IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 4 Issue 7, July 2017
ISSN (Online) 2348 – 7968 | Impact Factor (2016) – 5.264
www.ijiset.com

Data Mining Concepts And Techniques - A Survey

Dr.Sankar.K[1], Dr.Bhuvaneswari.C[2]

[1] Assistant Professor, Dept. of computer science, Thiruvalluvar university constituent arts and science college, siruvangur, Kallakurichi,villupuram district.

[2]Assistant Professor, Dept. of computer science, Thiruvalluvar university college of arts and science college, Thiruvennainallur, villupuram district.

Abstract

Data mining is the process of discovering interesting patterns from large amounts of data. It typically involves data cleaning, data integration, data selection, data transformation, pattern discovery, pattern evaluation, and knowledge presentation. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. In this paper the data mining process and its various techniques are discussed. This paper gives a complete overview of the concepts and technology.

Keywords: *Association, clustering, classification, Knowledge mining..*

1. Introduction

Data Mining is defined as the procedure of extracting information from huge sets of data. In other words, data mining is mining knowledge from data. In addition to that a similar meaning to data mining such as knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology, and data dredging. Data mining uses sophisticated mathematical algorithms to segment the data and evaluate the probability of future events. Data mining is also known as Knowledge Discovery in Data (KDD). Most companies collect and refine massive quantities of data and the data mining techniques can be implemented and the tools can also analyze the massive databases to enhance the value of existing information.

The key properties of data mining are:

- Automatic discovery of patterns
- Prediction of likely outcomes
- Creation of actionable information
- Focus on large data sets and databases.

2. The Scope Of Data Mining

Data mining technology can generate new business opportunities by providing the following capabilities:

- Automated prediction of trends and behaviors. Data mining automates the process of finding predictive information in large databases. Questions that traditionally required extensive hands-on analysis can now be answered directly from the data — quickly. A typical example of a predictive problem is targeted marketing. Data mining uses data on past promotional mailings to identify the targets most likely to maximize return on investment in future mailings. Other predictive problems include forecasting bankruptcy and other forms of default, and identifying segments of a population likely to respond similarly to given events.
- Automated discovery of previously unknown patterns. Data mining tools sweep through databases and identify previously hidden patterns in one step. An example of pattern discovery is the analysis of retail sales data to identify seemingly unrelated products that are often purchased together. Other pattern discovery problems include detecting fraudulent credit card transactions and identifying anomalous data that could represent data entry keying errors.

2.1 Processing Steps In Kdd

The following are the essential steps in the process of knowledge discovery.

1. Data cleaning (to remove noise and inconsistent data)

2. Data integration (where multiple data sources may be combined) A popular trend in the information industry is to perform data cleaning and data integration as a preprocessing step, where the resulting data are stored in a data warehouse.

3. Data selection (where data relevant to the analysis task are retrieved from the database)

4. Data transformation (where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations) Sometimes data transformation and consolidation are performed before the data selection process, particularly in the case of data warehousing. Data reduction may also be performed to obtain a smaller representation of the original data without sacrificing its integrity.

5. Data mining (an essential process where intelligent methods are applied to extract data patterns)

6. Pattern evaluation (to identify the truly interesting patterns representing knowledge based on interestingness measures

7. Knowledge presentation (where visualization and knowledge representation techniques are used to present mined knowledge to users)

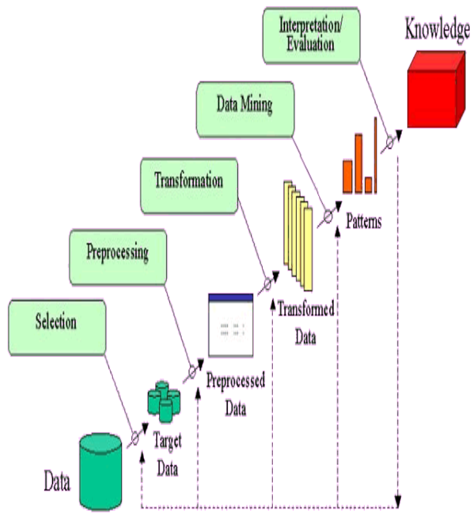

Fig 1. Datamining process

2.2 **Kinds of Patterns In Data Mining**

There are a number of data mining functionalities which includes characterization and discrimination, the mining of frequent patterns, associations, and correlations, classification and regression, clustering analysis; and outlier analysis. Data mining functionalities are used to specify the kinds of patterns to be found in data mining tasks. In general, such tasks can be classified into two categories: descriptive and predictive. Descriptive mining tasks characterize properties of the data in a target data set. Predictive mining tasks perform induction on the current data in order to make predictions.

