International Journal of Innovative Research in Science and Engineering
Vol. No.2, Issue 10, October 2016
www.ijirse.com
ISSN (O) 2454-9665
ISSN (P) 2454-0663

DISCOVERING UNSEEN VALUES WITH PREDICTIVE DATA MINING

C.Harinath[1], G.Mohammed Afzal [2], A.Kaleemullah[3], P.Rizwan Ahmed[4]

[1,2]Assistant Professor of CA, [3]Assistant Professor of CS & Head, [4]Assistant Professor & Head of CA(UG & PG), Mazharul Uloom College, Ambur

ABSTRACT

Data Mining is an analytic process to explore data (usually large amounts of data - typically business or market related) in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new sets of data. The main target of data mining application is prediction. Predictive data mining is important and it has the most direct business applications in world. The paper briefly explains the process of data mining which consists of three stages: (1) the Initial exploration, (2) Pattern identification with validation, and (3) Deployment (application of the model to new data in order to generate predictions). Data Mining is being done for Patterns and Relationships recognitions in Data analysis, with an emphasis on large Observational data bases. From a statistical perspective Data Mining is viewed as computer automated exploratory data analytical system for large sets of data and it has huge Research challenges in India and abroad as well. Machine learning methods form the core of Data Mining and Decision tree learning. Data mining work is integrated within an existing user environment, including the works that already make use of data warehousing and Online Analytical Processing (OLAP). The paper describes how data mining tools predict future trends and behaviour which allows in making proactive knowledge-driven decisions.

Keywords: ***Predictive Data mining, Data Warehousing, Decision tree learning, OLAP***

I. INTRODUCTION

Data mining is the process of automatically searching large amount of data for patterns and recognition of specific characteristics. A statistical perspective of Data Mining is viewed as computer automated exploratory data analytical system for large sets of data and it has huge research challenges in India and abroad as well. Machine learning methods form the core of Predictive Data Mining and Decision tree learning methods. Data mining work is integrated within an existing user environment, including the works that already make use of data warehousing and OLAP. Data mining provides:

1) Automated prediction of trends and behaviour.
2) Automated discovery of previously unknown patterns.

We have classified whole data mining concept in this paper, its application and its future trend technologies. The work so obtained is the result of schematic and systematic analysis and review.

ISSN (O) 2454-9665
ISSN (P) 2454-0663

1.1 Data Mining

Data mining methods is classified in two ways:

- Through Applied Functions.
- Through Class of Applications.

Evolution of Data Mining began when the collection of data especially for Business and Bioinformatics applications were first stored on computers and it got continued with improvements in data access generated technologies. Data mining takes evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery.

Data mining techniques is supported by three technologies i.e. massive data collection, Powerful multiprocessor computers and Data mining algorithms. Data mining tasks are Descriptive (discovering interesting patterns describing the data) and Predictive (predicting behaviour of model based available data). Data mining executes on the basis of computational techniques and methods from statistics machine learning and pattern recognition. Data mining techniques navigates through large amount of databases and extracts specific and required amount of data, which gets translated and transcribed into useful and predictive information as per needs. Various data mining software analyzes relationships and patterns in stored transaction of data and enables automatic detection of patterns in a database and accelerates the initial stages of information analysis. The data mining process can be characterized as a multi-stage iterative process involving data selection, data cleaning, application of data mining algorithms and evaluation. [1]

1.2 Exploring and Preprocessing

Initial steps of exploring, visualizing and querying data, to gain insight into data in an interactive manner. Preprocessing steps such as variable selection, data focusing, and data validation can also be included in these initial steps such as electing the model representations, selecting the score functions that score different models with respect to the data and specifying the computational methods and algorithms to optimize the score function. These components combined together specify the data mining algorithm which is to be used and the components may be precompiled into a specific algorithm (e.g., CART decision tree implementations) or can be integrated in a customized manner for a specific application such as:

- **Mining:** Step of implementing a particular data mining algorithm on a particular data set.
- **Evaluating:** Step of evaluating the quality of output of data mining algorithm from step 3, both predictions and interpretation of the model.
- **Deploying:** Step of putting a model from a data mining algorithm into predictive use for continuous data stream applications

II. METHOD

2.1 Data Mining Classifications:

They are being classified mainly into 6 broad categories as mentioned below:

1) **Classes:** Stored data can be used to locate data in predetermined groups.
2) **Clusters:** Data items can be grouped according to logical relationships or consumer preferences. E.g.

ISSN (O) 2454-9665
ISSN (P) 2454-0663

Data can be mined for identification of market segments.

3) **Associations:** Data can be mined to identify associations.
4) **Sequential patterns:** Data can be mined for anticipation of behaviour patterns and trends.
5) **Boolean Vector:** It is compared to analyze items frequently associated together.
6) **Forecasting:** It involves predictions of data.

