

Sequence Data Mining Processin Implementation Of Health Care Domain By Means of Association Rule.

G Shyama Chandra Prasad

Associate Prof, Matrusri Engineering College, Hyderabad,

ABSTRACT

Data mining is the analytic 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..Data mining plays vitalrole in various domains including health care, banking, retail, education, telecommunication, information security, finance etc. Association rule mining is one of the data mining techniques which plays vital role for analysis and decision making and it is used to find out the various patterns or association in the item set.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. The paper describes about the Association rule mining and an Apriori Algorithm. Also the paper discuss about the reviews of research work done in this filed by diverse researchers, scholars, organizations etc. 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,Text Mining, Association Rule, Apriori Algorithm, Diseases.*

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. An association rule mining is nothing but finding the frequent patterns, correlations, associations, or casual structures among the sets of items or objects in transaction database, relational database, and other information repositories. In data mining, Association Rule learning is a popular and well researched method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using different measures of interesting. 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.

ASSOCIATION RULE MINING

Association rule is having two main important properties namely Support and Confidence as $\text{Support}(AB) = P(A \cup B)$

$$\text{Confidence}(AB) = P(B|A)$$

If we correlate support and confidence then

$$\text{Confidence}(AB) = P(B|A) = \frac{\text{Support_count}(A \cup B)}{\text{Support_count}(A)}$$

(A) $\text{Support_count}(A \cup B)$: number of transaction containing the item sets A U B, $\text{Support_count}(A)$: number of transactions containing the item set A.

Association rule two steps, one is **Find all frequent item sets**: By definition, each of these item sets will occur at least as frequently as a predetermined minimum support count; and another is Generate strong association rules from the **frequent item sets**: By definition, these rules must satisfy minimum support and minimum support and minimum confidence.

Apriori Algorithm: (Finding Frequent Item sets Using Candidate Generation)

Apriori is an influential algorithm for mining frequent item sets. The name of the algorithms is based on the fact that the algorithm uses prior knowledge of frequent item sets properties. Apriori employs an iterative approach known as a level-wise search.

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.

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.

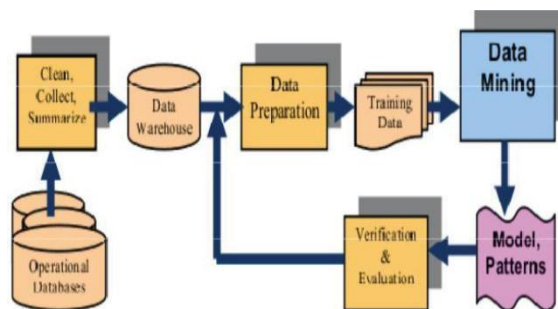
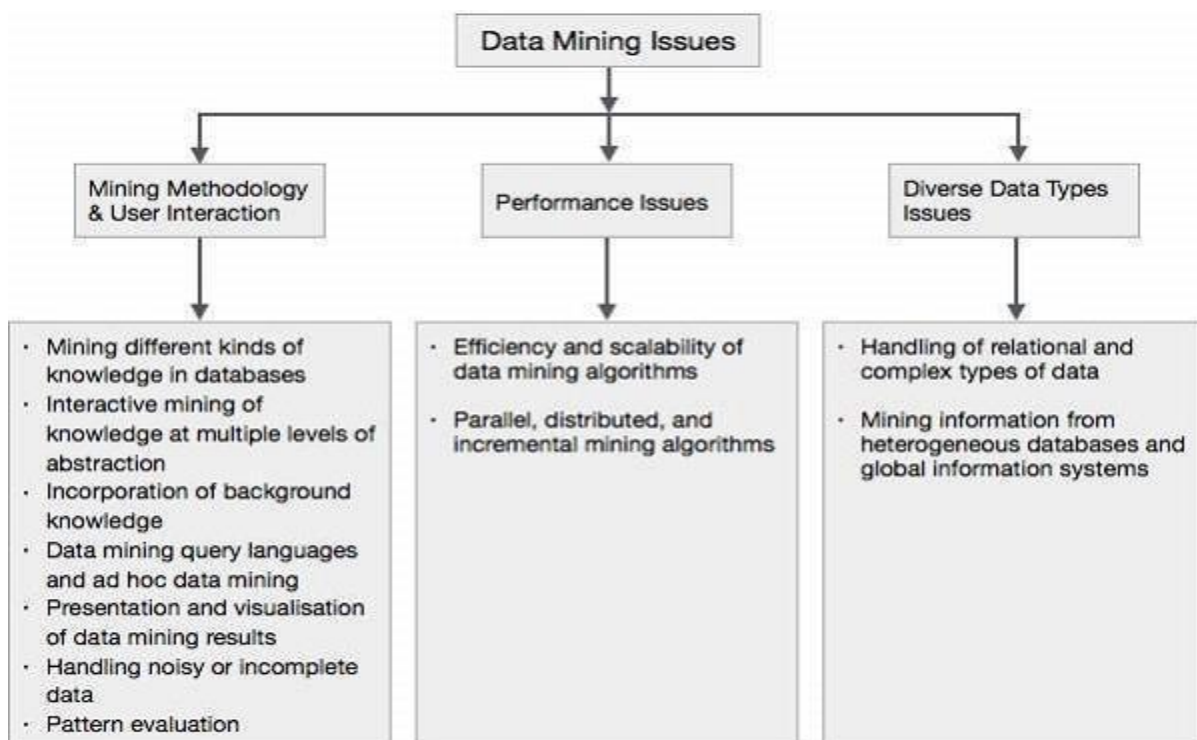


Figure 1: The KDD (Knowledge Discovery Process) and data mining process.

