering algorithm
B. K-means clustering algorithm

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 2, February 2015

C. Density Based Clustering algorithm
D. Partition clustering algorithm
E. Spectral clustering algorithm
F. Grid based clustering algorithm

A. 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 splited into number of clusters based on certain criteria, and this is called as top down approach[9]. Examples for this algorithms are LEGCLUST [10], BRICH [11] (Balance Iterative Reducing and Clustering using Hierarchies), CURE (Cluster Using Representatives) [12], and Chemeleon [13].

B. 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 [14]. The k-means algorithm has the following important properties:

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

C. Density Based Clustering Algorithm

Density based algorithm continue to grow the given cluster as long as the density in the neighborhood exceeds certain threshold [13]. This algorithm is suitable for handling noise in the dataset. The following points are enumerated as the features of this algorithm.

1. Handles clusters of arbitrary shape
2. Handle noise
3. Needs only one scan of the input dataset.
4. Needs density parameters to be initialized.

DBSCAN, DENCLUE and OPTICS [13] are examples for this algorithm.

D. 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 [15].

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 2, February 2015

E. 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 [17]. 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 [18]. Examples of this algorithm are, SM(Shi and Malik) algorithm, KVV (Kannan,VempalaandVetta) algorithm, and NJW (Ng, Jordan and Weiss)algorithm [16].

F. 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 [13] is an example of this algorithm.

IV. LITERATURE REVIEW CLUSTERING TECHNIQUES

S. Anupama Kumar and M. N. Vijayalakshm [3] 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 [20] 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. [21] anagement 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 [22] 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.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 2, February 2015

Johns, S., Santos, M.V [23] 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 [24].

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 [25].

V. 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. During the survey, we also points advanced concepts of clustering techniques.

REFERENCES

1. Heikki, Mannila. 1996. Data mining: machine learning,statistics, and databases, IEEE
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. 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.
4. Survey of classification techniques in data mining in Proceedings of the International Multi Conference of Engineers and Computer Scientists 2009 Vol I IMECS 2009, March 18 - 20, 2009, Hong Kong-classification
5. T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Data Mining, Inference and Prediction, Springer, New York, 2001.
6. BABU, G.P. and MARTY, M.N. 1994. Clustering with evolution strategies Pattern Recognition, 27, 2, 321-329.
7. DUDA, R. and HART, P. 1973. Pattern Classification and Scene Analysis. John Wiley & Sons, New York, NY.
8. P. Berkhin, 2002. Survey of Clustering Data Mining Techniques. Ttechnical report, AccrueSoftware, San Jose, Cailf.
9. S.Anitha Elavarasi and Dr. J. Akilandeswari and Dr. B. Sathiyabhama, January 2011, A Survey On Partition Clustering Algorithms.
10. 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.
11. 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.
12. 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.
13. Jiawei Han, Micheline Kamber, “Data Mining Concepts and Techniques” Elsevier Publication.
14. U. Boryczka, "Finding groups in data: Cluster analysis with ants," Applied Soft Computing Journal, vol. 9, pp. 61-70,2009.
15. P. Berkhin, 2002. Survey of Clustering Data Mining Techniques. Ttechnical report, AccrueSoftware, San Jose, Cailf.
16. 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.
17. M Meila, D Verma, 2001. Comparison of spectral clustering algorithm. University of Washington, Technical report.
18. A. Geva, “Hierarchical unsupervised fuzzy clustering,” IEEE Trans. Fuzzy Syst., vol. 7, no. 6, pp. 723–733, Dec. 1999.
19. 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
20. 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
21. 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).

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 2, February 2015

22. 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)
23. 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
24. 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.
25. 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.

BIOGRAPHY

N. Thinaharan is an Assistant Professor in the Computer Science Department, Thanthai Hans Roever College,Bharathidasan University, He finished M.Sc., M.Phil, B.Ed, (Ph.D) pursuing degrees . Her research interests are Image Processing, Data Mining and Neural Network Algorithms etc.

Vetriselvi.P is a M.Phil Research Scholar in the Computer Science Department, Thanthai Hans Roever College, Bharathidasan University. Her research interests are Image Processing, Neural Network Algorithms etc.