Rising Expenses on Data Mining

Kunal Deswal & Shweta Thakur

Dronacharya College of Engineering
Gurgaon, Harayana, India
kunaldeswal1993@gmail.com shwetathakur2424@gmail.com

ABSTRACT

Data mining is the new term relative to the technique used. For years, Associations have used powerful machines and computers to inquire through the volumes of supermarket scanner data and to analyze the research reports of market. Privacy issues hold the importance in data mining. If we talk about vertically partitioned, various data mining problems can be minimized to securely evaluating the scalar product. Association rule mining can be mentioned over vertically partitioned data, among these problems. Adaptability of a secure scalar product can be calculated by overhead of communication needed to assure this security. In vertically partitioned data, a few solutions have been suggested for privacy preserving association rule mining. Beside it, the excessive overhead communication required for assuring the privacy of data is the main drawback of these solutions. In this paper, we will target the issues related to privacy and new secure scalar product with an objective to minimize the expensive communication.

Keywords
Data Mining; Pre-processing; Data Mining Process; Clustering, Privacy; Overhead

1. INTRODUCTION

1.1 Definition

Data mining is a vast sub-field of computer science which means process of discovering patterns with help of computers, in large data sets involving methods at the intersection of artificial intelligent system, statistics, machine learning and database systems. The global objective of the data mining process is to obtain information from a data volume and transform it into a typical structure which could be easily understood and used further.

1.2 Overview

Data mining is also called as Data discovery or knowledge discovery. It is the process of examining data from different views or prospects and giving brief statements of main and useful information. Technically, data mining is the process of finding relations among various fields in large databases.

2. DATA MINING

2.1 What it does?

Data mining is chiefly used today by firms with a strong consumer focus - retail, financial, communication, and marketing organizations. Data mining enables these companies to resolve relationships among "internal" factors such as price, product

positioning, or staff skills, and "external" factors such as economic indicators, competition, and customer demographics. It facilitate them to determine the impact on sales, customer satisfaction, and corporate profits. Finally, it enables them to summarize the important information to view detail transactional data. With data mining, a retailer can use point-of-sale records of customer purchases to send targeted promotions based on an individual's purchase history. Through mining demographic data from comment or warranty cards, the retailer can develop products and promotions to appeal to specific customer segments. Let's take an example; Blockbuster Entertainment mines its video rental history database to recommend rentals to individual customers. Based on analysis of the monthly expenditures, American Express can suggest products to its cardholders. WalMart is beginning massive data mining to transform its supplier relationships. WalMart abducts point-of-sale transactions from over 2,900 stores in 6 countries and continuously transmits this data to its enormous 7.5 terabyte Teradata data warehouse. It allows more than 3,500 suppliers, to approach data on their products and implement data analyses. This data is used by suppliers to identify customer buying patterns at the store display level. They use this information to manage local store inventory and identify new merchandising opportunities. In 1995, WalMart computers were processed over 1 million complex data queries. The National Basketball Association (NBA) is exploring a data mining app that could be used in conjunction with image recordings of basketball games. The Advanced Scout software analyzes the movements of players to help coaches in coordinating and managing plays and strategies. For example, an analysis of the play-by-play sheet of the game played between the New York Knicks and the Cleveland Cavaliers on January 6, 1995 revealed that when Mark Price played the Guard position then John Williams attempted four jump shots and made each one! Advanced Scout not only able to find this pattern, but it also explained that it is interesting because it differs considerably from the average shooting percentage of 49.30% for the Cavaliers during that game. By using the NBA universal clock, the coach could automatically brought up the video clips showing each of the jump shots attempted by Williams with Price on the floor, without needing to comb through hours of that video footage. Those clips showed a very successful pick-and-roll play in which Price drew the Knick's defense and then found Williams for an open jump shot.

2.2. How does it work?

While large-scale information technology has been evolving separate transactions and analytical systems, data mining provides the link between the two. Data mining software analyze relationships and patterns in stored data based on open-ended user queries. Several types of analytical software are available. Generally, these types of relationships are sought:

2.2.1. Classes

Data is located in predetermined groups using the stored data. For example, a restaurant chain can mine customer purchase data to find out when customers visit and what they typically order. This information can be used to increase traffic by having daily specials.

2.2.2. Clusters

According to logical relationships or consumer preferences, data items are grouped. For example, To identify market segments or consumer affinities, data can be mined.

2.2.3. Associations

Data can be mined to identify associations. The best example of associative mining would be The beer-diaper example.

