A Review Paper on Data Mining Techniques and Algorithms

Shalu Swami[1], Ompal Jangir[2]
Assistant Professor
Department of Computer Application
Shekhawati Institute of Technology, Sikar,

Abstract: Data mining has made a great progress in recent year but the problem of missing data has remained a great challenge for data mining algorithms. It is an activity of extracting some useful knowledge from a large data base, by using any of its techniques.Data mining is used to discover knowledge out of data and presenting it in a form that is easily understood to humans. Data mining is the notion of all methods and techniques which allow analyzing very large data sets to extract and discover previously unknown structures and relations out of such huge heaps of details. This paper studied the classification and clustering techniques on the basis of algorithms which is used to predict previously unknown class of objects.

***Keywords:** - Data Mining, Classification, Clustering, Algorithms.*

*****

I. Introduction

Data mining is the process of extraction hidden knowledge from large volumes of raw data. Data mining has been defined as the nontrivial extraction of previously unknown, and potentially useful information from data. Data mining is used to discover knowledge out of data and presenting it in a form that is easily understood to humans[1]. Data mining is the notion of all methods and techniques which allow analyzing very large data sets to extract and discover previously unknown structures and relations out of such huge heaps of details.Data Mining is the process of extracting information from large data sets through the use of algorithms and techniques drawn from the field of Statistics, Machine Learning and Data Base Management Systems (Feelders, Daniels and Holsheimer, 2000). Traditional data analysis methods often involve manual work and interpretation of data that is slow, expensive an highly subjective (Fayyad, Piatsky Shapiro and Smyth, 1996). Data Mining, popularly called as knowledge discovery in large data, enablesfirmsandorganizations to make calculated decisions by assembling, accumulating, analyzing and accessing corporate data. It uses variety of tools like query and reporting tools, analytical processing tools, and Decision Support System (DSS) tools.

A)Data mining as a core process in KDD

The Knowledge Discovery in Database process comprises of a few steps leading from raw data collections to some form of new knowledge. It consists of the following steps as shown in figure 1.1

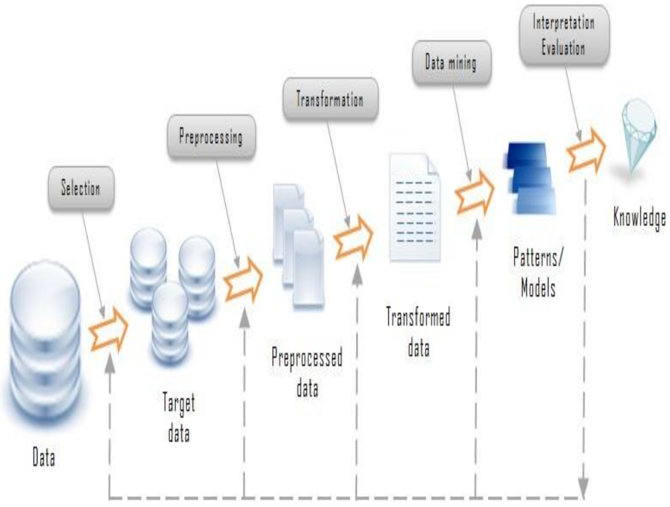

Figure 1.1:Data mining as a core process in KDD

- **Data cleaning:-**It is also known as the data cleansing, it is a phase in which noise data and relevant data removed from the collection.
- **Data integration:-**At this stage, multiple data sources, often heterogeneous, may be combined in a common source.
- **Data selection:-**At this step, the data relevant to the analysis is decided on and retrieved from the data collection.
- **Data transformation:-**It is also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.

- **Data mining:-**It is the crucial step in which clever techniques are applied to extract patterns potentially useful.

- **Pattern evaluation:-**In this step, strictly interesting patterns representing knowledge are identified based on given measures.

- **Knowledge representation:-**It is the final phase in which the discovered knowledge is visually represented to the user. This essential set uses visualization techniques to help users understand and interpret the data mining result[2]

