DATA MINING TECHNIQUES, APPLICATIONS AND CHALLENGES

Surajit Mohanty*, Sameer Kumar Das

***Corresponding author: mohanty.surajit@gmail.com**

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. Data mining applications have benefited the healthcare industry in terms of fraud and abuse detection by insurers, use in customer relationship management decisions by healthcare organizations and identification of effective treatments and best practices by physicians. The enormous data generated by healthcare transactions cannot be properly examined and practiced using traditional methods. Data mining presents the techniques and tools used in converting large healthcare data into useful information for decision-making. This study discussed data mining applications in healthcare areas such as the healthcare detection of fraud and abuse, appraisal of efficiency in treatment, hospital infection control, healthcare management, customer relationship management and identification of high-risk patients.

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

1. Overview of Data Mining

The development of Information Technology has generated large amount 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.[1]

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.[2]

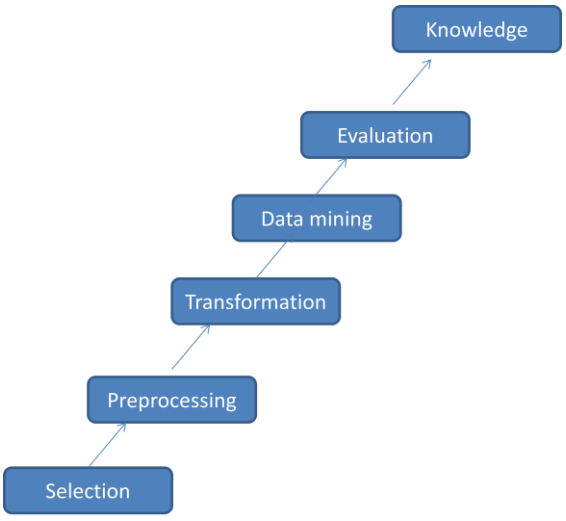

Fig 1: Overview of Data Mining

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.[3]
- Deployment: Patterns are deployed for desired outcome.

2. 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.

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.[5] 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.[4]

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

2.2. 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 pre-processing approach for attribute subset selection and classification.[6] 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. 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.[6]

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.

Types of association rule

- ➢ 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[7]. 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

3. 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.[8]

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

3.1. FBTO Dutch Insurance Company

Challenges

- ❖ To reduce direct mail costs.
- ❖ Increase efficiency of marketing campaigns.
- ❖ Increase cross-selling to existing customers, using inbound channels such as the company's sell centre.

Results

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

3.2. 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.

3.3. Provident Financial's 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.

3.4 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.[8]
- 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.

3.5. 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.

4. 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] Kevin McDonald, Andreas Wilmsmeier, David C. Dixon, W.H.Inmon:" Mastering SAP Business Information Warehouse", Wiley Publishing Inc., 2012.

[2] U. K. Pandey, and S. Pal, "A Data mining view on class room teaching language", (IJCSI) International Journal of Computer Science Issue, Vol. 8, Issue 2, pp. 277-282, ISSN:1694-0814, 2011.

[3]. Shaeela Ayesha, Tasleem Mustafa, Ahsan Raza Sattar, M. Inayat Khan, "Data mining model for higher education system", Europen Journal of Scientific Research, Vol.43, No.1, pp.24-29, 2010

[4]. Jayanthi Ranjan " Data mining in pharma sector:

Benefits" Information Management and Technology Area,Institute of Management Technology, Ghaziabad, India,28 April 2008

[5] M. Bray, The shadow education system: private tutoring and its implications for planners, (2nd ed.), UNESCO, PARIS, France, 2007

[6] Zuckerman and Alan, M. (2006)"Healthcare Strategic Planning", Prentice Hall of India.

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

[8] Adriaans Peiter, Zantinge Dolf. (2005) "Data Mining" Pearson Education, pp.6971.

Biographical Notes

Mr. Surajit Mohanty, Assistant professor in Dept. of CSE, DRIEMS, Tangi, Cuttack. He has 11 year industry and teaching experience. He completed his M-tech in the year 2010.His area of interest is in SAP, ERP, Data mining and BIG data analysis. He has published 12 paper in national and international journal.

Mr. Sameer Kumar Das working as an Asst. professor in CSE at GATE, Berhampur completed his M.Tech in computer science and engineering 2011. His area of interest is in computer networking and BIG data analysis.