2.2.4. Sequential patterns

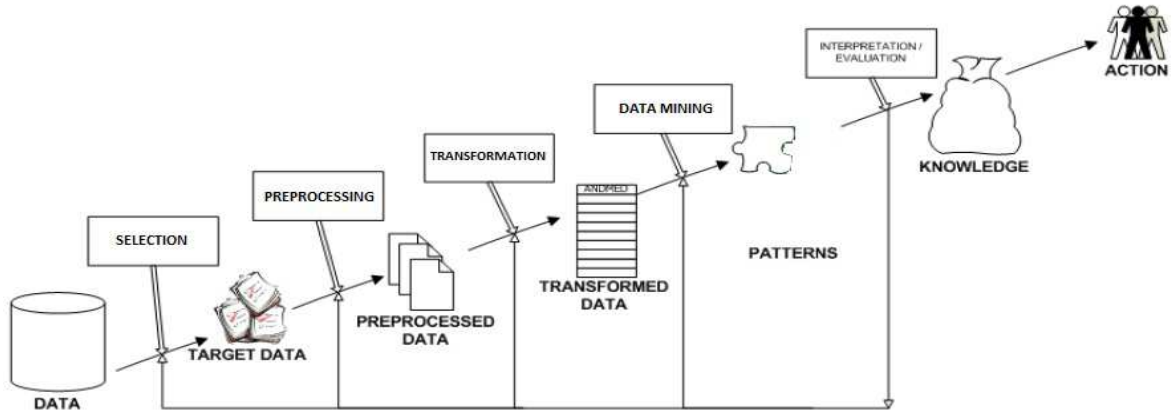
Data is mined to forecast behavior patterns and trends. For example, an outdoor equipment retailer can predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

Data mining consists of five major elements:

1. Extraction, transformation, and loading transaction data onto the data warehouse system.
2. Storing and managing the data in a multidimensional database system.
3. Providing data access to business analysts and information technology professionals.

Different levels of analysis are available:

1. Artificial neural networks: Non-linear predictive models which learn through training and resemble to biological neural networks in structure.
2. Genetic algorithms: The techniques of optimization which uses process such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.
3. Decision trees: These are Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a data volume. Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID) are included in Specific decision tree methods. CART and CHAID are decision tree approaches used for classification of a data set. They provides a set of rules which can be applied to a new (unclassified) dataset to forecast which record will have a given outcome. CART divides a dataset by creating 2-way splits while CHAID segments using chi square tests to create multi-way splits. CART typically need less data preparation than CHAID.



4. Analyzing the data by application software.
5. Presenting the data in a useful format, such as a graph or table.

4. Nearest neighbour method: It is a technique which classify each record in a dataset based on a combination of the classes of the k record(s) most alike to it in a

historical dataset (where k1). Sometimes called the k-nearest neighbor technique.
5. Rule induction: The extraction of useful if-then rules from data based on statistical significance.
6. Data visualization: The visual interpretation of complex relationships in multidimensional data. To illustrate data relationships, Graphics tools are used.

3. DATA MINING PROCESS

The Knowledge Discovery in Databases (KDD) process is commonly defined with the stages:
Selection
Pre-processing
Transformation
Data Mining
Interpretation/Evaluation

3.1 Pre-Processing

A target data set must be assembled, before data mining algorithms could be used. As data mining can only find patterns actually present in the data, the target data volume must be large enough to consist of these patterns while remaining concise enough to be mined within an acceptable time limit. A common source for data is a data mart or data warehouse. Pre-processing is necessary to analyze the multivariate data sets before data mining. The target set is then cleaned. Cleaning data helps in removing the observations containing noise and those with missing data.

3.2 Data Mining

Data mining involves six common classes of tasks:
Anomaly detection (Outlier/change/deviation detection) – The recognition of unusual data records, that may be interesting or data errors that need further investigation. Association rule learning (Dependency modelling) – It searches for relationships between variables. For example a supermarket may gather data on customer purchasing habits. Using association rule learning, the supermarket can find which products are frequently bought together and it can use this information for marketing purposes. Sometimes it is also referred to as market basket analysis.

3.2.1. Clustering
Clustering is the task of finding groups and structures in the data which are in some way or another "similar", without using known structures in the data.

3.2.2. Classification
It is the task of generalizing known structure for applying to new data. For example, an e-mail program may attempt to classify an e-mail as "legitimate" or as "spam".

3.2.3. Regression
It attempts to find a function which can model the data with the least error.

3.2.4. Summarization
It provides a more compact representation of the data set, including visualization and report generation.

3.2.5. Sequential pattern mining

Sequential pattern mining discover sets of data items which appear together frequently in some arrangement. Sequential pattern mining extracts the frequent subsequence from a sequence database. It has attracted a great deal of interest during the recent data mining research because it is the only basis of many applications, such as: web user analysis, stock trend prediction, DNA sequence analysis, finding language or linguistic patterns from natural language texts, and using the history of symptoms to predict certain kind of disease.