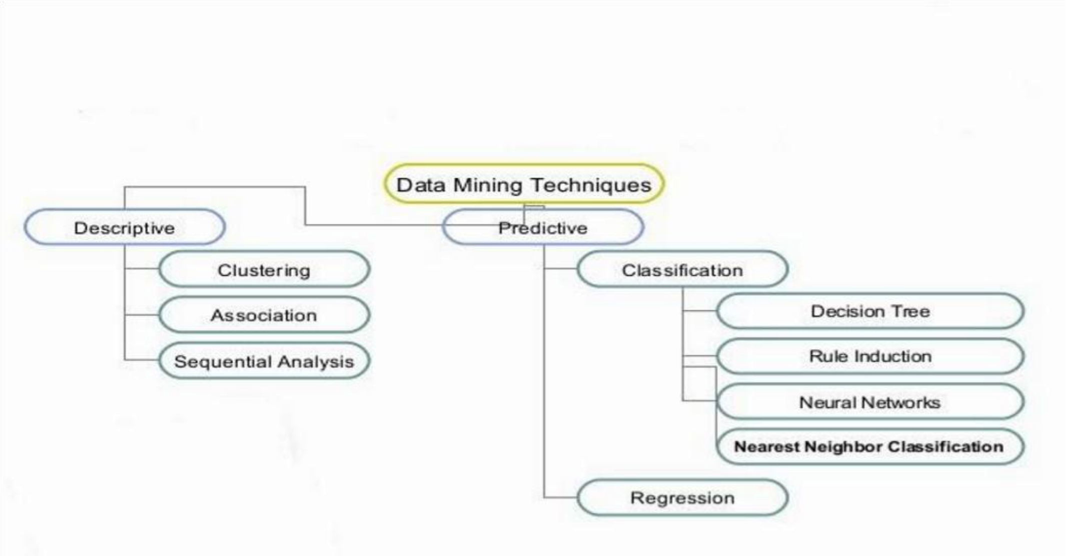

Figure.1 shows the descriptive and predictive data mining techniques.

Descriptive approach includes models for overall probability distribution of the data, partitioning of whole data into groups and models describing the relationships between the variables. Predictive approach permits the value of one attribute/variable is to predicted from the known values of other attribute/variable. This paper studies the one descriptive technique i.e. clustering and one predictive technique i.e. classification.

A) Classification Approach

Classification is a supervised learning method [3]. Data classification is two-step process. In the first step, a model is built by analyzing the data tuples from training data having a set of attributes. For each tuple in the training data, the value of class label attribute is known. Classification algorithm is applied on data training data to create the model. In the second step of classification, test data is used to check the accuracy of the model. If the accuracy of the model is acceptable then the model can be used to classify the unknown tuples
[4].Classification techniques were developed as an important componentofmachnine learning algorithms in order to extract rules and patterns from data that could be used for prediction. Classification techniques are used to classify data records into one among a set of predefined classes . They work by constructing a model of training dataset consisting of example records with known class labels[5].

B) Clustering Approach

Clustering is finding groups of objects such that the objects in one group will be similar to one another and different from the objects in another group. Clustering can be considered the most important unsupervised learning technique.Clustering can be considered the most important unsupervised learning technique so as every other problem of this kind. It deals with finding a structure in a collection of unlabeled data. Clusteringis the process of organizing objects into groups whose members are similar in some way [9]. Cluster analysis has been widely used in many applications such as business intelligence image pattern recoginition web search biology and security. In business intelligence clustering can be used to organize a large number of customers into groups where customers within a group share similar characteristics . This facilitates the development of business strategies for enhanced customer relationship management . In image recoginition clustering can be used to discover cluster or subclasses in handwritten character recoginition system. Suppose we have a data set of handwritten digits where each digit is labeled as either 1,2,3, and so on. Note that there can be a large variance in the way in which people write the same digit. Take the number 2, for example .some people may write it with a small cicle at the left bottom part , while some other may not. We can use clustering to determine sub classes for each of which represents a variation on the way in which 2 can be written. Using multiple models based on the subclasses can improve overall recognition accuracy[5].

The organization of this paper consists of following sections: Section 1 Lays the Basis of The Study, Section 2 Provides an overview of classification and clustering algorithm considered for study and Section 3 Concludes the study along with scope for future work.

II.Overview of Classification and clustering algorithms

A)Classification Algorithms:- A Classification Algorithm is a procedure for selecting a hypothesis from a set ofalternatives
that best fits a set of observations

1) Tree based CA:- tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.

1.1) Decision Stump:- A decision stump is a machine learning model consisting of a one-level decision tree.[1] That is, it is a decision tree with one internal node (the root) which is immediately connected to the terminal nodes (its leaves). A decision stump makes a prediction based on the value of just a
single input feature. Sometimes they are also called 1-rules

Decision stumps are often[6] used as components (called "weak learners" or "base learners") in machine learning ensemble techniques such as bagging and boosting. For example, a state-of-the-art Viola–Jones face detection algorithm employs AdaBoost with decision stumps as weak learners

1.2) J48:-J48 classifier is a simple C4.5 decision tree for classification. It creates a binary tree. The decision tree approach is most useful in classification problem. With this technique, a tree is constructed to model the classification process. Once the tree is built, it is applied to each tuple in the database and results in classification for that tuple[1][3].

2) Rules based classification algorithms:-Rule based classification algorithm also known as separate-and-conquer method. This method is an iterative process consisting in first generating a rule that covers a subset of the training examples and then removing all examples covered by the rule from the training set. This process is repeated iteratively until there are no examples left to cover [7].

Rule discovery or rule extraction from data is data mining techniques aimed at understanding data structures, providing comprehensible description instead of only black box prediction.

2.1) ZeroR:-ZeroR is the simplest classification method which relies on the target and ignores all predictors. ZeroR classifier simply predicts the majority category (class). Although there is no predictability power in ZeroR, it is useful for determining a baseline performance as a benchmark for other classification methods [8].Itis the simplest method which relies on the frequency of target. ZeroR is only useful for determining a baseline performance for other classification methods.

