International Journal of Research in Engineering, Science and Management
Volume-2, Issue-1, January-2019
www.ijresm.com | ISSN (Online): 2581-5792

Data Mining and Knowledge Discovery

C. P. Balasubramaniam[1], C. Janaki[2], P. K. Mangaiyarkarasi[3], S. Sindhuja[4]
[1,2,4]*Assistant Professor, Department of Computer Science, Kongu Arts and Science College, Erode, India*
[3]*Associate Professor, Department of Computer Science, Kongu Arts and Science College, Erode, India*

***Abstract*: Data mining is used to find or generate new useful information's from large amount of data base. It is a process of extracting previously unknown and processable information from large databases and using it to make important business decisions. Several emerging applications in information providing services, such as data warehousing and on-line services over the Internet, also call for various data mining and knowledge discovery techniques to understand user behavior better, to improve the service provided, and to increase the business opportunities Of an overview of knowledge discovery database and data mining**

***Keywords*: KDD–Knowledge Discovery in Data base, Data mining process.**

1. Introduction

In real-time information technology has generated and used large amount of databases and stored 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 [1]. Data mining is a process of extracting previously unknown and process able information from large databases and using it to make important business decisions. It is also called as knowledge discovery process, Data mining should be used exclusively for the discovery stage of the KDD process.

A. Knowledge discovery database

Some people don't differentiate data mining from knowledge discovery while others view data mining as an essential step in the process of knowledge discovery. Here is the list of steps involved in the knowledge discovery process,

- Data Selection: Here, data relevant to the analysis task are retrieved from the database.
- Data Transformation: In this step, data is transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.
- Data Mining: In this step, intelligent methods are applied in order to extract data patterns.
- Pattern Evaluation: In this step, data patterns are evaluated.
- Knowledge Presentation: In this step, knowledge is represented.

2. The KDD process

The knowledge discovery process is iterative and interactive, consisting of nine steps [3]. Note that the process is iterative at each step, meaning that moving back to previous steps may be required .So it is required to understand the process and the different needs and possibilities in each step. A typical knowledge discovery process is shown in Fig. 1, and the process is elaborated in each step.

- Developing an understanding of the application domain.
- Selecting and creating a data set on which discovery will be performed.
- Preprocessing and cleansing.
- Choosing the appropriate Data Mining task.
- Choosing the Data Mining algorithm.
- Employing the Data Mining algorithm.
- Evaluation.
- Using the discovered knowledge.

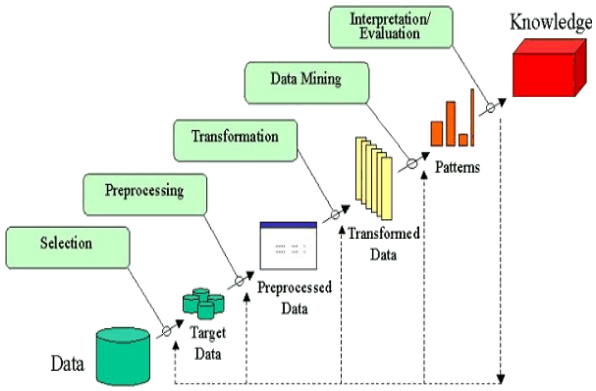

Fig. 1. The KDD process

The terms knowledge discovery and data mining are distinct. KDD refers to the overall process of discovering useful knowledge from data. It involves the evaluation and possibly interpretation of the patterns to make the decision of what qualifies as knowledge. It also includes the choice of encoding schemes, preprocessing, sampling, and projections of the data prior to the data mining step. Data mining refers to the application of algorithms for extracting patterns from data without the additional steps of the KDD process.

3. Data mining

Data mining is the process of discovering actionable information from large sets of data [4]. Data mining uses mathematical analysis to derive patterns and trends that exist in

data. These patterns and trends can be collected and defined as a data mining model. Mining models can be applied to specific scenarios, such as

Forecasting: Estimating sales, predicting server loads or server downtime

Risk and probability: Choosing the best customers for targeted mailings, determining the probable break-even point for risk scenarios, assigning probabilities to diagnoses or other outcomes

Recommendations: Determining which products are likely to be sold together, generating recommendations

Finding sequences: Analyzing customer selections in a shopping cart, predicting next likely events

Grouping: Separating customers or events into cluster of related items, analyzing and predicting affinities.

Building a mining model is part of a larger process that includes everything from asking questions about the data and creating a model to answer those questions, to deploying the model into a working environment [6]. This process can be defined by using the following basic steps:

- Defining the Problem
- Preparing Data
- Exploring Data
- Building Models
- Exploring and Validating Models

4. Data mining techniques

There are several major data mining techniques have been developing and using in data mining projects recently including association, classification, clustering, prediction, sequential patterns and decision tree. We will briefly examine those data mining techniques in the following sections.

A. Association

Association is one of the best-known data mining technique. In association, a pattern is discovered based on a relationship between items in the same transaction. That's is the reason why association technique 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 the customer and increase sales.

B. 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. In classification, we develop the software that can learn how 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.

C. 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. To make the concept clearer, we can take book management in the library as an 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.

D. 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, we can draw a fitted regression curve that is used for profit prediction.

