**DATA MINING TECHNIQUES AND APPLICATIONS**

Gazi Monirul Islam (151-15-4716)  
*Department of of CSE*  
*Daffodil International University* - Dhaka, Bangladesh  
[monirul15-4716@diu.edu.bd](mailto:monirul15-4716@diu.edu.bd)

***Abstract***

Data mining is a process which finds useful patterns from large amount of data. The paper discusses few of the data mining techniques, algorithms and some of the organizations which have adapted data mining technology to improve their businesses and found excellent results and the applications of data mining in general.

***Keywords:*** Data mining Techniques; Data mining algorithms; Data mining applications.

**1. Overview of Data Mining**

The development of Information Technology has generated large amounts of databases and huge data in various areas. 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. It helps to find information, pattern, logic and predict future events by analyzing previous record. There are 3 major steps are

* Exploration
* Pattern Identification
* Deployment

Exploration: In the first step of data exploration data is cleaned and transformed into another form and it has to be ready to be analyze friendly and important variables are kept to do our next step properly.

Pattern Identification: Once our data is explored, refined and defined for our specific purpose then second step is to find our desired pattern to find the best prediction possible.

Deployment: Patterns are deployed after we find the best prediction for particular topic that we want and get it as our desired output.

**2. Data Mining Algorithms and Techniques**

There are lots of types of algorithms for data mining purpose and every single one of them is good for different purposes. Some are good for Prediction, some are good for Identify something, some are good for categorize products and objects etc. But we all have to know the rules how to do all of them and apply them to find the best possible result we want from them and then we apply it in general purposes. They are:

**2.1 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:

* Decision Tree based classification
* Bayesian Classification
* Neural Networks
* Support Vector Machines (SVM)
* Association based classification
* Quantitative Association rule

**2.2 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

**2.3 Prediction Based**

For predictive analysis we use Regression techniques. It is used to model the relationship between one or more independent variable and dependent variables. In data mining independent variable are attributes that we already know and the response variable are what we want to predict. But some real-life instances they are not enough to be predicted by only regression rather more complex techniques like logistic regression, decision tree, neural network etc.



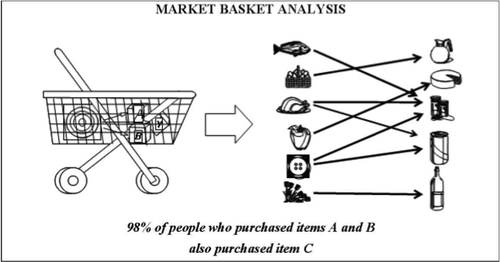
Fig: linear regression

Types of regression methods

* Linear regression
* Multivariate Linear Regression
* Nonlinear Regression
* Multivariate nonlinear regression

**2.4 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. Let’s see the image first to get some idea about Market analysis system and how to arrange products in customer’s buying test and set the related products according to the solution and relation of association.



Types of Association methods

* Multilevel Association Rule
* Multidimensional Association rule
* Quantitative association rule

**2.5 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 Association methods

* Back Propagation

**3. Major Components of Data Mining**

* Extract, transform and load transaction data into the data warehouse system
* Store and manage the data in a multi-dimensional database system
* Provide data access to business analysts and IT Professionals
* Analyze the data by application software
* Present the data in a useful format, such as a graph or table to visualize

**4. Predictive Model**

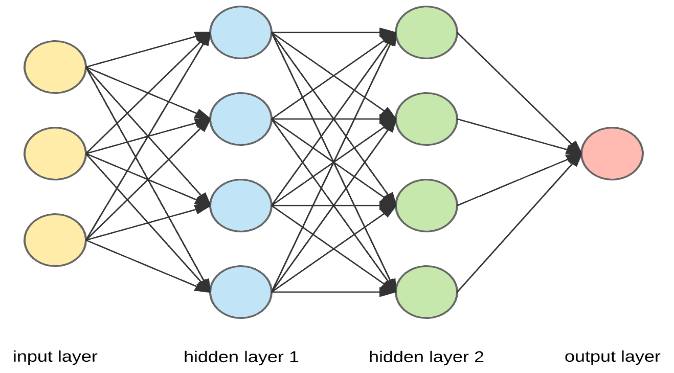
The purpose of a predictive model is to allow the data miner to predict and unknown value and it is often called as future value of a specific variable the target variable. If the target value is one of the predefined numbers of discrete (class) level, the data mining task is called classification. If the target value is a real number then the task is regression. The predictive model is thus created from given known values of variables, possibly including previous values of the target variable. The training data consists of pairs of measurements. Each consisting of an input vector x (i) with a corresponding target value y(i). The predictive model is an estimation of the function y = f(x: q) able to predict a value y, given and input vector pf measured value of x and a set of estimated parameters q for the model f. The process of finding the best q values is the core of the data mining technique.