2.2 Data Warehouse and OLAP

A data warehouse is a subject oriented, integrated, time variant, non volatile collection of large amount of data which supports Data Mining Techniques and related methods. It is an organized system of data derived from multiple data sources designed for decision making and its applications. Data warehouse maintains a central repository of all organizational data in which OLAP (On-Line Analytical Processing) is a method by which multidimensional analysis takes place. [2] OLAP facilities can be integrated into corporate data base systems and it allows analysts and managers to monitor the live performance of the business and market. OLAP techniques can be simple like frequency, tables, descriptive statistics, and simple cross-tabulations or complex as they involve seasonal adjustments and parameters and other forms of data. Data mining techniques can beoperated on unstructured information and it can be applied on summaries generated by OLAP to provide more in-depth and multidimensional knowledge in different steps as follows:

- 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 information technology and bioinformatics sectors.

Analyze the data with supporting application software and present the useful data in a format, like graph, table, charts, decision tree, cubes, etc.

2.3 Techniques Applied For Data Mining And Warehousing:

1) **Artificial neural networks:** Non-linear predictive models that function through training and resemble biological neural networks in its structure. Neural Networks are analytic techniques modeled after the processes of learning in the cognitive system and the neurological functions of the brain and capable of predicting new observations.
2) **Decision trees:** Tree-shaped non-linear structures represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID, A decision tree that uses contingency tables and the chi-square test to create the tree).
   [3] Also, decision tree of WEKA mining suite can be used to classify data in tree format.
3) **Genetic algorithms:** This is the new technique which is being prevalent now-a-days in Biotechnology and Bioinformatics fields. Optimization techniques that use process such as genetic combination, mutation, and natural selection in a design based on the concepts of evolution of genes.
4) **Rule induction:** The extraction of data by if-then rules based on statistical significance for retrieval of data.

International Journal of Innovative Research in Science and Engineering
Vol. No.2, Issue 10, October 2016
www.ijirse.com
ISSN (O) 2454-9665
ISSN (P) 2454-0663

After various techniques have been applied according to the need of user, chronological steps are to be followed for obtaining mined and required data as follows:

1) **Data Cleansing:** The power center's data cleansing technology improves data quality by validating, correctly naming and standardization of address data.
2) **Data Transformation:** Transforms the source data according to the requirements of the system. Transformations ensure the quality of the data being loaded into target and this can be done during the mapping process from source to target.
    - Sorting and Smoothing (binning, clustering & regression), aggregation, generalization and normalization. [6]
3) **Workflow:** Workflow loads the data from source to target in a sequential manner.
4) **Detailed study of pattern growth**: Adopts divide & conquer strategy for finding user defined data and compress data base into a frequent pattern tree (FP-tree) and stores associated information.

Through these methods and techniques the flow of Data from data warehouse to Assimilation of data can be done which is shown in *figure1*:

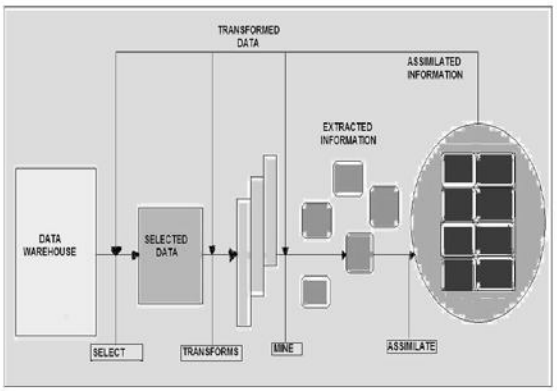

Figure1: Processed Data Mining, Process of Knowledge Discovery in Databases (K.D.D.)

Data warehouse integration with OLAP server provides End-User solution and can be implemented in relational database system like Sybase, Oracle and Redbrick. Technological infrastructure / driver requirement are size of the database, query complexity and dimensional of data. A data warehouse can be normalized or renormalized and it can be a relational database, multidimensional database, flat file and hierarchical database as shown in figure 2.

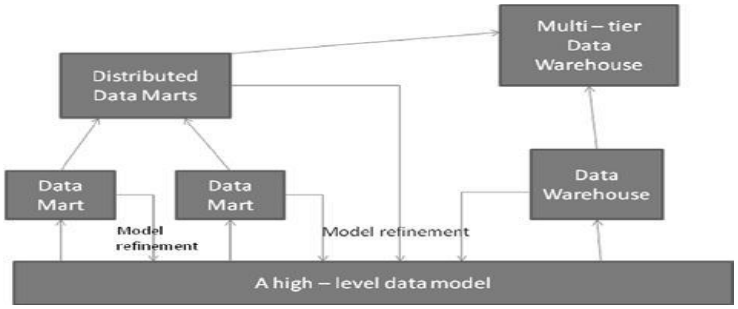

Figure2: Implementation of Data Warehouse in incremental and evolutionary manner.

