International Journal For Technological Research In Engineering *ISSN (Online): 2347 - 4718*
3rd International Conference on Emerging Technologies in Engineering, Biomedical, Medical and Science (ETEBMS – July 17)

ADVANCED DATA MINING: A SCIENTIFIC APPROACH

Priyanka Gautam[1], Prof. Dr.Yash Pal Singh[2], Pratibha Gautam[3], Parveen Shaikh[4]
[1,2]Research Scholar, Mewar University Chittorgarh , Rajasthan
[3]Professor of Computer and Information Sciences , N.C College of Engineering, Panipat
[4]Research Scholar, Sunrise University Alwar, Rajasthan

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. Compared with other DM techniques, Intelligent Systems (ISs) based approaches, which include Artificial Neural Networks, fuzzy set theory, approximate reasoning, and Derivative-free optimization methods such as Genetic Algorithms are tolerant of imprecision, uncertainty, partial truth, and approximation. This paper is concerned with the ideas behind design; implementation, testing and application of a novel ISs based DM technique. In this paper we have focused a variety of techniques, approaches and different areas of the research which are helpful and marked as the important field of data mining Technologies. This paper imparts more number of applications of the data mining and also focuses scope of the data mining which will helpful in the further research.

Keywords: Knowledge Data mining, clustering, classification, Data mining Lifecycle and process. IS intelligent system.ANN artificial Neural network, KDD.

I. INTRODUCTION

THE amount of data being generated and stored is growing exponentially, due in leaps and bounds innovations and advances in computer technology. This presents tremendous opportunities for those who can unlock the information embedded within this data, but also introduces new challenges. In this paper we discuss how the modern field of data mining can be used to extract useful knowledge from the data that surround us. Those that can master this technology and its methods can derive great benefits and gain a competitive advantage. In this introduction we begin by discussing what data mining is, why it developed now and what challenges it faces, and may be addressed.

II. DATA MINING

Data can be analyzed, summarized, understand and meet to challenges. Data mining is a powerful concept for data analysis and process of discovery interesting pattern from the huge amount of data, data stored in various databases such as data warehouse, World Wide Web, external sources .Interesting pattern that is easy to understand, unknown, valid, potential useful. Data mining is a type of sorting technique which is actually used to extract hidden patterns from large databases. The goals of data mining are fast retrieval of data or information, knowledge Discovery from the databases, to identify hidden patterns and those patterns which are previously not explored, to reduce the level of complexity, time saving, etc. Data mining refers extracting knowledge and mining from large amount of data. Sometimes data mining treated as knowledge discovery in database (KDD) .KDD shown in Figure 1.

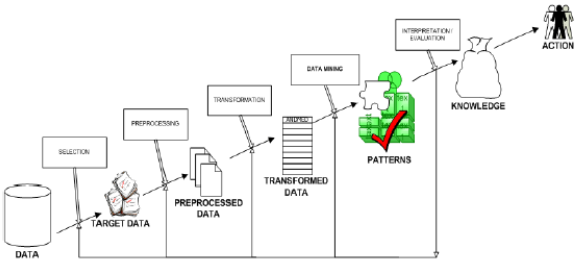


Figure 1 Knowledge data mining

- Selection: select data from various resources where operation to be performed.
- Preprocessing: also known as data cleaning in which remove the unwanted data.
- Transformation: transform /consolidate into a new format for processing.
- Data mining: identify the desire result.
- Interpretation / evaluation: interpret the result/query to give meaningful report/information

Data mining is the extraction of hidden predictive information from large databases; it is a powerful technology with great potential to help organizations focus on the most important information in their data warehouses .Data mining tools predict future trends and behaviors, helps organizations to make proactive knowledge-driven decisions . The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by prospective tools typical of decision support systems. Data mining tools can answer the questions that traditionally were too time consuming to resolve. Data mining is a new developing technology for enterprise data and information integration. It can reduce the operation cost, increase profit, and strengthen market competition of the enterprise.