At the core of the data mining process is the use of a data mining technique. Some data mining techniques directly obtain the information by performing a descriptive partitioning of the data. More often, however, data mining techniques utilize stored data in order to build predictive models. From a general perspective, there is strong agreement among both researchers and executives about the criteria that all data mining techniques must meet. Most importantly, the techniques should have high performance. This criterion is, for predictive modelling, understood to mean that the technique should produce models that will generalize well, i.e. models having high accuracy when performing predictions based on novel data.

Classification and prediction are two forms of data analysis that can be used to extract models describing the important data classes or to predict the future data trends. Such analysis can help to provide us with a better understanding of the data at large. The classification predicts categorical (discrete, unordered) labels, prediction model, and continuous valued function.

**5. Neural Network in depth**

Neural Network or an artificial neural network is a biological system that detects patterns and makes predictions. The greatest breakthroughs in neural network in recent years are in their application to real world problems like customer response prediction, fraud detection etc. Data mining techniques such as neural networks are able to model the relationships that exist in data collections and can therefore be used for increasing business intelligence across a variety of business applications [4]. This powerful predictive modelling technique creates very complex models that are really difficult to understand by even experts. Neural Networks are used in a variety of applications. It is shown in fig.1. Artificial neural network have become a powerful tool in tasks like pattern recognition, decision problem or predication applications. It is one of the newest signals processing technology. ANN is an adaptive, non linear system that learns to perform a function from data and that adaptive phase is normally training phase where system parameter is change during operations. After the training is complete the parameter are fixed. If there are lots of data and problem is poorly understandable then using ANN model is accurate, the non linear characteristics of ANN provide it lots of flexibility to achieve input output map. Artificial Neural Networks, provide user the capabilities to select the network topology, performance parameter, learning rule and stopping criteria.

– Fig:2 Neural Network

**6. Decision trees**

A decision tree is a flow chart like structure where each node denotes a test on an attribute value, each branch represents an outcome of the test and tree leaves represent classes or class distribution. A decision tree is a predictive model most often used for classification. Decision trees partition the input space into cells where each cell belongs to one class. The partitioning is represented as a sequence of tests. Each interior node in the decision tree tests the value of some input variable, and the branches from the node are labelled with the possible results of the test. The leaf nodes represent the cells and specify the class to return if that leaf node is reached. The classification of a specific input instance is thus performed by starting at the root node and, depending on the results of the tests, following the appropriate branches until a leaf node is reached [5].Decision tree is represented in figure 3

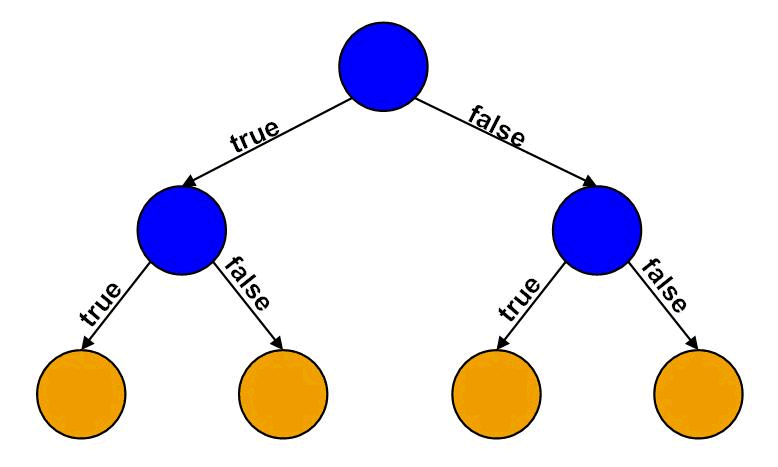


Fig:3Decision tree

Decision tree is a predictive model that can be viewed as a tree where each branch of the tree is a classification question and leaves represent the partition of the data set with their classification. The author defines a Decision Tree as a schematic tree-shaped diagram used to determine a course of action or show a statistical probability. Decision trees can be viewed from the business perspective as creating a segmentation of the original data set. Thus, marketing managers make use of segmentation of customers, products and sales region for predictive study. These predictive segments derived from the decision tree also come with a description of the characteristics that define the predictive segment. Because of their tree structure and skill to easily generate rules the method is a favoured technique for building understandable models.