International Journal of Innovative Research in Science and Engineering
Vol. No.2, Issue 10, October 2016
www.ijirse.com
ISSN (O) 2454-9665
ISSN (P) 2454-0663

Table 1: Comparison between without pattern Data Warehouse and with pattern data warehouse

Without Pattern Warehouse	With Pattern Warehouse:
Manipulate raw data to find patterns.	Patterns can be found with ease.
Analysis for pattern discovery is implemented.	Patterns can be stored in central repository.
Analytical reports are difficult to understand and have no explanations.	Graphic user interface is implemented for pattern query.
Analysis was performed on extract files.	The entire database is analyzed.

Data Mining and Information Warehouse framework Data mining tools discover useful and important facts buried in the raw data with the application known as discovery model. They complement use of queries, multidimensional analysis and visualization tools to gain a better understanding about data. Good facilities and software's are available to perform queries and data visualization as well as the availability of powerful data mining operators can be a part of well architected Decision Support Environment [4], [5]. It is much similar like a regular mining process, which digs out raw material as it may exist in a mine and through several steps extracts valuable and required metal from the ore. Data mining comprises three distinct phases: <*Data Preparation, Mining Operations and Presentation*>. The process of information discovery can be described as iteration over the three phases of this process.

The 1st phase: Data Preparation can be divided into two parts: Data Integration and Data Selection and Pre analysis. Data Integration is the process of merging data which resides in an operational environment having multiple files and databases. Resolving semantic ambiguities, handling missing values in data and cleaning dirty data sets are data integration issues. Issues can occur during integration which are specific to Data Mining and it deals with identifying the data required for mining and eliminating unwanted data. As a result of the Data Integration step, integrated data can be placed in a Data Warehouse. Data Selection and Pre-analysis are performed to subset the data after the integration of data. This sub-setting is done to improve the quality of mining results and to overcome limitations in current data mining products and tools.

The 2nd phase: Data mining process is the phase where the actual mining and extraction of data takes place. The Data Mining processor accesses a Data Warehouse that uses a relational database and this access can be done through a standard SQL interface. Following the completion of the second phase of Data Mining process, the *third phase* of presentation and explanation of facts and data discovered takes place.

International Journal of Innovative Research in Science and Engineering
Vol. No.2, Issue 10, October 2016
www.ijirse.com
ISSN (O) 2454-9665
ISSN (P) 2454-0663

2.4 Exploratory Data Analysis (Eda) And Predictive Modeling

The technique that is used in data mining can be known as modeling and the act of building a model in known situation and applying it to unknown situation is predictive modeling. [7] Various EDA`s associated are as mentioned below:

- Data, distributions, partitions and scales of data
- Visualizing and summarizing individual factors
- Histograms, line plots, multiples and augmentations
- transformations
- Visualizing and summarizing independence
- Visualizing and summarizing time-dependent behaviour and descriptive statistics.

Predictive modeling can be thought of as learning a mapping from an input set of vector measurements *x* to a scalar output *y*. Multi-level prediction are of 2 types. 1) Drill-down and 2) Roll-up analysis. [8] Various Prediction and manipulation of large amount of data can be solved according to given problems and obtained data is represented mainly in two ways: Numerical Data (*figure3*), Categorical Data (*figure4*)

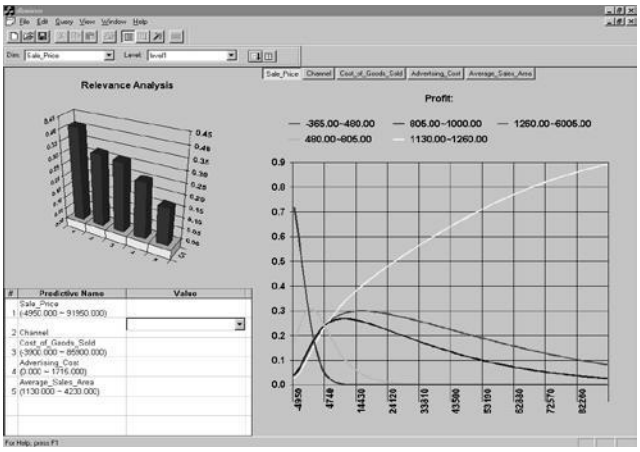

Figure3: Schematic flow of Prediction: Numerical Data.