III. OBJECTIVES OF STUDY

- Analytic study of data mining techniques.
- Train and test the data with different techniques.
- How to predict the unknown values i.e.analysis of output of different techniques.
- Comparing different techniques on different factors e.g. input and time taken to train and test.

IV. DATA MINING TECHNIQUES

Data mining means collecting relevant information from unstructured data. So it is able to help achieve specific objectives. 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. The purpose of a predictive model is to allow the data miner to predict an unknown value of a specific variable; the target variable. The goal of predictive and descriptive model can be achieved using a variety of data mining techniques as shown in figure 2.

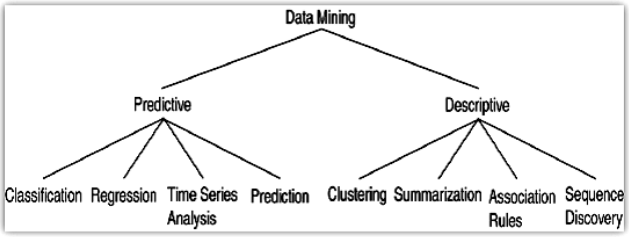


Figure 2 Data Mining Models

4.1 Data Mining Models

4.1.1 Predictive Modeling: This model permits the value of one variable to be predicted from the known values of other variables.

4.1.2 Classification: Classification based on categorical (i.e. discrete, unordered). This technique based on the supervised learning .It can be classifying the data based on the training set and values (class label). These goals are achieve using a decision tree, neural network and classification rule For example we can apply the classification rule on the past record of the student who left for university and evaluate them. Using these techniques we can easily identify the performance of the student.

4.1.3 Regression: Regression is used to map a data item to a real valued prediction variable. In other words, regression can be adapted for prediction. In the regression techniques target value are known. For example, you can predict the child behavior based on family history.

4.1.4 Time Series Analysis: Time series analysis is the process of using statistical techniques to model and explain a time-dependent series of data points. Time series forecasting is a method of using a model to generate predictions (forecasts) for future events based on known past events. For example stock market.

4.1.5 Prediction: It is one of a data mining techniques that discover the relationship between independent variables and the relationship between dependent and independent variables .Prediction model based on continuous or ordered value.

4.2 Descriptive Modeling:

It describe all the data, it includes models for overall probability distribution of the data,partitioning of the p-dimensional space into groups and models describing the relationships between the variables.

4.2.1Clustering: Clustering is a collection of similar data object. Dissimilar object is another cluster. It is way finding similarities between data according to their characteristic. This technique based on the unsupervised learning For example, image processing, pattern recognition, city planning.

4.2.2 Summarization: Summarization is abstraction of data. It is set of relevant task and gives an overview of data. For example, long distance race can be summarized total minutes, seconds and height.

4.2.3 Association Rule: Association is the most popular data mining techniques and fined most frequent item set. Association strives to discover patterns in data which are based upon elationships between items in the same transaction. Because of its nature, association is sometimes referred to as "relation technique". This method of data mining is utilized within the market based analysis in order to identify a set, or sets of products that consumers often purchase at the same time.

4.2.4 Sequence Discovery: Uncovers relationships among data. It is set of object each associated with its own timeline of events. For example, scientific experiment, natural disaster and analysis of DNA sequence.

V. ADVANCED DATA MINING PROCESS

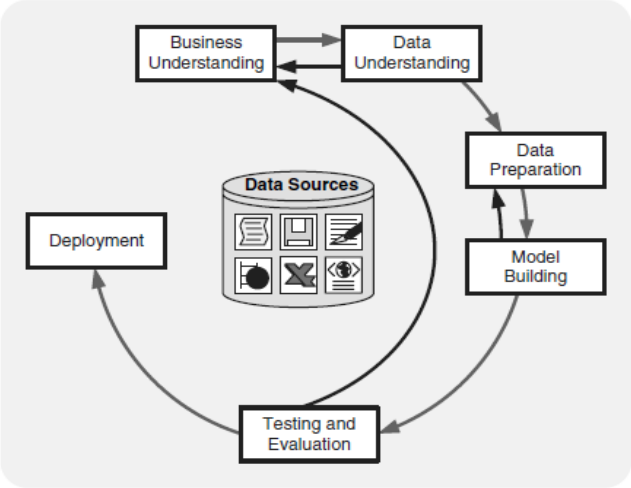
Data Mining Process ;in order to systematically conduct data mining analysis, an advance process is usually followed , two of which are described .