3.3. Result Validation

The final step of knowledge discovery from data is to check that the patterns produced by the data mining algorithms appear in the wider data set. Not all patterns found by the data mining algorithms are necessarily valid. It is common for the data mining algorithms to discover patterns in the training set which are not present in the general data set. This is what overfitting is called. The evaluation uses a test set of data to overcome this, on which the data mining algorithm was not trained and the resulting output is compared to the desired output.

For example, a data mining algorithm which is trying to differentiate "spam" from "legitimate" emails will be trained on a training set of sample e-mails. After being trained, the learned patterns will be applied to the test set of e-mails on which it had not been trained. The certainity of the patterns can then be calculated from how many e-mails they correctly

classify. A number of statistical methods may be used to evaluate the algorithm, such as ROC curves. If the learned patterns do not meet the desired standards, subsequently it is necessary to re-evaluate and change the preprocessing and data mining steps. If the learned patterns do meet the aspired standards, then the final step is to interpret the learned patterns and transforming them into knowledge.

4. PRIVACY ISSUES AND INCREASING OVERHEAD

4.1. Data mining for safety applications

Data mining is applying an applicable key technology for identifying doubtful activities. In this section, data mining will be discussed with respect to use in both ways for non real-time and for real-time applications. One have to gather data from several sources in order to complete data mining for counter terrorism activities. For example, the consecutive information on revolutionary attacks is desired at the very least: who, what, where, when, and how; personal and business data of the possible terrorists: place of birth, religion, education, ethnic origin, work history, finances, criminal record, relatives, friends and associates, and travel history; unstructured data: newspaper articles, video clips, dialogues, e-mails, and phone calls. The data needs to be included, warehoused and mined. One wants to draw depiction of terrorists, and activities/threats. The data needs to be mined to take out arrangements of possible terrorists and predict future activities and goals. Fundamentally one wants to find the “needle in the haystack”. Data integrity is essentially required and also

its methods have to SCALE. For several apps such as urgent situation response, one is required to complete real-time data mining. Data will be fetched from sensors and other strategy is that the form of nonstop data streams together with breaking news, video cassette releases, and satellite images. Caches may also consist of some serious data. One wants to quickly analyze and filter through the data and remove unnecessary data for shortly use and analysis (non-real-time data mining). The techniques of data mining require meeting timing restriction and may have to stick the quality of service (QoS) tradeoffs among suitability, accuracy and precision. The results have to be accessible and visualized in real-time. Additionally, alerts and triggers will also have to be employed. We need to first find out what our present capabilities are in order to efficiently apply data mining for safety applications and to develop suitable tools. For example, perform the profitable tools balance? Do they effort only on particular data and limited cases? Do they carry what they promise? We need a balanced objective study with display. At the same time, we also need to work on the large picture. For example, what do we want the data mining tools to carry out? What are our end results for the predictable future? What are the levels of quality for achievement? How do we evaluate the data mining algorithms? What test beds do we build? We need both a short-term as well as longer-term resolutions. For the future, we require affecting present efforts and filling the gaps in an objective aimed way and complete technology transfer. For the longer-term, we need a research and development depiction. In summary, data mining is very helpful to settle or find out a solution to security troubles. Tools can be utilized for audit data to discover any shortcomings and flag irregular behavior. There are many recent works on applying data mining to cyber safety applications, Tools are being inspected thoroughly in order to find out irregular patterns or arrangements for national security together with those based on categorization and link analysis. For fraud exposure and crime solving, law enforcement is also using these kinds of tools.

5. APPLICATIONS

5.1. Science and engineering

Recently, data mining has been used widely in the areas of science and engineering, such as bioinformatics, genetics, medicine, education and electrical power engineering. In the study of human genetics, mining helps to address the important objective of perceiving the intended meaning of the mapping relationship between the inter-individual variations
in human DNA sequence and the variability in disease susceptibility. In simple terms, it aims to discover out how the changes in an individual's DNA sequence alters the risks of developing common diseases such as cancer, which is of great significance to enhance methods of diagnosing, preventing, and treating these diseases. The data mining method which is used to perform this task is known as multifactor dimensionality reduction. In the area of electrical power engineering, data mining methods have been widely used for condition supervising of high voltage electrical equipment. The purpose of condition supervising is to attain valuable information on, for instance, the status of the insulation (or other important safety-related

parameters). Data clustering techniques – like the self-organizing map (SOM), have been enforced to vibration monitoring and analysis of transformer on-load tap-changers (OLTCS). Using vibration monitoring, it can be supervised that each tap change operation generates a signal that contains information about the condition of the tap changer contacts and the drive mechanisms. Definitely, different tap positions will produce different signals. However, there was a huge variability amongst normal condition signals for exactly the same tap position. SOM has been applied to catch abnormal conditions and to speculate about the nature of the abnormalities.

