A TECHNIQUE FOR MINING THE DATA TO GENERATE QUICK RESPONSE

Devarepu Navya Choudary
navyadevarapu13@gmail.com

Satya Sainath Mydavolu
sainathsatya99@gmail.com

J.Shree Harsha
Harsha1441@gmail.com

Lakshmi Sindhu Gorantla
indhu.sosweet@gmail.com

ABSTRACT

We know that data is the collection of information.A lot of data that can producevaluable information, The type of information is produced by the data and it decides the method for processing .Data mining techniques are mainly used to find the required information and relationships of large amount of data, which helps in decision making. Data mining helps in extracting meaningful new patterns that cannot be found necessarily by processing data. If the data is too big then the process of mining takes more time.so here we used to calculate the size of data and type of data and then identify which type of technique or algorithm or method is preferable to generate a quick solution or response.

KEYWORDS: KDD, Mining, Datamining, Data, Method.

1. INTRODUCTION

DataMining (formally termed Knowledge Discovery in Databases – KDD) is a process that aims to use existing data to invent new facts and to uncover new relationships previously unknown even to experts thoroughly familiar with the data. It is like extracting precious metal (say gold etc.) and/or gems, hence the term "mining", It is based on filtration and assaying of mountain of data "ore" in order to get "nuggets" of knowledge[4]. The data mining process is diagrammatically exemplified in Figure below
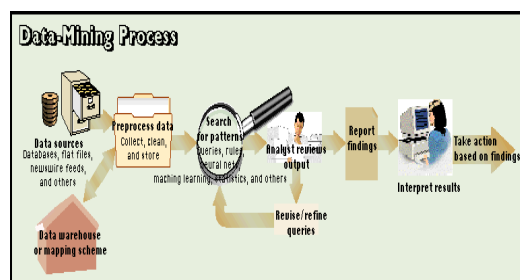
Data mining can be used in conjunction with a data warehouse to help with certain types of decisions.

• Data mining can be applied to operational databases with individual transactions.

• To make data mining more efficient, the data warehouse should have an aggregated or summarized collection of data.

• Data mining helps in extracting meaningful new patterns that cannot be found necessarily by merely querying or processing data or metadata in the data warehouse.

Data Mining Techniques Can Be Implemented Rapidly On Existing Software And Hardware Platforms Across D&B To Enhance The Value Of Existing Resources, And Can Be Integrated With New Products And Systems As They Are Brought On-Line. When Implemented On High Performance Client-Server Or Parallel Processing Computers, Data Mining Tools Can Analyze Massive Databases
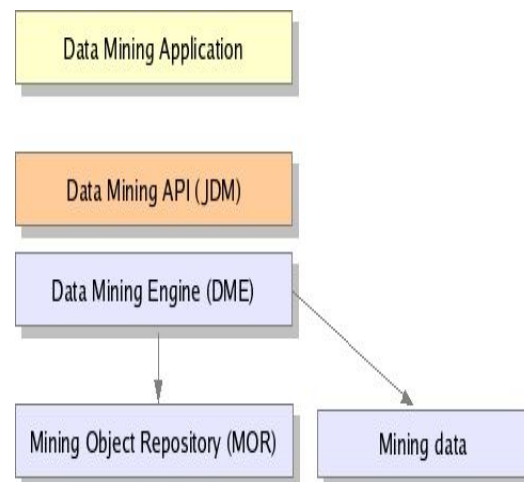

Figure 2: Extraction of mining data

Figure 1: process of Data Mining

A key challenge for data mining is tackling the problem of mining richly structured datasets, where the objects are linked in some way. Links among the objects may demonstrate certain patterns, which can be helpful for many data mining tasks and are usually hard to capture with traditional statistical models. Recently there has been a surge of interest in this area, fueled largely by interest in web and hypertext mining, but also by interest in mining social networks, security and law enforcement data, bibliographic citations and epidemiological records[1].

Data Mining (sometimes called data or knowledge discovery) has become the area of growing significancebecause it helps in analyzing data from different perspectives and summarizing it into useful information [5].

A Data mining query outputs aKDD object. A KDD object is either a rule, classifier or clustering [7]. These objects do not exist prior to executing the query nor or they part of the database being queried.

Data mining algorithms such as Decision trees, /regression, cluster analysis continue to form a traid of core algorithms for data miners. This has been very consistent over time. However, a wide variety if algorithms are being used as mentioned in the below

Figure 3: Different types of algorithms used for mining

An important aspect of data mining algorithms and systems is that they should scale well to large databases. A consequence of this is that most data mining tools are based on machine learning algorithms that work on datain attribute-value format. One way to enlarge the expressiveness is to generalize to support mining on full relational databases. The key step in such a generalization is to ensure that the search space does not explode and that efficiency and, thus, scalability are maintained. Different kinds of methods and algorithms are being used to generate mining purpose .