5.1 One (CRISP) is an industry standard process consisting of a sequence of steps that are usually involved in a data mining study. The other (SEMMA) is specific to SAS. While each step of either approach isn't needed in every analysis, this process provides a good coverage of the steps needed, starting with data exploration, data collection, data processing, analysis, inferences drawn, and implementation. CRISP-DM There is a Cross-Industry Standard Process for Data Mining (CRISP-DM) widely used by industry members. This model consists of six phases intended as a cyclical process (see Fig.):

• Business Understanding: Business understanding includes determining business objectives, assessing the current situation, establishing data mining goals, and developing a project plan.

• Data Understanding: Once business objectives and the project plan are established, data understanding considers data requirements.

• Data Preparation: Once the data resources available are identified, they need to be selected, cleaned, built into the form desired, and formatted. Data cleaning and data transformation in preparation of data modeling needs to occur in this phase. Data exploration at a greater depth can be applied during this phase, and additional models utilized, again providing the opportunity to see patterns based on business understanding.



2.1. CRISP-DM process

(Figure-3 CRISP-DM Process)

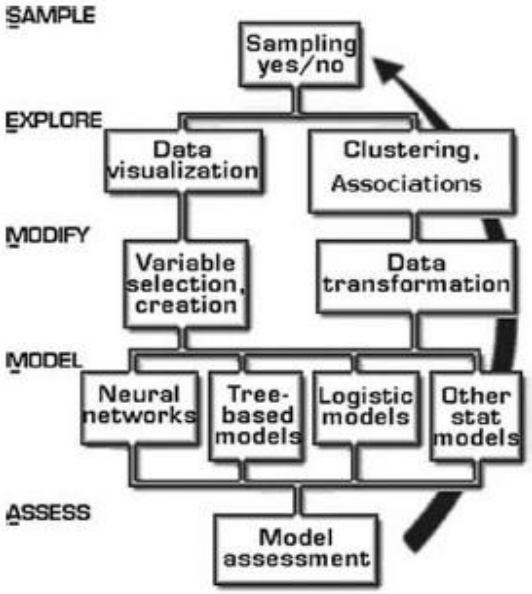
• Modeling: Data mining software tools such as visualization (plotting data and establishing relationships) and cluster analysis (to identify which variables go well together) are useful for initial analysis. Tools such as generalized rule induction can develop initial association rules. Once greater data understanding is gained (often through pattern recognition triggered by viewing model output), more detailed models appropriate to the data type can be applied. The division of data into training and test sets is also needed for modeling.

• Evaluation: Model results should be evaluated in the context of the business objectives established in the first phase (business understanding). This will lead to the identification of other needs (often through pattern recognition), frequently reverting to prior phases of CRISP-DM. Gaining business understanding is an iterative procedure in data mining, where the results of various visualization, statistical, and artificial intelligence tools show the user new relationships that provide a deeper understanding of organizational operations.

• Deployment : Data mining can be used to both verify previously held hypotheses, or for knowledge discovery (identification of unexpected and useful relationships). Through the knowledge discovered in the earlier phases of the CRISP-DM process, sound models can be obtained that may then be applied to business operations for many purposes, including prediction or identification of key situations. These models need to be monitored for changes in operating conditions

5.2 SEMMA

The acronym SEMMA stands for sample, explore, modify, model, assess. Beginning with a statistically representative sample of your data, SEMMA intends to make it easy to apply exploratory statistical and visualization techniques, select and transform the most significant predictive variables, model the variables to predict outcomes, and finally confirm a model's accuracy. A pictorial representation of SEMMA is given in Fig.. By assessing the outcome of each stage in the SEMMA process, one can determine how to model new questions raised by the previous results, and thus proceed back to the exploration phase for additional refinement of the data. That is, as is the case in CRISP-DM, SEMMA also driven by a highly iterative experimentation cycle.



