SEQUENCE DATA MINING TECHNIQUES AND APPLICATIONS

Ritika Goyal[1]
Assistant Professor[1]
Bhai Behlo Khalsa Girls College Phaphre Bhai Ke Punjab[1], Mansa[1],

ABSTRACT

Data mining is the analytics and knowledge discovery process of analyzing large volumes of data from various sources and transforming the data into useful information. Various disciplines have contributed to its development and is becoming increasingly important in the scientific and industrial world. Recent research trends focus more on large data sets and big data. Recently there have also been more applications in area of health informatics with the advent of newer algorithms.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.

Keywords:Data mining Techniques; Data mining ; Data mining applications

INTRODUCTION

Data Mining is defined as the procedure of extracting information from huge sets of data. In other words, we can say that data mining is mining knowledge from data. The tutorial starts off with a basic overview and the terminologies involved in data mining and then gradually moves on to cover topics such as knowledge discovery, query language, classification and prediction, decision tree induction, cluster analysis, and how to mine the Web.There is a huge amount of data available in the Information Industry. This data is of no use until it is converted into useful information. It is necessary to analyze this huge amount of data and extract useful information from it.Extraction of information is not the only process we need to perform; data mining also involves other processes such as Data Cleaning, Data Integration, Data Transformation, Data Mining, Pattern Evaluation and Data Presentation. Once all these processes are over, we would be able to use this information in many applications such as Fraud Detection, Market Analysis, Production Control, Science Exploration, etc.Data Mining is defined as extracting information from huge sets of data. In other words, we can say that data mining is the procedure of mining knowledge from data.

MINING METHODOLOGY AND USER INTERACTION ISSUES

It refers to the following kinds of issues −

- **Mining different kinds of knowledge in databases** − Different users may be interested in different kinds of knowledge. Therefore it is necessary for data mining to cover a broad range of knowledge discovery task.
- **Interactive mining of knowledge at multiple levels of abstraction** − The data mining process needs to be interactive because it allows users to focus the search for patterns, providing and refining data mining requests based on the returned results.

- **Incorporation of background knowledge** − To guide discovery process and to express the discovered patterns, the background knowledge can be used. Background knowledge may be used to express the discovered patterns not only in concise terms but at multiple levels of abstraction.
- **Data mining query languages and ad hoc data mining** − Data Mining Query language that allows the user to describe ad hoc mining tasks, should be integrated with a data warehouse query language and optimized for efficient and flexible data mining.
- **Presentation and visualization of data mining results** − Once the patterns are discovered it needs to be expressed in high level languages, and visual representations. These representations should be easily understandable.
- **Handling noisy or incomplete data** − The data cleaning methods are required to handle the noise and incomplete objects while mining the data regularities. If the data cleaning methods are not there then the accuracy of the discovered patterns will be poor.
- **Pattern evaluation** − The patterns discovered should be interesting because either they represent common knowledge or lack novelty.

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.

