An Issues, Challenges and Big Data Mining

Fatima[1], Dr Md. Jawed Ikbal Khan[2]
[1]*Research Scholar, Magadh University Bodh-Gaya, Gaya, Bihar, India*
[2]*Associate Professor, Dept. of Mathematics, Mirza Ghalib College, Gaya , Bihar, India*

***Abstract*- Data mining is a process which finds useful patterns from large amount of data by collection of data into information. The concept of data mining is centre of appeal for the users because of many factors as high availability of data which needs to be converted from masses of data to useful information. The list of sources that generate these data is endless. Businesses worldwide generate gigantic sets of data everyday that may include stock, transactions and many more of similar kinds. So there comes the need of powerful and most importantly automatic tools for uncovering valuable slots of organized information from tremendous amount of data. Considering any social networking site or a search engine, they receive millions of queries every day. Firstly, the Database Management Systems evolved to handle the queries of similar types. Then the approach was modified to advanced Database management system, Data Warehousing and Data mining for advance data analysis and web based databases.**

Index terms- Data mining, Issue and Challenges and Classification of algorithms.

I. INTRODUCTION

Data mining is used for exploring and analyzing large amounts of data to find patterns for big data. The advent of big data, the data mining is more prevalent. Four or five years ago, companies collected all data of transaction stored in a single database. Today, volume of data is collected have explode.

Marketers can also collect information about every conversation people are having about their brand. It requires the implementation of new Processes, technology and governance mechanisms that are collectively being referred to as big data. Today, big data is a big business. We can define big data is a process that allows companies to extract an information from large amount of data. Big data is used data mining techniques because size of information is larger. The last decade has practiced a revolution in info accessibility and exchange of it through web. Within the same strength additional business in addition as organizations began to gather knowledge associated with their own operations, whereas the info applied scientist are seeking economical mean of storing, retrieving and manipulating knowledge, the machine learning community centered on techniques that used for developing, learning and getting information from the info. data {processing} is that the process of analyzing knowledge from completely different views and summarizing it into helpful info. Data mining consists of extract, transform, and cargo dealing knowledge onto the info warehouse system, store and manage the info during a third dimensional info system, by victimization application computer code analyze the info, give knowledge access to business analysts and data technology professionals, gift the info during a helpful format, sort of a graph or table. Data processing involves the anomaly detection, association, classification, regression, rule learning, account

II. DATA MINING

Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. Data mining is an interdisciplinary subfield of computer science and statistics with an overall goal to extract information from a data set and transform the information into a comprehensible structure for further use. Data mining is the analysis step of the "knowledge discovery in databases" process, or KDD.

III. STEPS IN DATA MINING PROCESS

A) Classification Approach

Classification is a supervised learning method [1]. 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 [2].Classification techniques were developed as an important component of machine 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 [3].

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. Clustering is the process of organizing objects into groups whose members are similar in some way [4]. 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 circle 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 [3].

IV. ISSUES, CHALLENGES AND BIG DATA PROBLEMS IN DATA MINING

A. Problems

The main problems in big data has grown extremely. This large amount of data is beyond the of software tools to manage. The exploring a large amount of data, exacting a useful information from data sets and knowledge is a challenge, sometimes it is a major problems. Also big data is unstructured, large size and it is not easy to handle.

B. Issues

The main issues of data mining in big data are follows

a) Poor data quality e.g. noisy data, dirty data and inadequate size of data.
b) Redundant data is uploaded from various sources such as multimedia files.
c) Security, privacy of the companies
d) Algorithm of data mining is not effective.
e) Difficult to processing an unstructured data into structured data.
f) Higher cost, less flexibility.

C. Major challenges

a. Big Data Mining Platform

b. Dig Data Semantics and Application Knowledge

- Information Sharing and Data Privacy
- Domain and Application Knowledge

c. Big Data Mining Algorithm

- Local Learning and Model Fusion for Multiple Information Sources
- Mining from Sparse, Uncertain, and Incomplete Data
- Mining Complex and Dynamic Data

VII. CLASSIFICATION OF ALGORITHMS

A) Classification Algorithms: - A Classification Algorithm is a procedure for selecting a hypothesis from a set of alternatives 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.[3] 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[3][5].
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].It is 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 [4].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 . 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 items 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 squared 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 differs 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.

VIII. CONCLUSION

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. Data mining has wide application domain almost in every industry where the data is generated that's why data mining is considered one of the most important frontiers in database and information systems and one of the most promising interdisciplinary
Developments in Information Technology. 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.

REFERENCE

[1] 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
[2] Grabmeier. J, Rudolph. A, "Technique of Clustering Algorithms in Data Mining", Data Mining and Knowledge Discovery,2002.
[3] Kabra. R, Bichkar. R, "Perfoemance Prediction of Engineering Students using Decision Tree", International Journal of computer Applications, December ,2011
[4] VikramPudi,PRadha Krishna "Data Mining", Oxford University Press, First Edition,2009
[5] Tayel , Salma, et al. "Rule-based Complaint Detection using RapidMiner", Conference: RCOMM 2013, At Porto, Portugal, Volume: 141-149,2014
[6] Prajapati. D, Prajapat. J, "Handling missing values: Application to University Data Set", August, 2011.
[7] Margaret H. Dunham, "Data Mining Introductory and Advanced Topics", Dorling Kindersley Pvt. Ltd. India, Sixth Edition,2013.
[8] Phyu, Thair Nu. "Survey of classification techniques in data mining." International MultiConference of Engineers and Computer Scientists, 2009.
[9] http://www.saedsayad.com/zeror.htm
[10] Tayel , Salma, et al. "Rule-based Complaint Detection using RapidMiner", Conference: RCOMM 2013, At Porto, Portugal, Volume: 141-149,2014
[11] http://mydatamining.wordpress.com/2008/04/14/rule-learner-or-rule-induction/
[12] Vijayaran S, Sudha. "An Effective Classification Rule Technique for Heart Disease Prediction". International Journal of Engineering Associates, February 2013.
[13] Buddhinath, Gaya, and Damien Derry. "A simple enhancement to One Rule Classification." Department of Computer Science & Software Engineering. University of Melbourne, Australia (2006).
[14] Ali, Shawkat, and Kate A. Smith. "On learning algorithm selection for classification." Applied Soft Computing, 2006.