DATA MINING TECHNIQUES, TOOLS AND APPLICATIONS

Ms. Surbhi Sharma, Mrs. Ritu Tailor, Department of Computer Application, Assistant Professor of Geetanjali Institute of Technical Studies, Dabok, Udaipur (Raj)
Dr. Rajesh Kanja, Assistant Professor, Faculty of Computer Science, Pacific University, Udaipur (Raj)

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.

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

1. Overview of Data Mining

The evolution of Information Technology has generated large amount of databases and huge data in various areas. Data mining is one of the most useful techniques that help entrepreneurs, researchers, and individuals to extract valuable information from huge sets of data. Data mining is the process of sorting through large data sets to identify patterns and relationships that can help solve business problems through data analysis. Data mining techniques and tools enable enterprises to predict future trends and make more-informed business decisions. Data mining is also called Knowledge Discovery in Database (KDD). The knowledge discovery process includes Data cleaning, Data integration, Data selection, Data transformation, Data mining, Pattern evaluation, and Knowledge presentation.

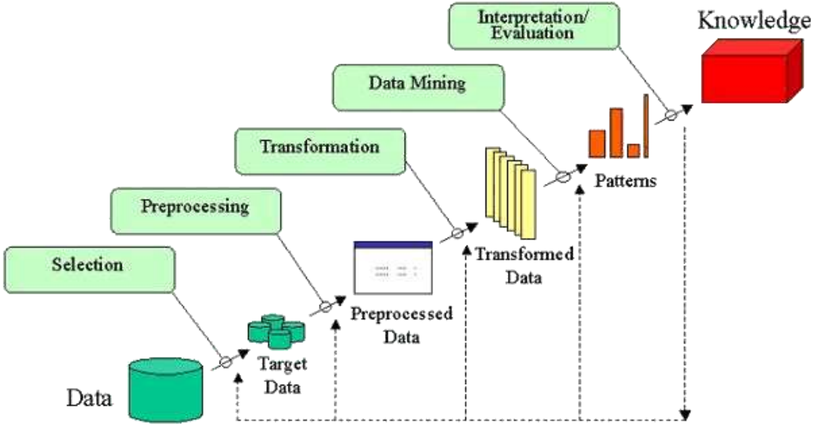

Figure1.KnowledgediscoveryProcess

Data Mining is a process used by organizations to extract specific data from huge databases to solve business problems. It primarily turns raw data into useful information. Data Mining is the process of investigating hidden patterns of information to various perspectives for categorization into useful data, which is collected and assembled in particular areas such as data warehouses, efficient analysis, data mining algorithm, helping decision making and other data requirement to eventually cost-cutting and generating revenue.[1]

Three steps involved are

- Data Exploration
- Pattern identification
- Deployment

- **Data Exploration:** Data exploration is the first step of data mining. It is a process in which data analysts clean and transform data and use various data visualization techniques to extract important variables. This step is also essential to understand the nature and characteristics of data. It helps analysts visualize data and classify variables before extracting relevant data for analysis
- **Pattern Identification –**

Once data analysts comprehensively view data through exploration,they use automated techniques to classify data further. This is done through pattern identification. Pattern identification is a process that helps identify important data trends, which help organizations prepare strategies to enhance their growth. Analyzing new trends and identifying patterns also allows organizations to make future predictions.

- **Deployment:**

Deployment is the final stage of data mining. It involves presenting and making use of data mining results. At last, Patterns are deployed for the desired outcome.

1. Data Mining Techniques

In data mining , various major data mining techniques have been developed and used, including 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]

1.1. Classification

Classification is the most commonly applied data mining technique which is used to assign items into predefined classes based on the values of their attributes. Classification involves the use of labeled training data, which the algorithm uses to build a model that can then be used to classify new items. By classifying items into predefined groups, the classification algorithm can help identify patterns and trends in the dataset that may not have been otherwise notified. 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

1.2. Clustering

Clustering is a data mining technique that does not require labeled data. Instead, clustering uses similarity measures between different data when collecting them. Clustering is often used for exploratory data analysis to find hidden patterns or collect data. It can also be used for division, which is the process of dividing a dataset into groups based on similarities. For example, to form group of customers based on purchasing patterns, to categories genes with similar functionality. [4]

Types of clustering methods

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

1.3. Regression

Regression is another way to find relationships in data sets, by calculating predicted data values based on a set of variables. Linear regression and multivariate regression are examples. Decision trees and some other classification methods can be used to do regressions, too. Neural networks too can create both classification and regression models. [5]

Types of regression methods

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

1.4. Association rule

Association rules are if-then statements that determine the possibility of interactions between data items within large data sets in various types of databases. This data mining technique helps to identify a link between two or more items. It finds an unknown pattern in the data set. The algorithm processes various data sets, For example, a list of grocery or daily expenditures to measure the percentage of items being purchased together.[6]