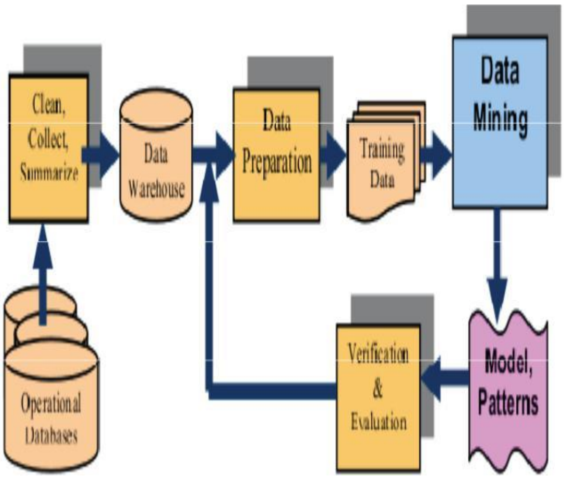
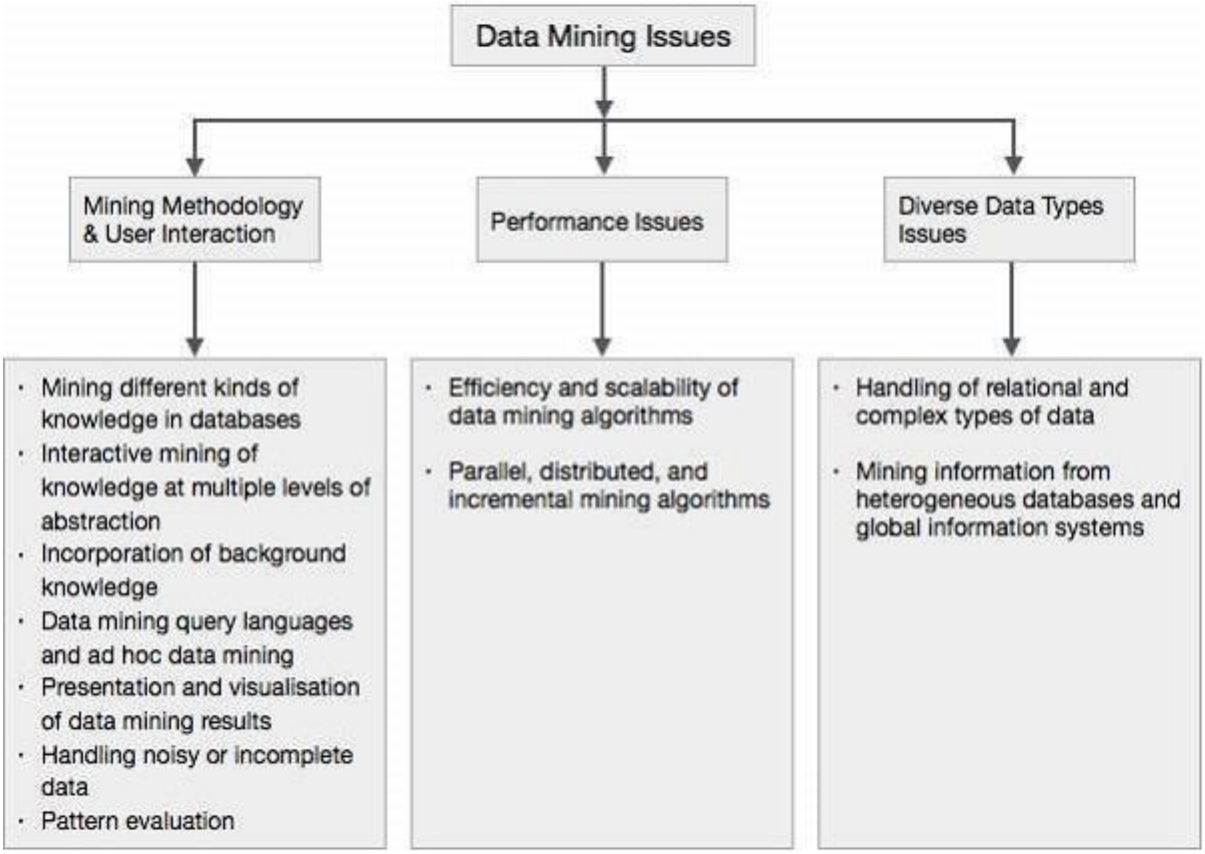


Figure 1: The KDD (Knowledge Discovery Process) and data mining process (Han & Kamber, 2002)

THE FOLLOWING DIAGRAM DESCRIBES THE MAJOR ISSUES


Data Mining Issues
Mining Methodology & User Interaction
Performance Issues
Diverse Data Types Issues
Mining different kinds of knowledge in databases
Interactive mining of knowledge at multiple levels of abstraction
Incorporation of background knowledge
Data mining query languages and ad hoc data mining
Presentation and visualisation of data mining results
Handling noisy or incomplete data
Pattern evaluation
Efficiency and scalability of data mining algorithms
Parallel, distributed, and incremental mining algorithms
Handling of relational and complex types of data
Mining information from heterogeneous databases and global information systems

PERFORMANCE ISSUES

There can be performance-related issues such as follows −

- **Efficiency and scalability of data mining algorithms** − In order to effectively extract the information from huge amount of data in databases, data mining algorithm must be efficient and scalable.
- **Parallel, distributed, and incremental mining algorithms** − The factors such as huge size of databases, wide distribution of data, and complexity of data mining methods motivate the development of parallel and distributed data mining algorithms. These algorithms divide the data into partitions which is further processed in a parallel fashion. Then the results from the partitions is merged. The incremental algorithms, update databases without mining the data again from scratch.

DIVERSE DATA TYPES ISSUES

- **Handling of relational and complex types of data** − The database may contain complex data objects, multimedia data objects, spatial data, temporal data etc. It is not possible for one system to mine all these kind of data.
- **Mining information from heterogeneous databases and global information systems** − The data is available at different data sources on LAN or WAN. These data source may be structured, semi structured or unstructured. Therefore mining the knowledge from them adds challenges to data mining.

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.

MARKET ANALYSIS AND MANAGEMENT

Listed below are the various fields of market where data mining is used −

1. **Customer Profiling** − Data mining helps determine what kind of people buy what kind of products.
2. **Identifying Customer Requirements** − Data mining helps in identifying the best products for different customers. It uses prediction to find the factors that may attract new customers.
3. **Cross Market Analysis** − Data mining performs Association/correlations between product sales.
4. **Target Marketing** − Data mining helps to find clusters of model customers who share the same characteristics such as interests, spending habits, income, etc.
5. **Determining Customer purchasing pattern** − Data mining helps in determining customer purchasing pattern.
6. **Providing Summary Information** − Data mining provides us various multidimensional summary reports.

