Journal of Advances in Science and Technology

Vol. IV, Issue No. VIII, February-2013, ISSN 2230-9659

A STUDY ON CONCEPTUAL FRAMEWORK OF DATA MINING

AN INTERNATIONALLY INDEXED PEER REVIEWED & REFEREED JOURNAL

www.ignited.in

Journal of Advances in Science and Technology
Vol. IV, Issue No. VIII, February-2013, ISSN 2230-9659

A Study on Conceptual Framework of Data Mining

Nisha Sharma[1] Dr. Chander Kant[2]

[1]Research scholar, Singhania University, Pacheri Bari, Jhunjhunu (Raj.), India

[2]Assistant Professor, Department of Computer Science and Applications, Kurukshetra University, Kurukshetra (Haryana)

Abstract – Data mining is a process which finds useful patterns from large amount of data. The research in databases and information technology has given rise to an approach to store and manipulate this precious data for further decision making. Data mining is a process of extraction of useful information and patterns from huge data. It is also called as knowledge discovery process, knowledge mining from data, knowledge extraction or data pattern analysis.

Keywords: Concept of Data mining, Data mining Techniques, Data mining algorithms, Data mining applications.

----------♦----------

INTRODUCTION TO DATA MINING

People have been collecting and organizing data from stone ages. In the earlier days data were collected and recorded in one way or the other mainly for record keeping purposes. With the advancement in computational technology in general and storage technology in particular data collection and their storage in large data warehouses have become an integral part of the data processing and decision-making environment of today's organizations. Over time people have learned to value data as an important asset. Reliable data in a database or a data warehouse could be used for decision-making purposes by appropriately analyzing the data and making them more meaningful and useful. In other words data could be analyzed to find hidden patterns and foresee trends. The process is broadly being called data mining.

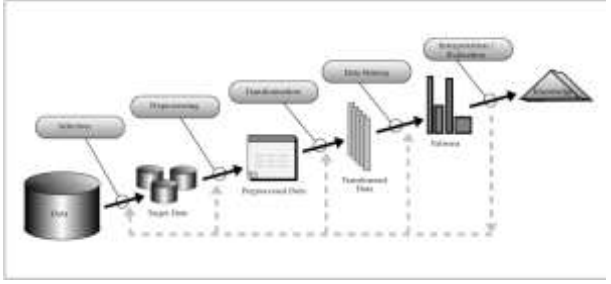
Data mining is a process of extraction of useful information and patterns from huge data. It is also called as knowledge discovery process, knowledge mining from data, knowledge extraction or data /pattern analysis.

THE EVOLUTION OF DATA MINING

Data mining is a natural development of the increased use of computerized databases to store data and provide answers to business analysts.

Evolutionary Step	Business Question	Enabling Technology
Data Collection (1960s)	"What was my total revenue in the last five years?"	computers, tapes, disks
Data Access (1980s)	"What were unit sales in New England last March?"	faster and cheaper computers with more storage, relational databases
Data Warehousing and Decision Support	"What were unit sales in New England last March? Drill down to Boston."	faster and cheaper computers with more storage, On-line analytical processing (OLAP), multidimensional databases, data warehouses
Data Mining	"What's likely to happen to Boston unit sales next month? Why?"	faster and cheaper computers with more storage, advanced computer algorithms

KNOWLEDGE DISCOVERY PROCESS



Data mining is a logical process that is used to search through large amount of data in order to find useful data. The goal of this technique is to find patterns that were previously unknown. Once these patterns are found they can further be used to make certain decisions for development of their businesses.

Three steps involved are

- Exploration
- Pattern identification
- Deployment

Exploration: In the first step of data exploration data is cleaned and transformed into another form, and important variables and then nature of data based on the problem are determined.

Pattern Identification: Once data is explored, refined and defined for the specific variables the second step is to form pattern identification. Identify and choose the patterns which make the best prediction.

Deployment: Patterns are deployed for desired outcome.

DATA MINING ALGORITHMS AND TECHNIQUES

Various algorithms and techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor method etc., are used for knowledge discovery from databases.

1. Classification

Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. Fraud detection and credit risk applications are particularly well suited to this type of analysis. This approach frequently employs decision tree or neural network-based classification algorithms. The data classification process involves learning and classification. In Learning the training data are analyzed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples. For a fraud detection application, this would include complete records of both fraudulent and valid activities determined on a record-by-record basis.

The classifier-training algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called a classifier.