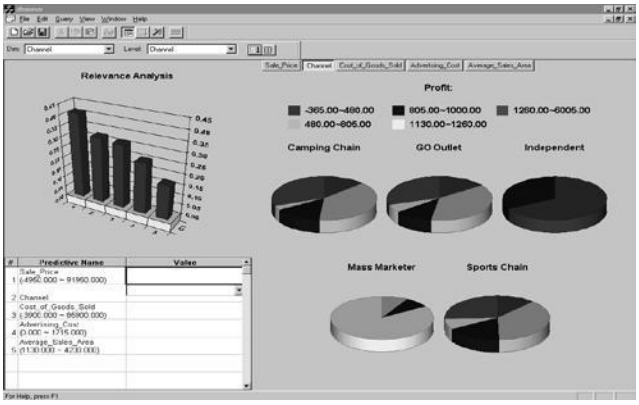

Figure4: Schematic flow of Prediction: Categorical Data.

ISSN (O) 2454-9665
ISSN (P) 2454-0663

In figure3 and figure4, the data are inputted in the WEKA software and it gives the result in numerical and categorical representation. These methods so obtained are the result of a long process of research and product development. [9] Different categorizations in data mining are prevalent as follows:

1) **Microeconomic Data Mining**
   - proposed by J. M. Kleinberg, C. H. Papadimitriou, P. Raghavan in *STOC 98* [10], evaluate data mining operation based on optimization of a profit function f(x) and market segmentation problem
2) **Capacitated location problem**
   - Similar to any SQL constraint and allow capacity and number to increase
3) **Mathematical programming**
   - Have millions of variables, need I/O devices and servers and approximation by prediction and manipulation of data.

III. DISCUSSION AND CONCLUSION

The objective of this review paper is to suggest possible solution paths to make data mining process more effective and less costly by eliminating and automating iterations as mentioned ahead. Replication, extraction, and decision support software can be used for predicting data during application of data mining which results into more precise obtained data. [11] New approach of data mining can be personalized as per specific requirements to generate the kind of information that is required for a particular application. The actual execution of a specific model or pattern can be obtained by working on process of deployment according to the specific requirement. Scalability in data mining is still an important issue for database applications: thus combining the classification of data with database and corresponding mining techniques.

In predictive data mining there occur the processes of data warehousing where data from operative systems is loaded into the warehouse for obtaining effective quality of data and to detect immediate deficiencies (raw data, unwanted data, etc.) that might become a problem during the execution of data. Special features are recommended while executing data mining in data warehouse which are as follows 1) Warehouse should be a joint quality maintenance project. 2) Implement an automated directory for information stored in warehouse. 3) Test the integrity and complex queries of data in warehouse before mining. 4) Coordinate system roll-out with network administration during data mining. Data mining and warehousing research should focus more on what happens before (exploration and modeling) and after (evaluation and deployment) the actual execution of a specific pattern-finding. These steps in the process often involve hard problems but provide interesting research opportunities which have significant potential scientific and economic impact such as parallel processing and software technology, Intensive market research, Business applications, Customer behaviour analysis, scientific discoveries and inventions. Control of Data Mining is achieved by following strategies such as dynamic system simulation, application of plan methods with the support of Artificial Intelligence (AI) and reinforcement. Large and numerous applications areas of data mining are there which includes business, retail marketing, banking, insurance and Health care, medicine, digital library, image archive, Bioinformatics, Finance and investment, Manufacturing and production, Telecommunication network, Scientific domain, World Wide Web (www), at large.

ISSN (O) 2454-9665
ISSN (P) 2454-0663

REFERENCES

[1] Sushmita Mitra, et al. Data Mining: Multimedia, Soft Computing and Bioinformatics, (2005) Wiley Publication.

[2] T. Dasu, et al. mining database structure; how to build a data quality browser, SIGMOD (2005), 240–255

[3] Michael Lloyd-Williams, Discovering the hidden secrets in your data - the data mining approach to information, "Information Research" (1997)

[4] David Hand et al. Principles of Data Mining, (2001) MIT Press Cambridge, Massachusetts

[5] S. Merugu and J. Ghosh, "Distributed Clustering Using Generative Models," Proceedings of the Third IEEE International Conference on Data Mining ICDM 2003.

[6] A. K. Jain and R. C. Dubes Algorithms for Clustering Data (2005) Prentice Hall

[7] ACM's Special Interest Group on Knowledge Discovery in Databases and their newsletter, "Explorations"

[8] Moe, W. W., and Fader, P. S. Capturing evolving visit behaviour in click stream data, (2000) Wharton School of Business, University of Pennsylvania.

[9] Agrawal, R. Data mining: Crossing the chasm. Invited Talk at the 5th ACM SIGKDD (1999) International Conference on Knowledge Discovery and Data Mining, San Diego, California.

[10] The Annual ACM Symposium on Theory of Computing (STOC), sponsored by SIGACT (the ACM Special Interest Group on Algorithms and Computation Theory), is held annually in April-May.

[11] Clark Glymour et. al Statistical Themes and lessons for Data Mining, Data Mining and Knowledge Discovery (1997) l; 11- 28, Kluwer Academic publishers.