**7. Genetic Algorithm**

Genetic Algorithm attempt to incorporate ideas of natural evaluation The general idea behind GAs is that we can build a better solution if we somehow combine the "good" parts of other solutions (schemata theory), just like nature does by combining the DNA of living beings [7]. Genetic Algorithm is basically used as a problem solving strategy in order to provide with an optimal solution. They are the best way to solve the problem for which little is known. They will work well in any search space because they form a very general algorithm. The only thing to be known is what the particular situation is where the solution performs very well, and a genetic algorithm will generate a high quality solution. Genetic algorithms use the principles of selection and evolution to produce several solutions to a given problem. Its fig. 4

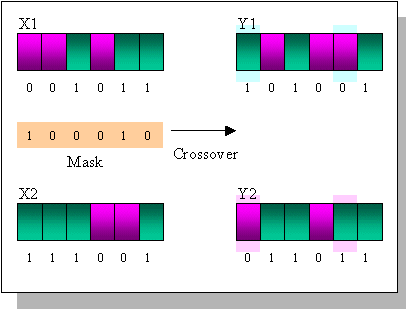


Fig4: Genetic Algorithm

Genetic algorithms (GAs) [8] are based on a biological application. It depends on theory of evolution. When GAs are used for problem solving, the solution has three distinct stages:

* The solutions of the problem are encoded into representations that support the necessary variation and selection operations; these representations, are called chromosomes, are as simple as bit strings
* A fitness function judges which solutions are the “best” life forms, that is, most appropriate for the solution of the particular problem. These individuals are favoured in survival and reproduction, thus giving rise to generation.
* Crossover and mutation produce a new generation of individuals by recombining features of their parents. Eventually a generation of individuals will be interpreted back to the original problem domain and the fit individual represents the solution

**8.Rule Extraction**

For the rules to be useful there are two pieces of information that must be supplied as well as the actual rule:

* Accuracy- How often is the rule correct
* Coverage- How often does this rule apply

Only because the pattern in the data base is expressed as rule, it does not mean that it is true always. So like data mining algorithms it is equally important to identify and make obvious the uncertainty in the rule. This is called accuracy. The coverage of the rule means how much of the database it “covers” or applies to. There are 5 criteria for rule extraction they are as follows:-

* **Comprehensibility**: The extent to which extracted representations are humanly comprehensible.
* **Fidelity**: The extent to which extracted representations accurately model the network from which they were extracted.
* **Accuracy**: The ability of extracted representations to make accurate predictions on previously unseen cases.
* **Scalability**: The ability of the method to scale to networks with large input spaces and large number of weighted connections.
* **Generality**: The extent to which the method requires special training

**9. 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. Here is overview of business problems and solutions found using data mining technology.

**9.1 FBTO Dutch Insurance Company**

Challenges

* To reduce direct mail costs
* Increase efficiency of marketing campaigns
* Increase cross-selling to existing customer, using inbound channels such as company’s sell center and the internet a one tear test of the solution’s effectiveness

Results

* Provided the marketing team with the ability to predict the effectiveness of its campaign
* Increased the efficiency of marketing campaign creation, optimization, and execution.
* Decreased mailing cost by 35 percent.
* Increased conversion rates by 40 percent

**9.2 ECtel Ltd, Israel**

Challenges

* Fraudulent activity in telecommunication services

Results

* Significantly reduced telecommunications fraud for more than 150 telecommunication companies worldwide.
* Saved money by enabling real-time fraud detection.

**9.3 Provident Financial Home Credit Division, United Kingdom**

Challenges

* No System to detect and prevent fraud.

Results

* Reduced frequency and magnitude of agent and customer fraud.
* Saved money through early fraud detection
* Saved investigator’s time and increased prosecution rate.