Methods and algorithms such as Classification, Clustering, Frequent Pattern Mining, Graphical Models, Spatial & TemporalMining, Data Stream Mining, Abnormality & Outlier Detection, Feature Extraction, Selection and Dimension Reduction, Mining with Constraints, Data Cleaning & Preprocessing, Computational Learning Theory, Multi-Task Learning, Online AlgorithmsBig Data, Scalable & High-Performance Computing Techniques, Mining with Data Clouds, Mining Graphs, Mining Semi Structured Data, Mining Complex Datasets, Mining on Emerging Architectures, Text & Web Mining, Optimization Methods, Other Novel Methods ,etc.[4],

Top most algorithms which will be used generally in data mining are C4.5, k-Means, SVM, Apriori, EM, Page Rank, AdaBoost, kNN, Naive Bayes, and CART. These top 10 algorithms are among the most influential data mining algorithms in the research community. With each algorithm, we provide a description of the algorithm, discuss the impact of the algorithm, andreview current and further research on the algorithm. These 10 algorithms cover classification, clustering, statistical learning, association analysis, and link mining, which are all among themost important topics in data mining research and development [2].

Classification is a well-studied problem in data mining. Classification performance was originally gauged almost exclusively using predictive accuracy, but as work in the field progressed, more sophisticated measures of classifier utility that better represented the value of the induced knowledge were introduced. Nonetheless, most work still ignored the cost of acquiring training examples, even though this cost impacts the total utility of the data mining process the relationship between the number of acquired training examples and the utility of the data mining process and, given the necessary cost information, we determine the number of train-
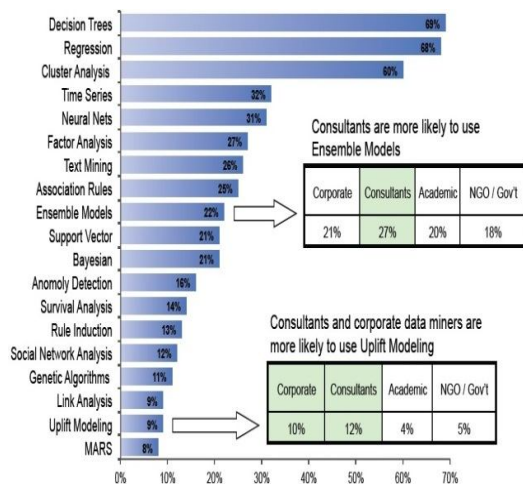
103

ing examples that yields the optimum overall performance. We then extend this analysis to include the cost of model induction--measured in terms of the CPU time required to generate the model. While our cost model does not take into account all possible costs, our analysis provides some useful insights and a template for future analyses using more sophisticated cost models. Because our analysis is based on experiments that acquire the full set of training examples, it cannot directly be used to find a classifier with optimal or near-optimal total utility. To address this issue we introduce two progressive sampling strategies that are empirically shown to produce classifiers with near-optimal total utility. [3].

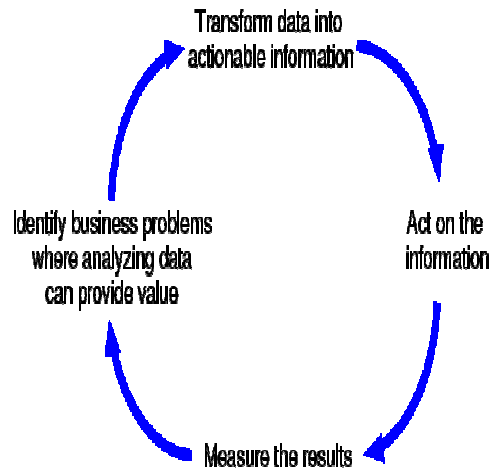The Process Of Data Mining Consists Of Three Steps such as Data Preparation, Model Building, Deployment

Figure 4: Virtuous cycle of data mining

An important step in the data mining process is data pre processing. One of the challenges that face the knowledge discovery process in meteorological data is poor data quality. For this reason we try to prepare our data carefully to obtain accurate and correct result [6].

To determine accurate number of clusters from gene expression data and validate the results using correctness ratio and sum of squares criteria. A new approach suggested to addresses the primary issue of k-means clustering algorithm that predefining number of clusters. This approach provides accurate number of clusters by minimizing the squared error function and maximizing the correctness ratio value. [9]

2. IDENTIFICATION OF DATA

The goals of data mining fall into the following classes

Prediction

Data mining can show how certain attributes within the data will behave in the future.

Identification

Data patterns can be used to identify the existence of an item, an event, or an activity.

Classification

Data mining can partition the data so that different classes orcategories can be identified based on combinations of parameters.

Optimization

One eventual goal of data mining may be to optimize the use of limited resources such as time, space, money, or materials and to maximize output variables such as sales or profits under a given set of constraints.

