International Journal of Enhanced Research in Management & Computer Applications, ISSN: 2319-7471
Vol. 2 Issue 8, Sept.-Oct., 2013, pp: (24-29), Available online at: www.erpublications.com

Data Mining Process and Its Applications for Knowledge Discovery

D. Kiran Kumar[1], Dr. Sudhir Dawra[2], Dr. K. Bhargavi[3]

[1]Research Scholar, Sun Rise University, IET Group, Alwar
[2]Professor in CSE, Sun Rise University, IET Group, Alwar
[3]Professor in CSE, Palamur University, Andhra Pradesh.

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. It is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.

Keywords: ***Data mining process; Data mining algorithms; Data mining applications, KDD, Patterns design.***

.

---

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.

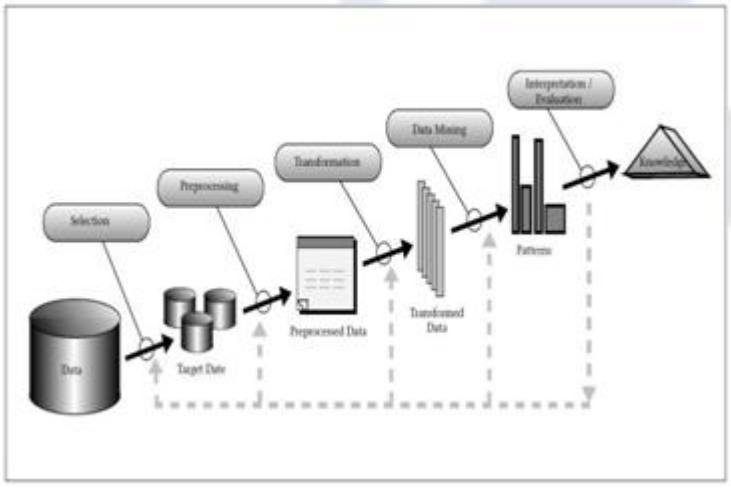

Figure 1. Knowledge discovery Process

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.

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.

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

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

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

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

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

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

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 center and the internet a one year test of the solution's effectiveness.

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.

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.

Provident Financials 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.

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.

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.

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.

Soft map Company Ltd., Tokyo

Challenges

Customers had difficulty making hardware and software purchasing decisions, which was hindering online sales.

Results

Page views increased 67 percent per month after the recommendation engine went live.

Profits tripled in 2001, as sales increased 18 percent versus the same period in the previous year.

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]. M. Halkidi, D. Spinellis, G. Tsatsaronis *et al.*, "Data mining in software engineering," *Intelligent Data Analysis*, vol. 15, no. 3, pp. 413-441, 2011.
[2]. A. E. Hassan, and R. C. Holt, "Predicting change propagation in software systems," in *Proceedings of the 20th IEEE International Conference on Software Maintenance*, 2004, pp. 284-293.
[3]. Chaturvedi K.K, Singh V.B, Singh P, "Tools in Mining Software Repositories", 13th International Conference on Computational Science and Its Applications, pp. 89-98, 2013
[4]. J. Huffman Hayes, A. Dekhtyar and J. Osborne, Improving requirements tracing via information retrieval. In *Proceedings of the International Conference on Requirements Engineering*, 2003.
[5]. J. Huffman Hayes, A. Dekhtyar and S. Sundaram, Text mining for software engineering: How analyst feedback impacts final results. In *Proceedings of International Workshop on Mining Software Repositories* (*MSR*), 2005.
[6]. D. German and A. Mockus, Automating the measurement of open source projects. In *Proceedings of the 3rd Workshop on Open Source Software Engineering, 25th International Conference on Software Engineering* (*ICSE03*), 2003.
[7]. C. Jensen andW. Scacchi, Datamining for software process discovery in open source software development communities. In *Proceedings of International Workshop on Mining Software Repositories* (*MSR*), 2004.
[8]. C.C.Williams and J.K.Hollingsworth,Automatingmining of source code repositories to improve bug finding techniques, *IEEE Transactions on Software Engineering* **31**(6) (2005), 466–480.
[9]. S.Morisaki,A.Monden andT.Matsumura, Defect data analysis based on extended association rulemining. In*Proceedings of International Workshop on Mining Software Repositories* (*MSR*), 2007.
[10]. R Chang, A. Podgurski and J. Yang, Discovering neglected conditions in software by mining dependence graphs, *IEEE Transactions on Software Engineering*, 2008.
[11]. W. Dickinson, D. Leon and A. Podgurski, Finding failures by cluster analysis of execution profiles, *International Conference on Software Engineering* (*ICSE*), 2001.
[12]. M. Last,M. Friedman and A. Kandel, *The Data Dimining Approach to Automated Software Testing*, In Proceeding of the SIGKDD Conference, 2005.
[13]. J. Bowring, J. Rehg and M.J. Harrold, Acive learning for automatic classification of software behavior, *International Symposium on Software Testing and Analysis* (*ISSTA*), 2004.
[14]. C. Liu, X Yan, and J. Han. Mining control ow abnormality for logical errors. In *Proceedings of SIAM Data Mining Conference* (*SDM*), 2006.

