International Journal of Computer Science and Software Engineering
Volume 1, Number 1 (2015), pp. 25-31

http://www.irphouse.com

Dynamic Data in terms of Data Mining Streams

Rupinder Kaur

Lecturer, Department of Computer Science & Engineering
Nehru Govt. College , Jhajjar.
rupi.vattu@gmail.com

ABSTRACT

Data mining is a process that takes data as input and outputs knowledge and is called KDD (Knowledge Discovery in Databases). KDD is an interdisciplinary area focusing upon methodologies for extracting useful knowledge from data.Data mining involves an integration of techniques from multiple disciplines such as database technology, statistics, machine learning, high-performance computing, pattern recognition, neural networks, data visualization, information retrieval, etc.It is technique used as association rule for detecting relationships or associations between specific values of categorical variables in the large data sets. From the enormous amount of record that is needed to be read and to extract the rules, if the inquest is for finding the new data or a need to modify some or a set of data is required during the process of data mining. As per the previous scenario, user needs to repeat the whole procedure, which is not only assured to be time consuming but lack in its efficiency. Hence forth, the significance of dynamic data mining process appears which is considered as a topic of my research. So the purpose of this study is to find a solution for dynamic data mining process which implicates all updates (insert, update and delete operations) into account.

1. INTRODUCTION

Historically, the notion of finding useful patterns in data has been given a variety of names, including data mining, knowledge extraction, information discovery, information harvesting, data archaeology, and data pattern processing. The term data mining has mostly been used by statisticians, data analysts, and the management information systems (MIS) communities. It has also gained popularity in the database field. Data mining involves an assimilation of techniques from multiple areas such as database technology, statistics, machine learning, neural networks, information retrieval, etc. Also defines Data mining as “the analysis of observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both

understandable and useful to the data owner". Data mining set the stage for KDD in two important ways: (1) data cleaning and (2) data access. Data cleaning: As organizations are forced to think about a unified logical view of the wide variety of data and databases they possess, they have to address the issues of mapping data to a single naming convention, uniformly representing and handling missing data, and handling noise and errors when possible. Data access: Uniform and well-defined methods must be created for accessing the data and providing access paths to data that were historically difficult to get to (for example, stored offline). The architecture of a typical data mining system may have the following major components : database, data warehouse, or other information repository; a server which is responsible for fetching the relevant data based on the user's data mining request, knowledge base which is used to guide the search.

The basic Data Mining Tasks consists of a number of processes:

- **Association analysis**
- **Classification**
- **Cluster analysis**
- **Class description**

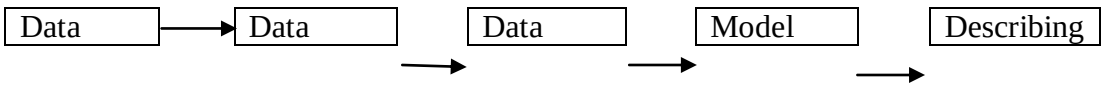
Data mining is one of the most hot research fields that are due to the stretch of both computer hardware and software technologies, which has enforce organizations to trust in these technologies. Data is considered as the top most asset of any organization, henceforth this asset should be used to envision forth coming decisions. Databases tend to be large and dynamic thus, their contents usually do change; new information might need to be inserted, current data might need to be updated and/or deleted. The problem with this from the data mining perspective is how to ensure that the rules are up-to-date and consistent with the most current information.

1.1 Static Data Mining Process

Data mining process is a step in Knowledge Discovery Process consisting of methods that produce useful patterns or models from the data . In some cases when the problem is known, correct data is available as well, and there is an attempts to find the models or tools which will be used, some problems might occur because of duplicate, missing, incorrect, outliers values and sometimes a need to make some statistical methods might arise as well.