3. TECHNIQUE

In the process of mining we use to reduce the response time for generating output for the requested query .so, in the pre processing method first identify the type of data whether it is related to clusted, classification etc and also the size of data whether the data which is subjected for mining is less or more .so basing on these results our system will identify a suitable algorithm or method or technique which helps to generate a quick response in the process of mining the data . The process of selecting the suitable technique or algorithm or method is represented below
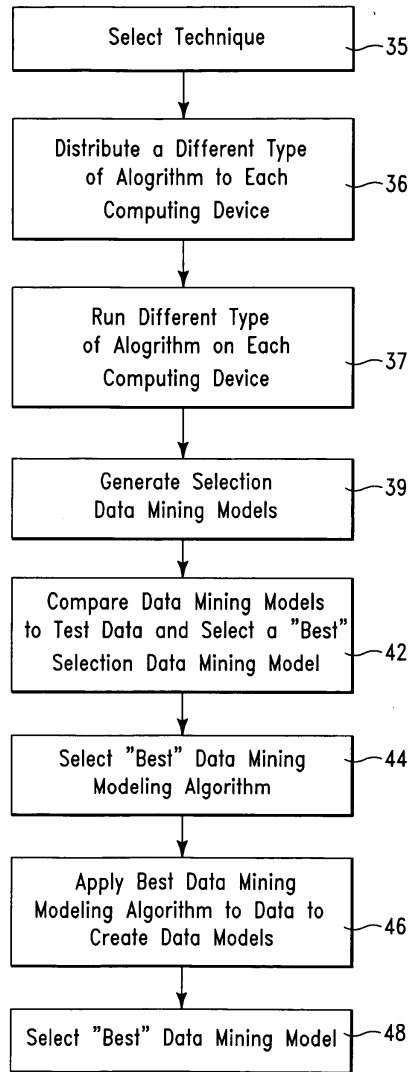
```
┌─────────────────────────┐
│     Select Technique     │──35
└─────────────────────────┘
            │
┌─────────────────────────┐
│  Distribute a Different Type │
│  of Alogrithm to Each    │──36
│   Computing Device       │
└─────────────────────────┘
            │
┌─────────────────────────┐
│   Run Different Type     │
│   of Alogrithm on Each   │──37
│   Computing Device       │
└─────────────────────────┘
            │
┌─────────────────────────┐
│   Generate Selection     │──39
│   Data Mining Models     │
└─────────────────────────┘
            │
┌─────────────────────────┐
│ Compare Data Mining Models │
│ to Test Data and Select a "Best" │──42
│ Selection Data Mining Model │
└─────────────────────────┘
            │
┌─────────────────────────┐
│  Select "Best" Data Mining │──44
│   Modeling Algorithm     │
└─────────────────────────┘
            │
┌─────────────────────────┐
│  Apply Best Data Mining  │
│ Modeling Algorithm to Data to │──46
│   Create Data Models     │
└─────────────────────────┘
            │
┌─────────────────────────┐
│ Select "Best" Data Mining Model │──48
└─────────────────────────┘
```

Figure 5: Procedure to select best data mining model

Generally every algorithm or technique will have its own specialization so we compare the pre-processed results with each and every techniques or algorithms so that we can identify which method will help to process easily

Most common data-mining techniques used in predictive analytics. Becoming familiar with them will go a long way toward enabling you to recognize patterns in customer preferences and buying behaviour. techniques such as regression analysis, choice modelling,rule induction , network/link analysis, clustering, neural n/w, mbr/cbr, decision tree, uplift modelling.[10].

Other data mining techniques used such as point estimation , models based on summerization, bayes theorem, and genetic algorithm [8].

The data which is subjected for mining will be preprocessed efficiently and effectively to get the suitable technique for mining such data .so that the processing time will be reduced than other methods .

4. CONCLUSION
The process of mining the data is to extract the required information from the large amount of data .In the process of mining we use different techniques and methods and algorithms but mostly use a particular method or algorithm only. sometimes inorder to get the response of mining it takes more time , so to reduce such drawbacks here in this paper we have proposed a new technique which will identify the size ,type of data which is to be mined so that for that particular data automatically the preferable method or algorithm which give quick response than other methods will be selected .so this helps to generate a quick response in mining the data .

REFERENCES

[1] LiseGetoor, Link mining: a new data mining challenge, ACM SIGKDD Explorations Newsletter, v.5 n.1, July 2003
[2]KnowlInf System (2008) 14:1–37DOI 10.1007/s10115-007-0114-2
[3] Data mining and knowledge discovery volume 17 issue 2, October 2008 pages 253-282
[4] Data mining concepts and Techniques jai-weihan and michelinekamber 2nd edition .
[5] C.Romero,S.Ventura,E.Garcia,"Datamining in course managementsystems: Moodle case study and tutorial", Computers &Education,Vol. 51, No. 1, pp. 368-384, 2008
[6] Implementation of Data Mining Techniquesfor Meteorological Data AnalysisInternational Journal of Information and Communication Technology ResearchVolume 1 No. 3, July 2011 ISSN-2223-4985
[7] Tomasz imielinski and heikki manila. A database perspective on knowledge discovery. Communications of the ACM ,39(11): 58-64 ,November 1996.
[8] Data mining techniques and algorithm partial draft of forth coming book fromprentice hall, Margaret H. Dunham august 23, 2000 .

[9] An Efficient Unified K-Means Clustering Technique for Microarray Gene Expression Data-Journal of Computer Science 7 (7): 954-957, 2011 ISSN 1549-3636

[10] Most common data mining techniques used in predictive analytics by Laura Paterson April 27, 2010