[15]. C. Liu, X. Yan, H. Yu, J. Han and P. Yu, Mining behavior graphs for 'backtrace' of noncrasinh bugs. In *SIAM Data Mining Conference* (*SDM*), 2005.
[16]. Y. Kannelopoulos, Y. Dimopoulos, C. Tjortjis and C. Makris, Mining source code elements for comprehending object oriented systems and evaluating their maintainability, *SIGKDD Explorations* **8**(1), 2006.
[17]. D. Engler, D. Chen, S. Hallem *et al.*, "Bugs as deviant behavior: A general approach to inferring errors in systems code," *ACM SIGOPS Operating Systems Review*, vol. 35, no. 5, pp. 57-72, 2001.
[18]. Z. Li, and Y. Zhou, "*PR-Miner: Automatically extracting implicit programming rules and detecting violations in large software code,*" in Proceedings of the 10th European software engineering conference held jointly with 13th ACM SIGSOFT international symposium on Foundations of software engineering, 2005, pp. 306-315.
[19]. S. Lu, S. Park, C. Hu *et al.*, *"MUVI: automatically inferring multi-variable access correlations and detecting related semantic and concurrency bugs,"* ACM SIGOPS Operating Systems Review, vol. 41, no. 6, pp. 103-116, 2007.
[20]. B. Baker, "On finding duplication and near-duplication in large software systems," in *Second IEEE Working Conf on Reverse Eng.(wcre)*, 1995, pp. 86-95.
[21]. T. Kamiya, S. Kusumoto, and K. Inoue, "CCFinder: a multilinguistic token-based code clone detection system for large scale source code," *IEEE Transactions on Software Engineering*, pp. 654-670, 2002.
[22]. V. Wahler, D. Seipel, J. Wolff *et al.*, "Clone detection in source code by frequent itemset techniques," in Fourth IEEE International Workshop on Source Code Analysis and Manipulation, 2004, pp. 128-135.
[23]. W. Qu, Y. Jia, and M. Jiang, "Pattern mining of cloned codes in software systems," *Information Sciences*, 2010.
[24]. H. A. Basit, and S. Jarzabek, "A data mining approach for detecting higher-level clones in software," *IEEE Transactions on Software Engineering*, pp. 497-514, 2009.
[25]. Z. Li, S. Lu, S. Myagmar *et al.*, "CP-Miner: A tool for finding copy-paste and related bugs in operating system code," in *Proceedings of the 6th conference on Symposium on Operating Systems Design & Implementation*, 2004, pp. 20.
[26]. A. Michail, "Data mining library reuse patterns using generalized association rules," in *Proceedings of 22nd International Conference on Software Engineering (ICSE'00)*, Limerick, Ireland, 2000, pp. 167-176.
[27]. N. Sahavechaphan, and K. Claypool, "XSnippet: mining for sample code," *ACM SIGPLAN Notices*, vol. 41, no. 10, pp. 413-430, 2006.
[28]. T. Xie, and J. Pei, "MAPO: Mining API usages from open source repositories," in *Proceedings of the international workshop on Mining software repositories*, 2006, pp. 54-57.
[29]. T. Zimmermann, P. Weisgerber, S. Diehl *et al.*, "Mining version histories to guide software changes," *IEEE Transactions on Software Engineering,31(6)*, pp. 429-445, June 2005.
[30]. A. E. Hassan, and R. C. Holt, "Predicting change propagation in software systems," in *Proceedings of the 20th IEEE International Conference on Software Maintenance*, 2004, pp. 284-293.