Types of association rule

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

1.5. Neural networks

Neural networks are a type of machine learning algorithm that is modeled after the human brain. They are designed to recognize patterns in data by using a network of interconnected nodes, or neurons. Neural networks are often used in data mining to classify data, make predictions, and identify relationships between variables.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. Data mining Tools widely used in industry.[7]

B. Rapid Miner:

It is open source data mining tools. Rapid Miner is one of the best predictive analysis system developed by the company with the same name as the Rapid Miner. It is written in Java programming language. It provides an integrated environment for deep learning & predictive analysis.This free data mining software offers a range of products to build new data. Mining processes and predictive setup analysis. The tool can be used for over a vast range of applications including for business applications, commercial applications, training, education, research, application development, machine learning.

C. Orange:

It is open-source software written in python language. Orange is the best software for analyzing data and machine learning. These components are called widgets. These widgets are used for reading data, analyzing components, allowing users to select the features, and showing the data. With orange, data formatting and moving them with the help of widgets becomes fast and easy.

D. Weka:

It is open-source software used for predictive modeling and analysis of data. Weka has a GUI interface that provides easy and interactive access to users. It supports SQL and allows a user to connect to the database, and performs operations by firing query. It stores data in a flat-file format.

E. KNIME:

It is open source data mining tools. It developed by KNIME.com AG. It operates on the concept of the modular data pipeline. KNIME constitutes of various machine learning and data mining components embedded together. KNIME has been used widely for pharmaceutical research.

2. 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 the list of areas where data mining is widely used

A. Financial Data Analysis
B. Retail Industry
C. Other Scientific Applications

A. Financial Data Analysis:- The financial data in banking and financial industry is generally reliable and of high quality which facilitates systematic data analysis and data mining. Some of the typical cases are as follows

– Design and construction of data warehouses for multidimensional data analysis and data mining.
– Loan payment prediction and customer credit policy analysis.
– Classification and clustering of customers for targeted marketing.
– Detection of money laundering and other financial crimes.

B. Retail Industry:- Data mining has its great application in Retail Industry because it collects large amount of data from on sales, customer purchasing history, goods transportation, consumption and services. Data mining in the retail industry helps to identify the customer buying patterns and trends that lead to improved quality of customer service and good customer support and fulfillment. Here is the list of examples of data mining in the retail industry.

– Design and Construction of data warehouses based on the benefits of data mining.
– Multidimensional analysis of sales, customers, products, time and region
– Customer Retention

C. Other Scientific Applications:- The applications discussed above tend to handle relatively small and different data sets for which the statistical techniques are appropriate. Huge amount of data have been collected from scientific domains such as geosciences, astronomy, etc., A large amount of data sets is being generated because of the fast numerical simulations in various fields such as climate and ecosystem modeling, chemical engineering, fluid dynamics, etc., Following are the applications of data mining in the field of Scientific Applications

– Data warehouses and data preprocessing.
– Graph-based mining.
– Visualization and domain specific knowledge.

3. Conclusion

Data mining has significance 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 broad application domain almost in every industry where the data is produced 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.S. Chen, J. Han, and P.S. Yu. Data Mining: An Overview from a database Prespective. IEEE Transactions on Knowledge and Data Engineering, Vol.8, pp.866-883,1996.
[2] Jiawei Han and Micheline Kamber (2006), Data Mining Concepts and Techniques, published by

Morgan Kauffman, 2nd ed.

[3] Arun K Pujari, "Data Mining Techniques", Universities India Private Limited, Hyderabad, 2001.

[4] P. Usha Madhuri and S.P. Rajagopalan," An Overview of Basic Clustering Algorithms", International Journal of computer Science and System Analysis, vol. 4, no. 1, January-June 2010, pp. 15-23.

[5] R. Kaur, S. Kaur, A. Kaur, R.Kaur, A. Kaur, "An Overview of Database management System, Data warehousing and Data Mining". IJARCCE, Vol.2, issue.7, July 2013.

[6] Jiawei Han, Member and Yongjian Fu, Member, "Mining Multiple-Level Association Rules in Large Databases", ieee transactions on knowledge and data engineering, vol 11, no.5, September/October, 2000.

[7] Nen-Fu Huang, Chia-Nan Ka, Hsien-Wei Hun, GinYuan Jai and Chia-Lin Lin," Apply Data Mining to Defense-in-Depth Network Security System". Proceedings of the 19th International Conference on Advanced Information Networking and Applications (AINA'05), 2005.

[8] Y. Fu, Data Minig: Tasks, Techniques and Applications.