The KDD procedures are explained bellow in a way to help us focus on data mining process. It includes five processes:

1) Data mining problem definition
2) Data Collection
3) Data detection and correction
4) Model estimation and building
5) Describing and validating model (as seen in Figure.1)



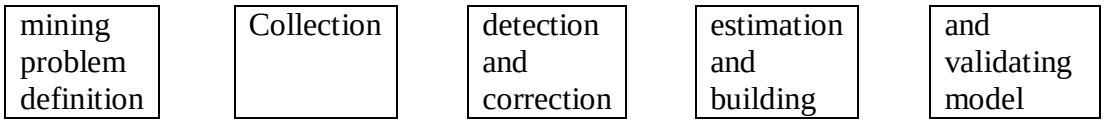


Figure 1. Data mining process

1.1 Data Mining Problem Definition

Before we can use data mining models and algorithms we have to find the most suitable strategy. In order to do that, we have to detect the problem type. Usually, data mining project involves a combination of different problem types, which together solve the problem. In this step, a modeler usually specifies a set of variables for the unknown dependency and, if possible, a general form of this dependency as an initial hypothesis. There may be several hypothesis formulated for a single problem at this stage.

1.2 Data Collection

This process is concerned with the data collection from different sources and locations. The process of gathering and measuring information usually with software.There are loads of different data collection techniques and procedures, but when you're talking about it in terms of Big Data (which most buzzword lovers are) they usually mean electronic (or online) data collection.

1.3 Data Detection and Correction

All raw data sets which are initially prepared for data mining are often huge; many are related to humans and have the potential for being messy. Real-world databases are subject to noise, missing, and inconsistent data due to their typically huge size, often several gigabytes or more. Data preprocessing is commonly used as a preliminary data mining practice. It transforms the data into a format that will be easily and effectively processed by the users.

1.4 Model estimation and building

This process includes four parts:

1. Task (s) Selection

Selecting which task to use depends on the model whether it is predictive or descriptive. predictive models predict the values of data using known results and/or information found large data sets, historical data, or using some variables or fields in the data set to predict unknown, classification, regressions, time series analysis, prediction, or estimation are tasks for predictive model. A descriptive model identifies patterns or relationships in data and serves as a way to explore the properties of the data examined.

2. Data mining method (s) Selection

Once the task is decided and goals are codified, a concrete method (or set of methods) needs to be chosen for searching patterns in the data. Depending on the choice of

techniques, parameter optimization may or may not be required.There are number of methods for model estimation includes these but not limited to neural networks, Decision trees, Association Rules, Genetic algorithms, Cluster Detection, Fuzzy Logic.

3. Selection of suitable algorithm

The next step is to construct a specific algorithm that implements the general methods. All data mining algorithm include three primary components these are:

(1) Model representation,
(2) Model evaluation, and
(3) Search.

4. Extracting knowledge

This is the last step in building the model which is the results (or the answers for the problem solved in data mining) after making the simulation for the algorithm. This can be best explained by presenting an example of Auction Fraud.

1.5 Model Description, Validation

In all cases, data mining models should assist users in decision making. Hence, such models need to be interpretable in order to be useful because humans are not likely to base their decisions on complex "black-box" models; the goals of the accuracy of the model and accuracy of its interpretation are somewhat contradictory. Modern data mining methods are expected to yield highly accurate results using high dimensional models. The ultimate goal of a data mining process should not be just to produce a model for a problem at hand, but to provide one that is sufficiently credible, acceptable and implemented by the decision-makers; this type would need to consider all the data i.e. using a dynamic database.

2. Dynamic Data Mining Process

Many researchers and developers have specified a process model designed to guide the user through a sequence of steps that will lead to good results. Many have been reported for data mining process. Some of these assumed that this is possible for Dynamic data mining process. Up-to-date most of the data mining projects have been dealing with verifying the actual data mining concepts. Since this has now been established most researchers will move into solving some of the problems that stand in the way of data mining, this research will deal with such a problem, in this case the research is to concentrate on solving the problem of using data mining dynamic databases. Crespoa, and Weberb presented a methodology for dynamic data mining using fuzzy clustering that assigns static objects to dynamic classes. Changes that they have studied are movement, creation, and elimination of classes and any of their combination.One of the problem arises is of query processing, specifically on how to define and evaluate continuous queries over data streams, address semantic issues as well as efficiency concerns; they specified a general and flexible architecture for query processing in the presence of data streams. They also used their basic architecture as tool to clarify alternative semantics and processing techniques for

continuous queries. In many situations, new information is more important than old information, such as in publication database, stock transactions, grocery markets, or web-log records. Consequently, a frequent item set in the dynamic database is also important even if it is infrequent in the updated database.

2.1 Defining the Data Mining Problem

Since this process is concerned with the definition of the problem and the data is not yet built, no change will be carried out in this step, unless the analyst changes the problem goals.