THE FOLLOWING DIAGRAM DESCRIBES THE MAJOR ISSUES



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.

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

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.

ASSOCIATION RULE GENERATION FOR VIRAL INFECTIVE DISEASES

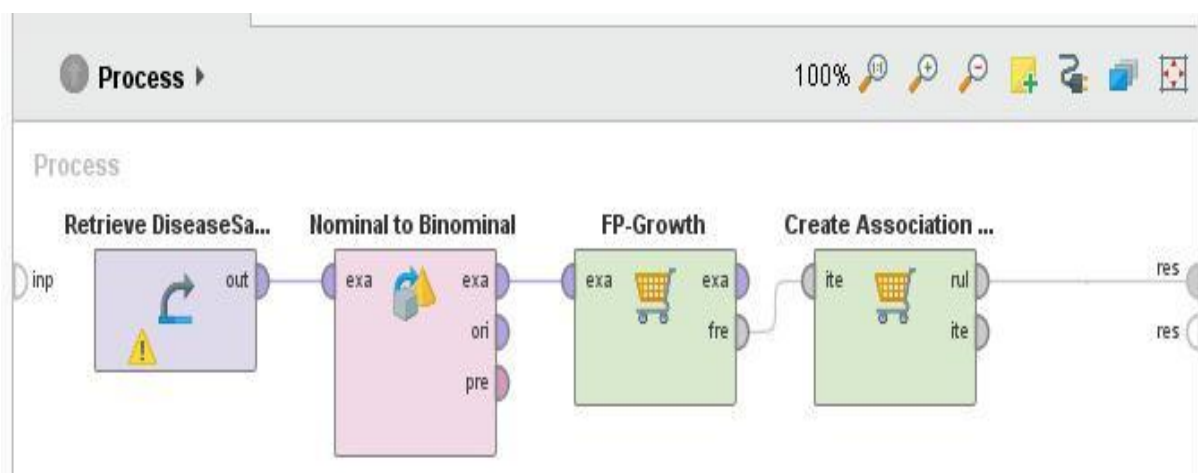


Diagram : Dataset for Association Mining in Rapid Miner.

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.

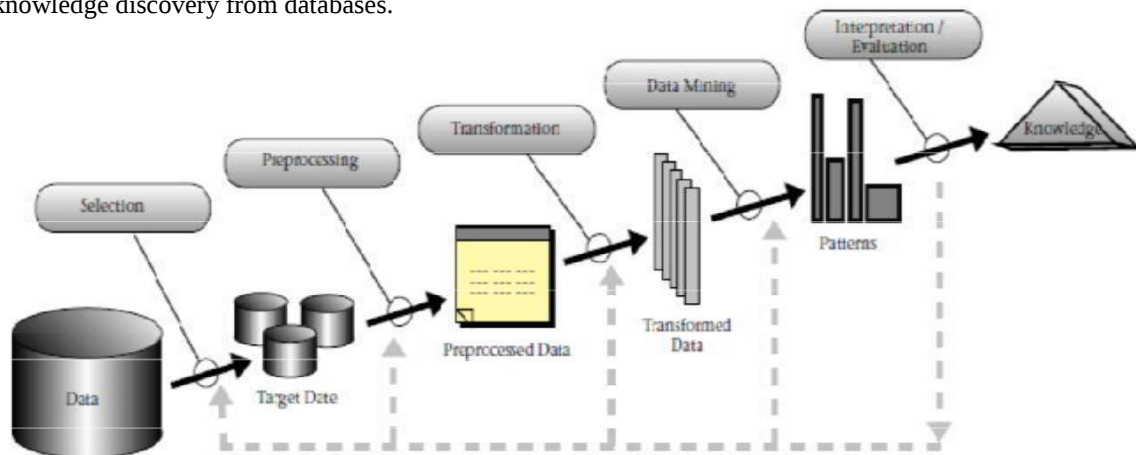


Figure : Data mining as a step in the process of Knowledge Discovery

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.

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 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 weak. 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. The data mining works with the structured data that is the data which is in row and column format. The future scope of the present study is to work on text mining of health care data and generating the various patterns from it. This research used an Apriori algorithm an association rule generation algorithm for various viral infective diseases and their symptoms. 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. 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.
11. https://www.tutorialspoint.com/data_mining/dm_overview.htm