E. 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.

F. Decision trees

The 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:

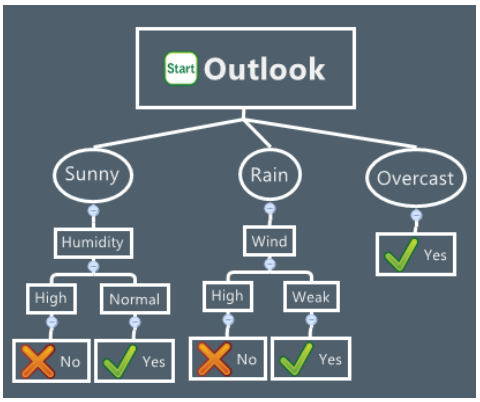

Fig. 2. Decision tree

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.

We often combine two or more of those data mining techniques together to form an appropriate process that meets the business needs.

5. Data mining algorithms

C4.5 and beyond: Systems that construct classifiers are one of the commonly used tools in data mining [8]. Such Systems take as input a collection of cases, each belonging to one of a small number of Classes and described by its values for a fixed set of attributes, and output a classifier that can accurately predict the class to which a new case belongs.

The k-means algorithm: The k-means algorithm is a simple iterative method to partition a given dataset into a user specified Number of clusters, k. this algorithm has been discovered by several researchers across different disciplines. The algorithm operates on a set of _d-dimensional vectors, D = {xi| i = 1 . . . N}, where xi _d denotes the ith data point. The algorithm is initialized by picking k points in _d as the initial k cluster representatives or “centroids”.

Step 1: Data Assignment. Each data point is assigned to its closest centroid, with ties Broken arbitrarily. This results in a partitioning of the data.

Step 2: Relocation of “means”. Each cluster representative is relocated to the center (Mean) of all data points assigned to it.

Support vector machines: The machine learning applications, Support Vector Machines (SVM) are considered A must try, it offers one of the most robust and accurate methods among all well-known Algorithms. Therefore it has a sound theoretical foundation, requires only a dozen examples for training, and is insensitive to the number of dimensions.

The Apriori algorithm: One of the most popular data mining approaches is to find frequent item sets from a transaction Dataset and derive association rules. Therefore Finding frequent item sets is not trivial because of its combinatorial explosion.

The EM algorithm: Finite mixture distributions provide a flexible and mathematical-based approach to the modeling and clustering of data observed on random phenomena. Therefore we focus here on the use of Normal mixture models, which can be used to cluster continuous data and to estimate the Underlying density function.

Page Rank: The most popular Page Ranking algorithm issued by Google search engine. The algorithm assigns ranks for each hyperlink on the web. Based on this algorithm, they built the search engine Google, which has been a huge success. Nowadays every search engine has its own hyperlink based ranking method.

Ada Boost: Ensemble learning deals with methods which employ multiple learners to solve a problem. This generalization ability of an ensemble is usually significantly better than that of a single learner, so ensemble methods are very attractive.

kNN: k-nearest neighbor classification: One of the simplest and rather trivial classifiers is the Rote classifier, which memorizes the entire training data and performs classification only if the attributes of the test object match one of the training examples exactly.

6. Conclusion

Data mining has the most important and promising features of interdisciplinary developments in Information technology. This review would help the researchers to focus on the various issues of data mining. An overview of knowledge discovery database and data mining techniques has provided an extensive study on data mining techniques. Data mining is useful for both public and private sectors for finding patterns, forecasting, discovering knowledge in different domains such as finance, marketing, banking, insurance, health care and retailing. Data mining is commonly used in these domains to increase the sales, to reduce the cost and enhance research to reduce costs & enhance research.

References

[1] Bharati M. Ramageri, “Data Mining Techniques and Applications,” Indian Journal of Computer Science and Engineering, Vol. 1 No. 4, pp. 301-305

[2] Hemlata Sahu, Shalini Shrma and Seema Gondhalakar, “A Brief Overview on Data Mining Survey,” International Journal of Computer Technologyand Electronics Engineering (IJCTEE)., Vol.1, Issue 3, pp.114-121

[3] Kalyani M Raval, “Data Mining Techniques,” International Journal of Advanced Research in Computer Science and Software Engineering,Vol. 2Issue 10,pp. 439-442

[4] Sangeeta Goele, NishaChanana, “Data Mining Trend in Past, Current and Future,” International Journal of Computing & Business Research, in Proc. I-Society 2012, 2012.

[5] Kalyani M Raval, “Data Mining Techniques,” International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 2, Issue 10, pp. 439-442.

[6] Sangeeta Goele, and Nisha Chanana, “Data Mining Trend In Past, Current And Future,” International Journal of Computing & Business Research, in Proc. I-Society 2012, 2012

[7] S. P. Deshpande and V. M. Thakare, "Data Mining System and Applications: A Review," International Journal of Distributed and Parallel systems (IJDPS) Vol.1, No.1, September 2010, pp. 32-44

[8] Y. Ramamohan, K. Vasantharao, C. Kalyana Chakravarti, and A. S. K. Ratnam, "A Study of Data Mining Tools in Knowledge Discovery Process," International Journal of Soft Computing and Engineering (IJSCE), Vol. 2, Issue-3, July 2012, pp.19 1-1994.