An Overview of Data Mining Techniques and Applications

R. Tamilselvi[1], S. Kalaiselvi[2]

[1]Assistant Professor, Department of Computer Science
Dr.SNS Rajalakshmi College of Arts & Science
Coimbatore, India
rtamilu@gmail.com

[2]Assistant Professor, Department of Computer Science
Dr.SNS Rajalakshmi College of Arts & Science
Coimbatore, India
kulandaikalai@gmail.com

Abstract: ***Data mining may be defined as the science of extracting useful information from databases. It also called knowledge discovery. Using a combination of machine learning, statistical analysis, modeling techniques and database technology, data mining finds patterns and subtle relationships in data and infers rules that allow the prediction of future. In this paper overview of data mining, Types and Components of data mining algorithms have been discussed. Data mining tasks like Decision Trees, Association Rules, Clustering, Time-series and its related data mining algorithms have been included. The working style and the data required for the algorithms are explained. Each algorithm has its own set of merits and demerits. We have also incorporated the various application domains of Decision Trees and Clustering algorithms.***

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, typically deals with data that have already been collected for some purpose rather than the data mining analysis. This means that the objectives of data mining exercise play no role in the data collection strategy. The data sets examined in data mining are often large.

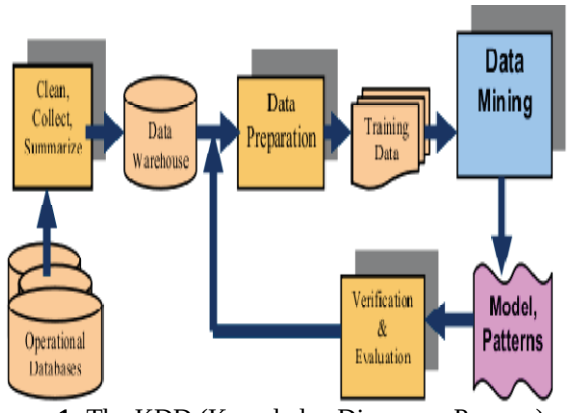

Figure 1: The KDD (Knowledge Discovery Process) and data mining process (Han & Kamber, 2002)

2. Data Mining Terminologies

Data: Data are any facts, numbers, or text that can be processed by a computer.

Information: The patterns, associations, or relationships among all this data can provide information.

Knowledge: Information can be converted into knowledge about historical patterns and future trends.

Data Warehouses: Data Warehouse is a repository of information collected from multiple sources, stored under a unified schema, and that usually resides at a single site.

Association Analysis: Association analysis is the discovery of association rules showing attribute-value conditions that occur frequently together in a given set of data.

Data Mining: It is the extraction of hidden predictive information from large databases. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions.

3. Data Mining Types

Predictive data mining: It produces the model of the system described by the given data. It uses some variables or fields in the data set to predict unknown or future values of other variables of interest.

Descriptive Data Mining: It produces new, non trivial information based on the available data set. It focuses on finding patterns describing the data that can be interpreted by humans.

3. Data Mining Tasks

- Data processing [descriptive]
- Prediction [predictive]
- Regression [predictive]
- Clustering [descriptive]
- Classification [predictive]

- Link analysis/ associations [descriptive]
- Evolution and deviation analysis [predictive]

4. Basic Facts in KNN

Data mining has attracted a great attention in the information industry and in society as a whole in recent years, due to wide availability of huge amount of data and the imminent need for turning such data into useful information and knowledge. The information and knowledge gained can be used for application ranging from market analysis, fraud detection, to production control, disaster management and science exploration. Data mining can be viewed as a result of the natural evolution of information technology. The database system industry has witnessed an evolutionary path in the development of various functionalities: data collection and database creation, database management (including data storage and retrieval, and database transaction processing and advance data analysis Knowledge discovery as a process consists of an iterative sequence of following steps:

1. **Data cleaning**, that is, to remove noise and inconsistent data.
2. **Data integration**, that is, where multiple data sources are combined.
3. **Data selection**, that is, where data relevant to the analysis task are retrieved from the database.
4. **Data transformation**, that is, where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.
5. **Data mining**, that is, an essential process where intelligent methods are applied in order to extract the data patterns.
6. **Knowledge presentation**, that is, where visualization and knowledge representation techniques are used to present the mined knowledge to the user.

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.

5. 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.