**9.4 Standard Life Mutual Financial Services Companies**

Challenges

* Identify the key attributes of clients attracted to their mortgage offer.
* Cross sell Standard Life Bank products to the clients of other Standard Life companies.
* Develop a remortgage model which could be deployed on the group Web site to examine the profitability of the mortgage business being accepted by Standard Life Bank.

Results

* Built a propensity model for the Standard Life Bank mortgage offer identifying key customer types that can be applied across the whole group prospect pool. Saved money through early f
* Discovered the key drivers for purchasing a remortgage product.
* Achieved, with the model, a nine times greater response than that achieved by the control group.
* Secured £33million (approx. $47 million) worth of mortgage application revenue.

**9.5 Shenandoah Life insurance company United States**

Challenges

* Policy approval process was paper based and cumbersome
* Routing of these paper copies to various departments, there was delays in approval.

Results

* Empowered management with current information on pending policies.
* Reduced the time required to issue certain policies by 20 percent.
* Improved underwriting and employee performance review processes.

**Conclusion:**

Data mining has importance regarding finding the patterns, forecasting, 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] Jiawei Han and Micheline Kamber (2006), Data Mining Concepts and Techniques, published by Morgan Kauffman, 2nd ed.

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

[3] Crisp-DM 1.0 Step by step Data Mining guide from <http://www.crisp-dm.org/CRISPWP-0800.pdf>.

[4] Customer Successes in your industry from <http://www.spss.com/success/?source=homepage&hpzone=nav_bar>.

[5] https://www.allbusiness.com/Technology /computer-software-data-management/ 633425-1.html, last retrieved on 15th Aug 2010. 6. <http://www.kdnuggets.com/>.

[6] Xingquan Zhu, Ian Davidson, “Knowledge Discovery and Data Mining: Challenges and Realities”, ISBN 978- 1-59904-252, Hershey, New York, 2007.

[7] Joseph, Zernik, “Data Mining as a Civic Duty – Online Public Prisoners Registration Systems”, International Journal on Social Media: Monitoring, Measurement, Mining, vol. - 1, no.-1, pp. 84-96, September2010.

[8] Zhao, Kaidi and Liu, Bing, Tirpark, Thomas M. and Weimin, Xiao,”A Visual Data Mining Framework for Convenient Identification of Useful Knowledge”, ICDM '05 Proceedings of the Fifth IEEE International Conference on Data Mining, vol.-1, no.-1,pp.- 530- 537,Dec 2005.

[9] R. Andrews, J. Diederich, A. B. Tickle,” A survey and critique of techniques for extracting rules from trained artificial neural networks”, Knowledge-Based Systems,vol.- 8,no.-6, pp.-378-389,1995.

[10] Lior Rokach and Oded Maimon,“Data Mining with Decision Trees: Theory and Applications(Series in Machine Perception and Artificial Intelligence)”, ISBN: 981-2771-719, World Scientific Publishing Company, , 2008.

[11] Venkatadri.M and Lokanatha C. Reddy ,“A comparative study on decision tree classification algorithm in data mining” , International Journal Of Computer Applications In Engineering ,Technology And Sciences (IJCAETS), Vol.- 2 ,no.- 2 , pp. 24- 29 , Sept 2010.

[12] AnkitaAgarwal,”Secret Key Encryption algorithm using genetic algorithm”, vol.-2, no.-4, ISSN: 2277 128X, IJARCSSE, pp. 57-61, April 2012.

[13] Li Lin, Longbing Cao, Jiaqi Wang, Chengqi Zhang, “The Applications of Genetic Algorithms in Stock Market Data Mining Optimisation”, Proceedings of Fifth International Conference on Data Mining, Text Mining and their Business Applications,pp- 593-604,sept 2005.

[14] Fu Xiuju and Lipo Wang “Rule Extraction from an RBF Classifier Based on Class-Dependent Features “, ISNN’05 Proceedings of the Second international conference on Advances in Neural Networks ,vol.-1,pp.- 682-687,2005

[15] H. Johan, B. Bart and V. Jan, “Using Rule Extraction to Improve the Comprehensibility of Predictive Models”. In Open Access publication from Katholieke Universiteit Leuven, pp.1-56, 2006

[16] M. Craven and J. Shavlik, “Learning rules using ANN “, Proceeding of 10th International Conference on Machine Learning, pp.-73-80, July 1993