Data Mining Techniques

There are several data mining techniques have been developing which includes association, classification, clustering, prediction, sequential patterns and decision tree.

(i) Association

Association is one of the best-known data mining techniques. In association, a pattern is discovered based on a relationship between items in the same transaction is also known as relation technique. The association technique is used in market basket analysis to identify a set of products that customers frequently purchase together.

Retailers are using association technique to research customer's buying habits. Based on historical sale data, retailers might find out that customers always buy crisps when they buy beers, and, therefore, they can put beers and crisps next to each other to save time for customer and increase sales.

Types of association rule

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

(ii) Classification

Classification is a classic data mining technique based on machine learning. Basically, classification is used to classify each item in a set of data into one of a predefined set of classes or groups. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics. The classification, helps to classify the data items into groups. For example, we can apply classification in the application that "given all records of employees who left the company, predict who will probably leave the company in a future period." In this case, we divide the records of employees into two groups that named "leave" and "stay". And then we can ask our data mining software to classify the employees into separate groups.

Types of classification models:

- Classification by decision tree induction
- Bayesian Classification
- Neural Networks
- Support Vector Machines (SVM)
- Classification Based on Associations

(iii) Clustering

Clustering is a data mining technique that makes a meaningful or useful cluster of objects which have similar characteristics using the automatic technique. The clustering technique defines the classes and puts objects in each class, while in the classification techniques, objects are assigned into predefined classes. Example, In a library, there is a wide range of books on various topics available. The challenge is how to keep those books in a way that readers can take several books on a particular topic without hassle. By using the clustering technique, we can keep books that have some kinds of similarities in one cluster or one shelf and label it with a meaningful name. If readers want to grab books in that topic, they would only have to go to that shelf instead of looking for the entire library.

Types of clustering methods

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

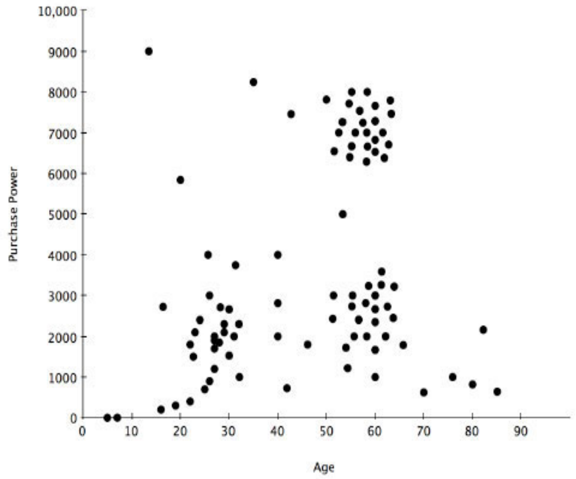

Fig 2 :Clustering

(iv) **Prediction**

The prediction, as its name implied, is one of a data mining techniques that discovers the relationship between independent variables and relationship between dependent and independent variables. For instance, the prediction analysis technique can be used in the sale to predict profit for the future if we consider the sale is an independent variable, profit could be a dependent variable. Then based on the historical sale and profit data, a fitted regression curve that is used for profit prediction is drawn.

Types of regression methods

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

(v) **Sequential Patterns**

Sequential patterns analysis is one of data mining technique that seeks to discover or identify similar patterns, regular events or trends in transaction data over a business period.

In sales, with historical transaction data, businesses can identify a set of items that customers buy together different times in a year. Then businesses can use this information to recommend customers buy it with better deals based on their purchasing frequency in the past.

