DESIGN AND QUANTIFICATION OF PRIVACY PRESERVING DISTRIBUTED DATA MINING

PAWAN KUMAR

Research Scholar
Presently working as Asstt. Prof. Computer Science at HRIT
Ghaziabad

ABSTRACT

The goal of data mining is to extract or “mine” knowledge from large amounts of data. Privacy, legal and commercial concerns restrict centralized access to this data. Theoretical results from the area of secure multiparty computation in cryptography prove that assuming the existence of trapdoor permutations, one may provide secure protocols for any two-party computation as well as for any multiparty computation with honest majority.

INTRODUCTION

Data mining is a field that has emerged in response to analysis of large data sets. Data mining involves various statistical methods for searching relationships among variables. Data Mining is the analysis of observational data sets to find un-suspected relationships and to summarize the data in novel ways that are both understandable and useful to the owner. Data characterization is the summarization of the general characteristics or features of a target class of data. Data Discrimination, on the other hand, is a comparison of the general features of target class data objects with the general features of objects from one or a set of contrasting classes. Association analysis is the discovery of association rules showing attribute- value conditions that occur frequently together in a given set of data. Classification is the process of finding a set of models (or functions) that describe and distinguish data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown.

REVIEW OF LITERATURE

This chapter reviewed literature of past and contemporary research on data mining. Important data mining and statistical terminologies will be reviewed. Next, the supervised learning methods for data mining were discussed. Then, the data mining techniques that were applied to determine the important of the simulation conditions and characteristics of the procedures in the generation of p-values were illuminated.

Distributed Data Mining

In contrast to the centralized model, the Distributed Data Mining (DDM) model assumes that the data sources are distributed across multiple sites. Algorithms developed within this field address the problem of efficiently getting the mining results from all the data across these distributed sources. Since the primary focus is on efficiency, most of the algorithms developed to date do not take security consideration into account. However, they are still useful in framing the context of the thesis. A simple approach to data mining over multiple sources that will not share data is to run existing data mining tools at each site independently and combine the results. However, this will often fail to give globally valid results. Issues that cause a disparity between local and global results include:

- Values for a single entity may be split across sources. Data mining at individual sites will be unable to detect cross-site correlations.

- The same item may be duplicated at different sites, and will be over-weighted in the results.
- Data at a single site is likely to be from a homogeneous population. Important geographic or demographic distinctions between that population and others cannot be seen on a single site.

RESEARCH METHODOLOGY

Data mining tools can answer business questions that traditionally are time consuming to resolve. In general, wherever data exist, powerful data mining techniques can help reveal important data patterns that would otherwise remain unnoticed when using simple type of analysis. Thus, data mining are applicable in many areas (refer to Figure 1.1)

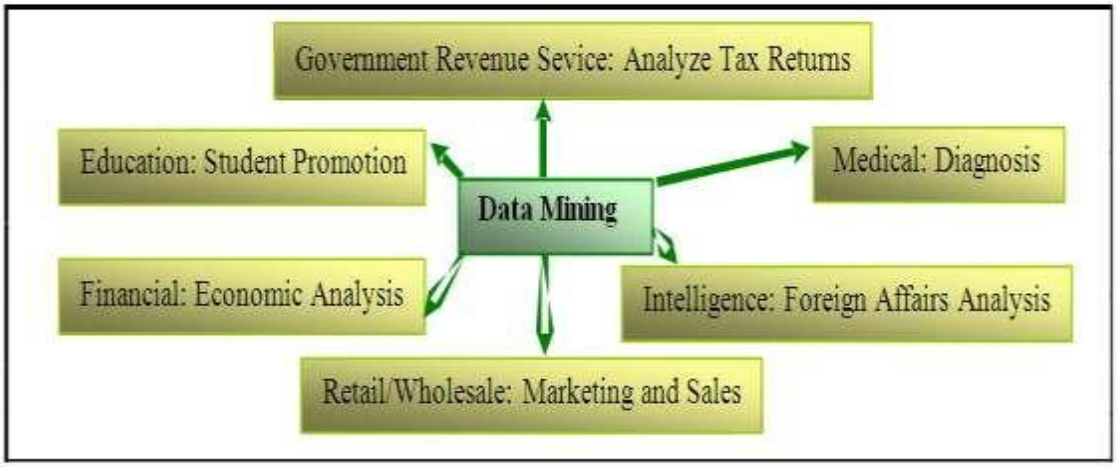
Government Revenue Sevice: Analyze Tax Returns
Education: Student Promotion
Data Mining
Medical: Diagnosis
Financial: Economic Analysis
Intelligence: Foreign Affairs Analysis
Retail/Wholesale: Marketing and Sales

Figure 1.1. Some data mining application areas.

Data mining techniques can yield the benefits of automation on existing software and hardware platforms, and can be implemented on new systems as existing platforms are upgraded and new products developed. When data mining tools are implemented on high performance parallel processing systems, it can analyze massive databases in minutes. Faster processing means that more models can be automatically experiment to understand complex data. Therefore, high speed makes it practical to analyze huge quantities of data. In addition, via data mining, it is possible to discover patterns and build models automatically. The models are both descriptive and prospective. They explain why things happened and predict the future. "What-if" questions can be posted to a data-mining model that cannot be queried directly from the database.

On the other hand, visualization of the data mining output in a meaningful way allows analysts to see the data mining results. Visualization will enable analysts to see plausible relationship between variables that were tested (Thearling, 1997).

Data Mining Procedures

The steps in data mining are important keys to a successful data mining of a data set. Data mining projects require substantial initial effort in data preparation. In particular, the knowledge discovery process in databases consists of several steps. The overall statistical process, from data sources to model application involved the following data mining process.