Types of classification models:

Classification by decision tree induction

Bayesian Classification

Neural Networks

Support Vector Machines (SVM)

Classification Based on Associations

2. Clustering

Clustering can be said as identification of similar classes of objects. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes. Classification approach can also be used for effective means of distinguishing groups or classes of object but it becomes costly so clustering can be used as preprocessing approach for attribute subset selection and classification. For example, to form group of customers based on purchasing patterns, to categories genes with similar functionality.

Types of clustering methods

- Partitioning Methods
- Hierarchical Agglomerative (divisive) methods
- Density based methods
- Grid-based methods
- Model-based methods

3. Predication

Regression technique can be adapted for predication. Regression analysis can be used to model the relationship between one or more independent variables and dependent variables. In data mining independent variables are attributes already known and response variables are what we want to predict.

Unfortunately, many real-world problems are not simply prediction. For instance, sales volumes, stock prices, and product failure rates are all very difficult to predict because they may depend on complex interactions of multiple predictor variables. Therefore, more complex techniques (e.g., logistic regression, decision trees, or neural nets) may be necessary to forecast future values. The same model types can often be used for both regression and classification. For example, the CART (Classification and Regression Trees) decision tree algorithm can be used to build both classification trees (to classify categorical response variables) and regression trees (to forecast continuous response variables). Neural

networks too can create both classification and regression models.

Types of regression methods

- Linear Regression
- Multivariate Linear Regression
- Nonlinear Regression
- Multivariate Nonlinear Regression

4. Association rule

Association and correlation is usually to find frequent item set findings among large data sets. This type of finding helps businesses to make certain decisions, such as catalogue design, cross marketing and customer shopping behavior analysis. Association Rule algorithms need to be able to generate rules with confidence values less than one. However the number of possible Association Rules for a given dataset is generally very large and a high proportion of the rules are usually of little (if any) value.

Types of association rule

- Multilevel association rule
- Multidimensional association rule
- Quantitative association rule

5. Neural networks

Neural network is a set of connected input/output units and each connection has a weight present with it. During the learning phase, network learns by adjusting weights so as to be able to predict the correct class labels of the input tuples. Neural networks have the remarkable ability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. These are well suited for continuous valued inputs and outputs. For example handwritten character reorganization, for training a computer to pronounce English text and many real world business problems and have already been successfully applied in many industries.

Neural networks are best at identifying patterns or trends in data and well suited for prediction or forecasting needs.

Types of Neural Networks

- Back Propagation

DATA MINING APPLICATIONS

Data mining is a relatively new technology that has not fully matured. Despite this, there are a number of industries that are already using it on a regular basis. Some of these organizations include retail stores, hospitals, banks, and insurance companies. Many of these organizations are combining data mining with such things as statistics, pattern recognition, and other important tools. Data mining can be used to find patterns and connections that would otherwise be difficult to find. This technology is popular with many businesses because it allows them to learn more about their customers and make smart marketing decisions.

CONCLUSION

Data mining has importance regarding finding the patterns, forecasting, and discovery of knowledge etc., in different business domains. Data mining techniques and algorithms such as classification, clustering etc., helps in finding the patterns to decide upon the future trends in businesses to grow. 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.

REFERENCES

1. Clifton, Christopher (2010). "Encyclopedia Britannica: Definition of Data Mining". Retrieved 2010-12-09.

2. Crisp-DM 1.0 Step by step Data Mining guide from http://www.crisp-dm.org/CRISPWP-0800.pdf.

3. Dr. Gary Parker, vol 7, 2004, Data Mining: Modules in emerging fields, CD-ROM.

4. Fayyad, U.M., Piatetsky-Shapiro, G. and Smyth, P. Advances in knowledge discovery and data mining. *Data Mining to Knowledge Discovery: An Overview*, 1-34, AAAI/MIT Press, 1996.

5. Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2009). "The Elements of Statistical Learning: Data Mining, Inference, and Prediction". Retrieved 2012-08-07.

6. Jiawei Han and Micheline Kamber (2006), Data Mining Concepts and Techniques, published by Morgan Kauffman, 2nd ed.

7. Kantardzic, Mehmed (2003). *Data Mining: Concepts, Models, Methods, and Algorithms*. John Wiley & Sons. ISBN 0-471-22852-4. OCLC 50055336

8. Sowa, J.F. *Conceptual Structures, Information Processing in Mind and Machine*, Addison-Wesley, Reading, Massachusetts, 1984.