(vi) **Decision Trees**

The decision tree is one of the most common used data mining techniques because its model is easy to understand for users. In decision tree technique, the root of the decision tree is a simple question or condition that has multiple answers. Each answer then leads to a set of questions or conditions that help us determine the data so that we can make the final decision based on it.

Applications Of Data Mining

(i) data mining in sales/marketing

Data mining enables businesses to understand the hidden patterns inside historical purchasing transaction data, thus helping in planning and launching new marketing campaigns in prompt and cost effective way. The following illustrates several data mining applications in sale and marketing.

- Data mining is used for market basket analysis to provide information on what product combinations were purchased together when they were bought and in what sequence. This information helps businesses promote their most profitable products and maximize the profit. In addition, it encourages customers to purchase related products that they may have been missed or overlooked.
- Retail companies use data mining to identify customer's behavior buying patterns.

(ii) Banking / Finance

- Several data mining techniques e.g., distributed data mining have been researched, modeled and developed to help credit card fraud detection.
- Data mining is used to identify customers loyalty by analyzing the data of customer's purchasing activities such as the data of frequency of purchase in a period of time, a total monetary value of all purchases and when was the last purchase. After analyzing those dimensions, the relative measure is generated for each customer. The higher of the score, the more relative loyal the customer is.
- To help the bank to retain credit card customers, data mining is applied. By analyzing the past data, data mining can help banks predict customers that likely to change their credit card affiliation so they can plan and launch different special offers to retain those customers.
- Credit card spending by customer groups can be identified by using data mining.
- The hidden correlation's between different financial indicators can be discovered by using data mining.
- From historical market data, data mining enables to identify stock trading rules.

(iii) Health Care and Insurance

The growth of the insurance industry entirely depends on the ability to convert data into the knowledge, information or intelligence about customers, competitors, and its markets. Data mining is applied in insurance industry lately but brought tremendous competitive advantages to the companies who have implemented it successfully. The data mining applications in insurance industry are listed below:

- Data mining is applied in claims analysis such as identifying which medical procedures are claimed together.

- Data mining enables to forecasts which customers will potentially purchase new policies.
- Data mining allows insurance companies to detect risky customers' behavior patterns.
- Data mining helps detect fraudulent behavior.

(iv) TRANSPORTATION

- Data mining helps determine the distribution schedules among warehouses and outlets and analyze loading patterns.

(v) MEDICINE

- Data mining enables to characterize patient activities to see incoming office visits.
- Data mining helps identify the patterns of successful medical therapies for different illnesses.

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

[2] S. M. Weiss and C. A. Kulikowski. Computer Systems that Learn: Classication and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems. Morgan Kaufman, 1991.

[3] J. D. Ullman. Principles of Database and Know ledge-Base Systems, Vol. 1. Computer Science Press, 1988.

[4] P. Adriaans and D. Zantinge. Data Mining. Addison-Wesley: Harlow, England, 1996.

[5] S. M. Weiss and N. Indurkhya. Predictive Data Mining. Morgan Kaufmann, 1998..

First Author [1]Dr.K.shankar is working as a assistant professor and Head in the Department of computer science, Thiruvalluvar university constituent arts and science college, siruvangur, Kallakurichi. He received UG and PG degrees from Nehru memorial college puthanampatti, Bharathidasan university India. M.Phil from Vinayaka mission university, Salem and Ph.D from karpagam university. He has 18 years of teaching experience and has published 7 international papers. His research interests are Computer Networks ,image processing and Programming languages.

Second Author [2]Dr.C.Bhuvaneswari is working as a assistant professor and Head in the Department of computer science, Thiruvalluvar university college of arts and science, Thiruvennainallur, Villupuram. She received UG from Madras university and PG from Bharathidasan university,M.Phil from Madurai kamaraj university, MCA & Ph.D from Annamalai university, Chidambaram. She has 13 years of teaching experience and 8 years of research experience. She has published 14 international papers and attended 7 international conferences and 14 national publications, 1 book chapter to her credit. Her research interests are data mining and image processing.