- Data Cleaning and Data Quality Assessment
- Data Integration and Consolidation
- Data Selection

- Data Transformation
- Data Mining
- Pattern Evaluation
- Knowledge Presentation

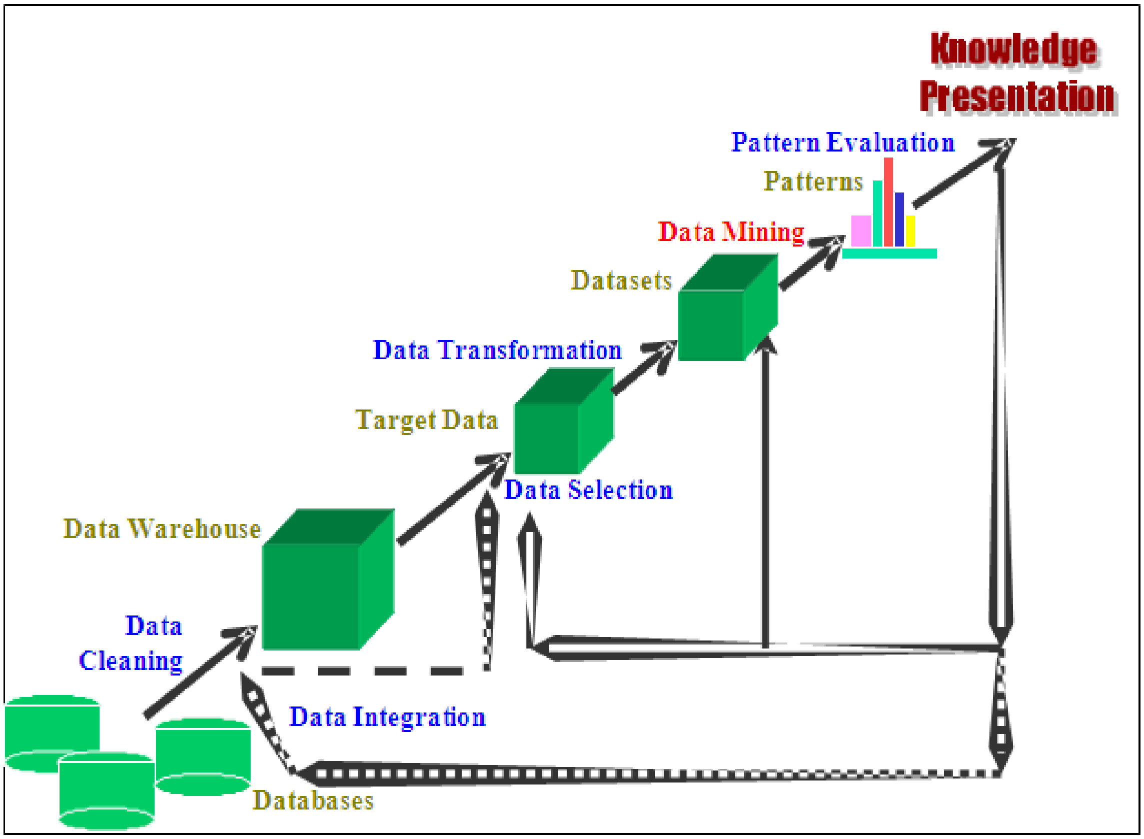

Figure 1.2. An overview of data mining procedures.

SCOPE OF THE STUDY

Data Selection

In the data selection step, this research aims to extract information needed by the data mining step from the database. Hence, data which are relevant to the analysis task are retrieved based on the objectives of this study. Base on one‟ s data selection criteria, one may wish to include or exclude other sets of data. Data selection is important because the existence of a lot of data for the data mining process. If inappropriate data were retrieved, the process could be time consuming and more money will be spend and also increase risk. Data selection is different from sampling the database and choosing predictor variables. It is an elimination of irrelevant or unneeded data during data analysis to achieve the research objectives.

Data Transformation

Data transformation is the application of a mathematical modification to the values of a variable.

Data Mining

Having very large databases is becoming standard practice. Data mining is an intelligent method applied to extract data patterns, i.e. applying a concrete algorithm to find useful and novel patterns in the data. Generally, data mining is the stage of finding universal patterns or principles that summarize and explain a set of observations.

CONCLUSION

Two primary goals of this research were:

- Create a database of all p-values for tests of spread procedures from Keselman, et al. (in press) and
- Determine important simulation conditions and characteristics of the spread procedures that correspond to the target p-value of 0.05 or a set of value that is „close" to 0.05 using the predictive modeling approach.

REFERENCES

[1] Security-control methods for statistical databases: A comparative study. ACM Computing Surveys, 21(4):515–556, December 1989.

[2] Information sharing across private databases. In Proceedings of ACM SIGMOD International Conference on Management of Data, San Diego, CA, June 9-12 2003.

[3] Ramakrishnan Srikant. Privacy-preserving data mining. In Proceedings of the 2000 ACM SIGMOD Conference on Management of Data, pages 439–450, Dallas, TX, May 14-19 2000.

[4] V. Verykios. Disclosure limitation of sensitive rules. In Knowledge and Data Engineering Exchange Workshop (KDEX'99), pages 25–32, Chicago, IL, November 8 1999.

[5] K. Yelick. Runtime Support for Portable Distributed Data Structures. Workshop on Languages, Compilers, and Runtime Systems for Scalable Computers, May 1995.

[6] Consistent Linear Speedups to a First Solution in Parallel State-Space Search. In Proceedings of the 1990 National Conf. on Artificial Intelligence AAAI-90, July 1990.