2.2) OneR:-OneR or “One Rule” is a simple algorithm proposed by Holt. The OneR builds one rule for each attribute in the training data and then selects the rule with the smallest error rate as its one rule. The algorithm is based on ranking all the attributes based on the error rate [9].To create a rule for an attribute, the most frequent class for each attribute value must be determined [10]. The most frequent class is simply the class that appears most often for that attribute value. A rule is simply a set of attribute values bound to their majority class. OneR selects the rule with the lowest error rate. In the event that two or more rules have the same error rate, the rule is chosen at random [11]. The OneR algorithm creates a single rule for each attribute of training data and then picks up the rule with the least error rate [12].

2.3)PART:- PART is a partial decision tree algorithm, which is the developed version of C4.5 and RIPPER algorithms .PART is a separate-and-conquer rule learner proposed by Eibe and Witten [54]. The algorithm producing sets of rules called decision lists which are ordered set of rules. A new data is compared to each rule in the list in turn, and the item is assigned the category of the first matching rule (a default is applied if no rule successfully matches). PART builds a partial C4.5 decision tree in its each iteration and makes the best leaf into a rule. The algorithm is a combination of C4.5 and RIPPER rule learning [13].

B)Clustering Algorithms:-Clustering is a data mining technique to group the similar data into a cluster and dissimilar data into different clusters.

1) K-Mean algorithm:-K-mean is an iterative clustering algorithm in which iteams are moved among sets of clusters unit the desired set is reached. As such, it may be viewed as a type of squared error algorithm, although the convergence criteria need not be defined based on the sequared error. A high degree of similarity among elements in clusters is obtained, while a high degree of similarity among elements in clusters is obtained while a high degree of dissimilarity among elements in different clusters is achieved simultaneously[6].

Sets of algorithm:-

a) First it selects the initial k prototypes arbitrarily.
b) The squared error criterion is used to determine the clustering quality.
c) In each iteration the prototype of each cluster is re-computed to be the cluster mean.
d) The basic version of k- means does not include any sampling techniques to scale to huge databases.

2) Hierarchical Algorithm:-Hierarchical clustering algorithms actually creates sets of clusters. Hierarchical algorithm differ in how the sets are created. A tree data structure called a dendrogram can be used to illustrate the hierarchical clustering technique and the sets of different clusters.

The root is a dendrogram tree contains one cluster where all elements are together. The leaves in the dendrogram each consist of a single element cluster. Internal nodes in the dendrogram represent new clusters formed by merging the clusters that appear as its children in the tree. Each level in the tree is associated with the distance measure that was used to merge the clusters. All clusters created at a particular level were combined because the children clusters had a distance between them less than the distance value associated with this level in the tree[6].

III. Conclusions and Future Work

This paper presents a detailed description of data mining techniques and algorithms. Therefore, Data Mining is the process of discovering interesting knowledge from large amounts of data stored either in databases, data warehouses, or other information repositories. The various algorithms used for the mining of data are specified in detail. The future scope provides enhancement and efficiency of data in the system. They could lead to better, faster and qualitative exaction of data with better tools and techniques.

References

[1]. Prajapati. D, Prajapat. J, “Handling missing values: Application to University Data Set”, August, 2011.

[2]. Grabmeier. J, Rudolph. A, “Technique of Clustering Algorithms in Data Mining”, Data Mining and Knowledge Discovery,2002.

[3]. Han. J, Kamber. M, Pei. J, “ Data Mining Concepts and Techniques”, Third edition The Morgan Kaufmann Series in Data Management Systems Morgan Kaufmann Publishers, July 2011

[4]. Kabra. R, Bichkar. R, “Perfoemance Prediction of Engineering Students using Decision Tree”, International Journal of computer Applications, December ,2011

[5]. VikramPudi,PRadha Krishna “Data Mining”,Oxford University Press, First Edition,2009

[6]. Margaret H. Dunham, “Data Mining Introductory and Advanced Topics”,Dorling Kindersley Pvt.Ltd.India,Sixth Edition,2013.

[7]. Phyu, Thair Nu. "Survey of classification techniques in data mining."InternationalMultiConferenceof Engineers and Computer Scientists, 2009.

[8]. http://www.saedsayad.com/zeror.htm

[9]. Tayel , Salma, et al. “Rule-based Complaint Detection using RapidMiner”, Conference: RCOMM 2013, At Porto, Portugal, Volume: 141-149,2014

[10]. http://mydatamining.wordpress.com/2008/04/14/rule-learner-or-rule-induction/

[11]. Vijayaran S, Sudha. *“An Effective Classification Rule Technique for Heart Disease Prediction”*.International Journal of Engineering Associates, February 2013.

[12]. Buddhinath, Gaya, and Damien Derry."A simple enhancement to One Rule Classification." *Department of Computer Science & Software Engineering. University of Melbourne, Australia* (2006).

[13]. Ali, Shawkat, and Kate A. Smith."On learning algorithm selection for classification." *Applied Soft Computing*, 2006.