(Figure-4 Steps in SEMMA Process)

VI. SCOPE OF DATA MINING:

Data mining derives its name from the similarities between searching for valuable business information in a large database for example, finding linked products in gigabytes of store scanner data and mining a mountain for a vein of valuable ore. Both processes require either sifting through an immense amount of material, or intelligently probing it to find exactly where the value resides. Given databases of efficient size and quality, data mining technology can generate new business opportunities by providing these capabilities:

6.1 Automated prediction of trends and behaviors.

Data mining automates the process of finding predictive information in large databases. Questions that traditionally required extensive hands-on analysis can now be answered directly from the data quickly. A typical example of a predictive problem is targeted marketing. Data mining uses data on past promotional mailings to identify the targets most likely to maximize return on investment in future mailings. Other predictive problems include forecasting bankruptcy and other forms of default, and identifying segments of a population likely to respond similarly to given events.

6.2 Artificial neural networks:

Non-linear predictive models that learn through training and resemble biological neural networks in structure.

6.3 Decision trees:

Tree-shaped structures that 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).

6.4Genetic algorithms:
Optimization techniques that use process such as genetic combination, mutation, and natural selection in a design based on the concepts of evolution.

6.5Nearest neighbor method:
A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k ³ 1). Sometimes called the k-nearest neighbor technique.

6.6Rule induction:
The extraction of useful if-then rules from data based on statistical significance. Many of these technologies have been in use for more than a decade in specialized analysis tools that work with relatively small volumes of data. These capabilities are now evolving to integrate directly with industry-standard data warehouse and OLAP platforms.

VII. CONCLUSION

This paper provides a general idea of data mining, data techniques and data mining in various fields. The main objectives of data mining techniques are to discover the knowledge from active data. These applications use classification, Prediction, clustering, Association techniques and so on. Hopefully in future work we review various classifications and clustering algorithm and its significance's. Data mining initially generated a great deal of excitement and press coverage, and, as is common with new "technologies", overblown expectations. However, as data mining has begun to mature as a discipline, its methods and techniques have not only proven to be useful, but have begun to be accepted by the wider community of data analysts.

REFERENCES

[1] Yongjian Fu "data mining: task, techniques and application" .
[2] Er. RimmyChuchra "Use of Data Mining Techniques for the Evaluation of Student Performance: A Case Study" International Journal of Computer Science and Management Research Vol 1 Issue 3 October 2012 .
[3] J. Han and M. Kamber."Data Mining, Concepts and Techniques", Morgan Kaufmann, 2000.
[4] AakankshaBhatnagar, Shweta P. Jadye, Madan Mohan Nagar" Data Mining Techniques & Distinct Applications: A Literature Review" International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 9, November- 2012 .
[5] Brijesh Kumar Baradwaj, Saurabh Pal "Mining Educational Data to Analyze Students Performance" (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011 .
[6] Data mining white paper, www.ikanow.com .
[7] Nikita Jain, Vishal Srivastava "DATA MINING TECHNIQUES: A SURVEY PAPER" IJRET: International Journal of Research in Engineering and Technology, Volume: 02 Issue: 11 | Nov-2013 .
[8] Dr. M.H.Dunham, "Data Mining, Introductory and Advanced Topics", Prentice Hall, 2002.
[9] Time Series Analysis and Forecasting with Weka , http://wiki.pentaho.com/display/DATAMINING/ .
[10] Umamaheswari. K, S. Niraimathi "A Study on Student Data Analysis Using Data Mining Techniques" International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 8, August 2013
[11] Industry Application of data mining , ttp://www.pearsonhighered.com.
[12] David L Olson, DursunDelen" Advance data minig techniques" springer 2008
[13] G. V. Otari, Dr. R. V. Kulkarni, "A Review of Application of Data Mining in Earthquake Prediction" G. V. Otari et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 3 (2) , 2012,3570-3574 .
[14] D Ramesh , B Vishnu Vardhan, "Data Mining Techniques and Applications to Agricultural Yield Data" International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 9, September 2013