CORPORATE ANALYSIS AND RISK MANAGEMENT

Data mining is used in the following fields of the Corporate Sector

1. **Finance Planning and Asset Evaluation** − It involves cash flow analysis and prediction, contingent claim analysis to evaluate assets.
2. **Resource Planning** − It involves summarizing and comparing the resources and spending.
3. **Competition** − It involves monitoring competitors and market directions.

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

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.

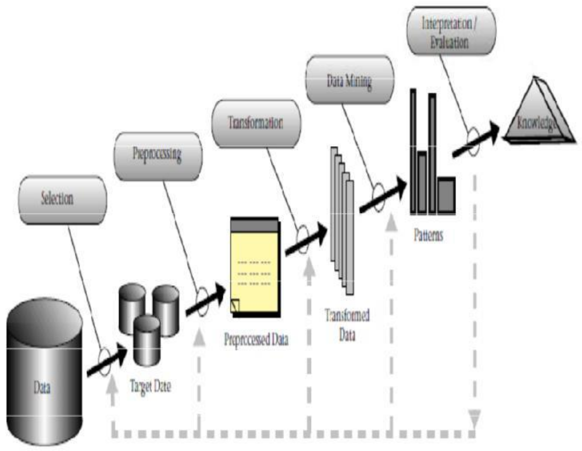


Figure 2: Data mining as a step in the process of Knowledge Discovery

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.

PREDICATION

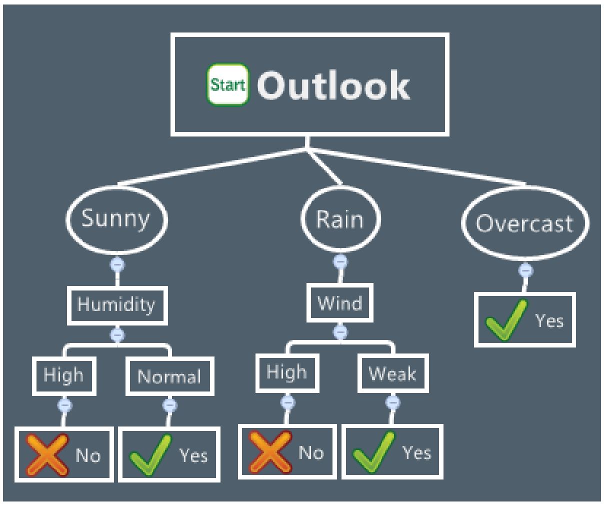
Regression technique can be adapted for predication. Regression analysis can be used to model the relationship between one or more independent variables and dependent variables. In data mining independent variables are attributes already known and response variables are what we want to predict. Unfortunately, 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.

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.

Decision trees

A decision tree is one of the most commonly 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. For example, We use the following decision tree to determine whether or not to play tennis:



Starting at the root node, if the outlook is overcast then we should definitely play tennis. If it is rainy, we should only play tennis if the wind is the week. And if it is sunny then we should play tennis in case the humidity is normal.

CONCLUSION

The use of data mining in enrollment management is a fairly new development. Current data mining is done primarily on simple numeric and categorical data. In the future, data mining will include more complex data types. In addition, for any model that has been designed, further refinement is possible by examining other variables and their relationships. Research in data mining will result in new methods to determine the most interesting characteristics in the data. As models are developed and implemented, they can be used as a tool in enrollment management. 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. U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “From Data Mining to Knowledge Discovery in Database,” AI Magazine, Vol.17, pp. 37-54, 1996.
2. M. Kantardzic, “Data Mining: Concepts, Models, Methods and Algorithms,” John Wiley & Sons, Inc., 2002.
3. Jiawei Han and Micheline Kamber (2006), Data Mining Concepts and Techniques, published by Morgan Kauffman.
4. Paulraj Ponnian, .Data Warehousing Fundamentals. John Wiley.
5. M. H. Dunham, “Data mining introductory and advanced topics,” Pearson Education, Inc., 2003.
6. G. C. Nsofor, “A Comparative Analysis of Predictive Data-Mining Techniques,” Master of Science, The University of Tennessee, Knoxville, 2006.
7. F. Gorunescu, “Data Mining Concepts, Models and Techniques,” Vol.12, Springer-Verlag Berlin Heidelberg, 2011.
8. 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.
9. Arun K Pujari, “Data Mining Techniques”, Universities India Private Limited, Hyderabad, 2001.
10. 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.
11. https://www.tutorialspoint.com/data_mining/dm_overview.htm