AJCSE
ASIAN JOURNAL OF COMPUTER SCIENCE ENGINEERING

www.ajcse.info

Asian Journal of Computer Science Engineering 2016;1(1):13-16

REVIEW ARTICLE

Data Mining

*Jesna Mathew[1], Nitisha Bansal[2], Lovleen Bhan[3]

Students of CSE (III SEM) Mandsaur University, Mandsaur

Received on: 19/07/2016, Revised on: 25/10/2016, Accepted on: 30/10/2016

ABSTRACT

"Computers have permitted us a fountain of wisdom but refused to handle a huge amount of data". To get over this data, the scholars of 1990s introduced the advanced problem solving methodology named "Data Mining" which has given us abundant of opportunities for exploring and analyzing new key type of data in new ways. In simple, Data Mining is a process of employing one or more computer technique to automatically analyze & extract knowledge from large scale data. For this paper we overviewed, the process model of data mining and its tasks.

Keywords- Data mining, pattern

INTRODUCTION

The amount of data kept in computer is increasing day by day. And the user are using it for getting more and more satisfactory information. A shopkeeper is not happy with just the knowledge of its customers, but he too want to know maximum about them. And the SQL languages are not enough to fulfill the raising demands for information.

Data mining is working as blessing to solve the problems of people. Data mining is something which help us to extract our useful data from huge amount of data.

In a broad sense, data mining is a four step process to perform a data mining session.

We:

- Assemble a collection of data to analyze.
- Present this data to a data mining software program.
- Interpret the result
- Apply the result to a new problem or situation.

At the same instant, Data mining perform many useful tasks.

PROCESSING MODEL

- Assembling the data
- The data warehouse
- Relational database flat files
- Interpreting the result
- Result application

Assembling the Data:

To assemble the data, it should be accessed through data mining. Data can be assembled by multiple records in one file or several files. A common misconception is that in order to build an effective model a data mining algorithm must be presented with thousands or millions of instances. In fact, most data mining tools work best with a few hundred or a few thousand pertinent records. Therefore once a problem has been defined, a first step of data mining process is to extract or assemble a relevant subset of data for processing. Many times this first step requires a great amount of human time and effort. There are three common ways to access data for mining.

Through Data Warehouse, Relational database and via Spreadsheets.

The Data Warehouse:

A common scenario for data assembly shows data originating in one or more operational databases. Operational databases are transaction based and frequently designed using a relational database model. An operational database fixed on the relational model will contain several normalized tables. The tables have been normalized to reduce data redundancy and promote quick access to individual records. For example, a specific customer might have data appearing in several relational tables where each table views the customer from a different perspective.

Data is transferred from operational environment to a data warehouse. A data warehouse is a historical database designed for decision support rather than transaction processing (Kimball et al., 1996). Thus only data useful for decision support is extracted from the operational environment and entered into the warehouse database. Data transfer from operational database to warehouse is

Corresponding Author: Jesna Methew, Email: *mathewjes17@gmail.com

an ongoing process usually accomplished a daily basis after the close of a regular business day. Before each data items enters into warehouse, the item is time stamped, transformed as necessary, ad checked for errors. The transfer process can be complex, especially when several operational databases are involved. Once entered, the records in the data warehouse become read-only and are subject to change only under special conditions.

A data warehouse stores all data relating to the same subject (such as customers) in the same table. This distinguishes the data warehouse from an operational database, which stores information so as to optimize transaction processing. Because the data warehouse is subject-oriented rather than transaction-oriented, the data will contain redundancies. It is redundancy stored n data warehouse that is used by data mining algorithm to develop pattern representing discovered knowledge.

Relational database flat files:

If a data warehouse does not exist, you can make use of database query language such as SQL to write one or more queries to create a table suitable for data mining. From where ever we extract the data we have to transform it into its respective data mining tool either its query language or data warehouse. Finally, if a database structure to store the data has not being designed, and the amount of collected data is minimal, the data will likely be stored in a flat file or spreadsheet.

Mining of Data:

Prior to giving a data mining tool to a data we have several choices:

- Should learning be supervised or un-supervised
- Which instances in the assembled data will be used for building the model and which instances will test the model.
- Which attributed will be selected from the list of available attributes.
- Data mining tool requires the user to specify one or more learning parameters. What parameter setting should be used to build a model to best represent data?

Interpreting the Result:

Interpreting result symbolizes us to extract the meaningful data and to find out outcomes of latest discovery of data, If the results are not satisfactory we can repeat our extraction process to get an optimized result.

Result Application:

Our ultimate goal is to apply what has been discovered to new situations. Suppose through the process of a data mining market analysis we find that the product X is almost always purchased with product Y. A classic example of this is the discovery that an unusually high percentage of people who purchase baby diapers in Thursdays also purchase beer. An initial surprise reaction to this findings make sense when we realize that couples with a young baby at home are not likely to go out on Friday or Saturday night but instead but prefer to enjoy the weekend by relaxing at home. A market analyst can take advantage of this finding by making beer an obvious display items for customers buying diapers.