2.2 Collecting the Data Mining Data

Data is collected from different sources and/or locations, if we assume that an update was carried out on the date after the data was collected by the algorithm, the following will take place:

1) If a new source of data (new database) and the data included in this database is a main source, this source will be used. This can be achieved by recollecting this data again, and/or replace existing data partially or totally.

2) New updated data (insert, update, delete) If the data used is changed in any way during a current run, then the data being used will be considered invalid and the new version of this data should be collected, a new run will be initiated, with the same source of data, but it must update the data immediately either by inserting new data, updating current data, or deleting data completely.

2.3 Detecting and Correcting the Data

If it is realized that the current data is a new source of data and there is a need to recollect the data, this takes us back to the Collecting data process or if it is a new source of data and there is a need to include this data then, there is no need to go back, from the beginning, it should contain the new source of data and decide which of the new data items it needs; it should then carry out data cleaning to remove noise, correcting inconsistencies, integration Afterwards, it simply combines the new set of data (the detected and corrected data) with the main data. Note here some times after combining data one could have another detecting and correcting process i.e. (data reduction or integration) but this will not flow the same procedures as starting from scratch or if it is a new updated version of the data, the steps mentioned above can be used iff it is necessary and then combine the new updated data with the main data, taking into consideration with inserting new record, updating, and/ or deleting an existing record.

2.4 Estimating and Building the Model

In estimating and building the model process in dynamic data mining; there are three main parts : task (s) selection, select Data Mining method (s), and selecting the suitable algorithm that will not change if a new version of data is introduced (since they are concerned with the definition of the problem). But in this case the extraction knowledge part will change depending on the new version of the data.

2.5 Model Description and Validation

Modern data mining methods are expected to yield highly accurate results using high dimensional models, with new updated version of the data, it will surely change the results; say for example the number of rules in association rules could be changed; it may add a new rules, replace existing number of rules or change the percentage of the rules, and this change could affect the decision making process.

3. Conclusion

On the account, dynamic data is very essential for making right decision. The streaming of data can be useful for data mining and is still in its infancy. Up-to-date most of the data mining projects have been dealing with verifying the actual data mining concepts. Thus, we can reflect the ongoing changes of insertion, deletion and modification of the data with the intellectual use of techniques on data streams.

References:

[1] Maria Halkidi, "Quality assessment and Uncertainty Handling in Data Mining Process" http://www.edbt2000.unikonstanz.de/phd-workshop/papers/Halkidi.pdf.

[2] Fayyad, U. M., G. P. Shapiro, P. Smyth. "From Data Mining to Knowledge Discovery in Databases", 0738-4602-1996, AI Magazine (Fall 1996): 37–53.

[3] Jiawei Han, Micheline Kamber. "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers, Champaign:CS497JH, fall 2001, www.cs.sfu.ca/~han/DM_Book.html.

[4] Claude Seidman. "Data Mining with Microsoft SQL Server 2000 Technical Reference", ISBN:0-7356-1271-4,amazon.com/Mining-Microsoft-Server-Technical-Reference/dp/0735612714.

[5] David Hand, Heikki Mannila, Padhraic Smyth. "Principles of Data Mining", ISBN: 026208290 MIT Press, Cambridge,MA, 2001.

[6] Fernando Crespoa, Richard Weberb. "A methodology for dynamic data mining based on fuzzy clustering", Fuzzy Sets and Systems 150 (2005) 267–284.

[7] Papadimitriou, J.Sun, C.Faloutsos, "Streaming Pattern Discovery in Multiple Time-Series", Proceedings of the 31st VLDB Conference, Trondheim, Norway, 2005,p697-708.

[8] Mohamed Medhat Gaber, Arkady Zaslavsky and Shonali Krishnaswamy. "Mining Data Streams: A Review", VIC3145, Australia, ACM SIGMOD Record Vol. 34, No. 2; June 2005.

[9] Two Crows Corporation. "Introduction to Data Mining and knowledge Discovery", ISBN: 1-892095-02-5

[10] Ruoming Jin and Gagan Ag awal. "A Middleware for Developing Parallel Data Mining Applications", Proc. of the 1-stSIAM Conference on Data Mining, 2000 - cs.ubc.ca.