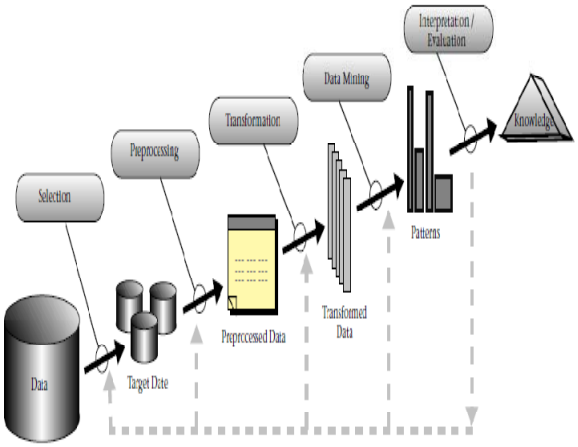

Figure 2: Data mining as a step in the process of Knowledge Discovery

6.1 Classification

Discovery of a predictive learning function that classifies a data item into one of several predefined classes. 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
- Back propagation
- Support Vector Machines (SVM)
- Classification Based on Associations
- K-Nearest Neighbor Classifies
- Case Based Reasoning
- Genetic Algorithm
- Rough Set Approach
- Fuzzy Set Approach

6.2 Clustering

A common descriptive task in which one seeks to identify a finite set of categories or cluster to describe the data. 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

- Hierarchical (divisive) Methods
- Partitioning Methods
- Density Based Methods
- Grid-Based Methods
- Model Based Algorithms

Categorization of Clustering Algorithms

Algorithms are key step for solving the techniques. In these clustering techniques, various algorithms are currently in the life, still lot more are evolving. But in general, the algorithm for clustering is neither straight nor canonical:

Hierarchical methods:
- Agglomerative Algorithms
- Divisive Algorithms

Partitioning methods:
- Relocation Algorithms
- Probabilistic Clustering
- K-Medoids Methods
- K-Means Methods

Density-based algorithms:
- Density-based connectivity clustering
- Density functions clustering

Grid-based methods:
- Methods based on co-occurrence of categorical data
- Constraint-based clustering
- Clustering algorithms used in machine learning
- Gradient descent and artificial neural networks
- Evolutionary methods

Model Based Algorithms:
- Algorithms for high dimensional data
- Subspace Clustering
- Projection Techniques
- Co-Clustering Techniques

6.3 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.

Types of Regression Methods

- Multivariate Linear Regression
- Nonlinear Regression
- Multivariate Nonlinear Regression

6.4 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

6.5. 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.

6. Data Mining Applications

Data Mining For Financial Data Analysis

In Banking Industry data mining is used:

1. Predicting Credit fraud
2. Evaluation Risk
3. Performing trend analysis
4. Analyzing profitability
5. Helping with direct marketing campaigns

In financial markets and neural networks data mining is used:

1. Forecasting stock prices
2. Forecasting commodity-price prediction
3. Forecasting financial disasters

Data Mining for Telecommunications Industry used:

1. How does one retain customers and keep them loyal as competitors offer special offers and reduced rates?
2. When is a high-risk investment, such as new fiber optic lines, acceptable?
3. How does one predict whether customers will buy additional products like cellular services, call waiting, or basic services?
4. What characteristics differentiate our products from those of our competitors?

Data Mining for the Retail Industry:

The retail industry is a major application area for data mining since it collects huge amounts of data on sales, customer-shopping history, goods transportation, consumption patterns, and service records.

1. What are the best types of advertisements to reach certain segments of customers?
2. What is the optimal timing at which to send mailers?
3. What types of products can be sold together?
4. How does one retain profitable customers?
5. What are the significant customer segments that buy products?

7. Conclusion

Data mining is a broad area that integrates techniques from several fields including machine learning, statistics, pattern recognition, artificial intelligence, and database systems, for the analysis of large volumes of data. There have been a large number of data mining algorithms rooted in these fields to perform different data analysis tasks. Data mining techniques and algorithms such as classification, clustering etc., helps in finding the patterns to decide upon the future trends in businesses. In this study, the basic concept of clustering and clustering techniques are given. 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.
[2] Paulraj Ponnian, .Data Warehousing Fundamentals. John Wiley.
[3] M.H. Dunham, .Data Mining Introductory and Advanced Topics., Pearson Education.
[4] 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.
[5] Arun K Pujari, "Data Mining Techniques", Universities India Private Limited, Hyderabad, 2001.
[6] 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.

Author Profile

R. Tamilselvi received the B.Sc (CS) and M.Sc from Bharathiar University 2001 and 2003, respectively. She received her M.Phil in year of 2006. She presented more than 20 papers in various National Conferences. She published 3 Articles and Area of Interest is data Mining, RDBMS. At Present She Working As a Assistant Professor in Department of Computer Science at Dr.SNS Rajalakshmi College of Arts & Science, Coimbatore, Tamilnadu, India

S. Kalaiselvi received the B.Sc (CS) and M.Sc from Bharathiar University 2001 and 2003, respectively. She received her M.phil in year of 2008 and Area of Interest is networking. At Present she is working as Assistant Professor in department of Computer Science at Dr.SNS Rajalakshmi College of Arts & Science, Coimbatore, Tamilnadu, India