DATA MINING TASKS

It consist of two models**:** predictive model and descriptive model as shown in fig 1.1,

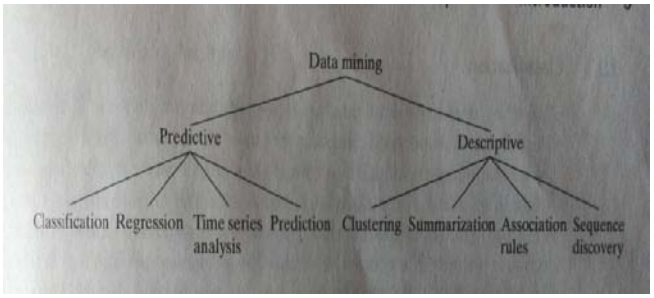

Fig 1.1 Data mining tasks

Classification

Classification maps data into predefined groups or classes. It is often referred to as supervised learning because the classes are determined before examining the data. Two example of classification application are determining whether to make a bank loan and identifying credit risks. Classification algorithms require that the classes be define based on the data attribute values. They often describe these classes by looking at the characteristics of data already known to belong to the classes. *Pattern recognition* is a type of classification where an input pattern is classified into one of several classes based on its similarity to these predefined classes.

Regression

Regression is used to map a data item to a real valued prediction variable. in actuality, regression involves the learning of the function that does this mapping. Regression that the target data fit into some known type of function (e.g., linear logistic, etc) and then determines the bet function of these type that models the given data. Some type of error analysis is used to determine which function is "best".

Time Series Analysis

With *Time Series Analysis,* the value of an attribute is examined as it various overtime. The values usually are obtained as evenly space time points (daily, weekly, hourly, etc). There are three basic functions perform in time series analysis. In one case, distance measures are used to determine the similarity between different time series. In second case, the structure of the line is examined to determine (and perhaps classify) its behavior. The third application would be to use the historical time series plot to predict future values.

Prediction

Many real-world data mining application can be seen as predicting future data states based on past and current data. *Prediction* can be viewed as a type of classification. (Note: This is the data mining task that bias different from prediction model, although the prediction task is a type of prediction model.) The difference is that prediction is predicting a future state rather than a current state. Here we are referring to a type of application rather than to a type of data mining modeling approach, as discussed earlier. Prediction applications include flooding, speech recognition, machine learning, and pattern recognition. Although future values may be predicted using time series analysis or regression techniques.

Clustering

Clustering is similar except that the groups are not predefine, but rather defined by the data alone. Clustering is alternatively referred to as unsupervised learning or segmentation. It can be thought of as partitioning or segmenting the data into groups that might or might not be disjointed. The clustering is usually accomplished by determining the similarity among the data on predefined attribute the most similar data are grouped into clusters. Since the clusters are not predefined, a domain expert is often required to interpret the meaning of the created cluster. A special type of clustering is called segmentation.With segmentation a database is partitioned into disjointed groupings of similar tuples called *segments*. Segmentation is often viewed as being identical to clustering. In other circles segmentation is viewed as a specific type of clustering applied to a database itself.

Summarization

Summarization maps data into subsets with associated simple descriptions. Summarization is also called characterization or generalization. It extracts or derives representative information about the database. This may be accomplished by actually retrieving portions of the data. Alternatively, summary type information (such as the meaning of some numeric attribute) can be derived from the data. The summarization succinctly characterizes the contents of the database.

Association Rules

Link analysis, alternatively referred to as *affinity analysis* or association, refers to the data mining task of uncovering relationships among data. The best example of this type of application is to determine association rules. An *association rule* is a model that identifies specific types of data association. These associations are often used in retail sales community to identify items that are frequently purchased together. Associations are also used in many other applications such as predicting the failure of telecommunication switches. Users of association rules must be cautioned that these are not casual relationships. They do not represent any relationship inherent in the actual data (as is true with functional dependencies) or in the real world. There probably is no relationship between bread and pretzels that causes them to be purchased together. And there is no guarantee that this association will apply in the future. However, association rules can be used to assist retail store management in effective advertising, marketing and inventory control.

Sequence Discovery

Sequential analysis or sequence discovery is used to determine sequential patterns in data. These patterns are based on a time sequence of actions. These patterns are similar to associations in the data (or events) are found to be related, but the relationship is based on time. Unlike the market analysis, which requires the items to be purchased at the same time, in sequence discovery the items purchased over time in some order. A similar type of discovery can be seen in the sequence within which data are purchased. For example, most people purchase CD players may be found to purchase CDs within one week. As we will see, temporal association rules really fall into this category.

CONCLUSION

Today's people is drowning in data but starving for knowledge. Therefore Data mining involves extracting meaningful information, rules, patterns from huge deposited data. In today's era many useful techniques are available to extract data. Data mining is a method to search hidden data and to acknowledge the entities.

AJCSE, Sep-Oct, 2016, Vol. 1, Issue 1

REFERENCE

1. DATAMININGIntroductoryand Advanced Topicsby Margaret H. Dunham, S. Sridhar.
2. Anand V . S aurkar, V aibhav Bhaujade, Priti B hagat , R eview pa per on va rious data mining techniques.
3. M. S . S ousa, M . L. Q. Mattoso a nd N . F . F.Ebecken, Data m ining: a d atabase perspective