5.2. Business

Data mining is the detailed examination of elements of historical business activities, stocked as static data in data warehouse databases, to acknowledge hidden patterns and trends. Advanced pattern recognition algorithms is used by Data mining software to sift through large amounts of data in order to assist in discovering previously unknown strategic business information.

5.3. Games

Since the early 1960s, with the availability of oracles for certain combinatorial games, also called table bases (e.g. for 3x3-chess) with any beginning configuration, small-board dots-and-boxes, small-board-hex, and certain endgames in chess, dots-and-boxes, and hex; a new area for data mining has been unlocked. This is the derivation of human-usable plans of action from these oracles. Presently, pattern recognition approaches do not seem to fully attain the high level of extraction required to be applied successfully.

6. RELIABILITY/VALIDITY

Data mining can be used in wrong way for wrong purposes and can also accidentally produce results which appear important but which do not actually predict future behavior and cannot be reproduced on a new sample of data. See Data dredging.

7. PRIVACY CONCERNS

Some people rely on that data mining itself is socially neutral. While the term "data mining" has no ethical meanings, it is often related with the mining of information in relation to peoples' behavior (ethical and otherwise). To be accurate, data mining is a statistical method which is applicable to a set of information (i.e., a data set). Associating these data sets with people is an extreme narrowing of the types of data that are available. Examples could vary from a set of crash test data for passenger vehicles, to the performance of a group of stocks. These types of data volumes make up a great relative amount of the information available to be acted on by data mining methods, and hardly have ethical concerns associated with them. However, the ways in which data mining can be used can in some cases and contexts raise questions regarding privacy, legality, and ethics. In peculier, data mining government or commercial data sets for national security or law enforcement purposes, such as in the Total Information Awareness Program or in ADVICE, has increased privacy concerns.

8. CONCLUSIONS

Data mining is the new term relative to the technique used. For years, Associations have used powerful machines and computers to

inquire through the volumes of supermarket scanner data and to analyze the research reports of market. However, regular innovations in computer processing power, disk storage, and statistical software are dramatically rising the accuracy of analysis while driving down the cost. In this paper we have also examined data mining applications in security and their implications for privacy. We have examined the idea of privacy and then discussed about the developments particularly those on privacy preserving data mining. We then presented an agenda for research on privacy and overheads on data mining. Here are our conclusions. There is no collective definition for privacy, each organization must clear cut what it indicates by privacy and develop suitable privacy policies. Technology only is not enough for privacy; we require Technologists, Policy expert, Legal experts and Social scientists to put work on Privacy. Some well acknowledged people have believed 'Forget about privacy" Therefore, should we follow research on Privacy? We trust that there are attractive research problems; therefore we are required to carry on with this research. Moreover, some privacy is better than nil. One more school of consideration is trying to avoid privacy destructions and if destructions take place then put on trial. We are required to put into effect suitable policies and inspect up the legal aspects. Privacy from all directions is needed to be undertaken.

9. SUMMARY OF REVIEW

Although data mining is a relatively new term, the technology is not. There is no collective definition for privacy, each organization must clear cut what it indicates by privacy and developed suitable privacy policies. Technology only is not enough for privacy; we require Technologists, Policy expert, Legal experts and Social scientists to put effort on Privacy. Some well acknowledged people have believed 'Forget about privacy". Therefore, should we pursue research on Privacy? We trust that there are attractive research problems; therefore we need to carry on with this research. Additionally, some privacy is much better than nil.

10. FUTURE ISSUES

Data mining is primarily used today by companies with a strong consumer focus - retail, financial, communication, and marketing organizations. It authorize these companies to determine relationships among "internal" factors such as price, product positioning, or staff skills, and "external" factors such as economic indicators, competition, and customer demographics. And, it allows them to find the impact on sales, customer satisfaction, and corporate profits. Finally, it enables them to "drill down" into summary information to view detail transactional data. With data mining, a retailer could use point-of-sale records of customer purchases to send targeted promotions based on an individual's purchase history. By mining demographic data from comment or warranty cards, products and promotions can be developed by retailer to appeal to specific customer segments.

11. DISCLOSURE STATEMENT

No financial support was taken for this research work from the funding agency.

12. ACKNOWLEDGMENTS

Much thanks to our guide for his constructive criticism, and assistance towards the successful completion of this research work.

13. REFERENCES

[1]. Azevedo, In Proceedings of the IADIS European Conference
on Data Mining, 2008.
[2]. Bouckaert, Remco R.; Frank, Eibe; Hall, WEKA Experiences
with a Java open-source project, Journal of Machine
Learning Research, 2011
[3]. Clifton, Christopher, Encyclopædia Britannica: Definition of
Data Mining, 2010.
[4]. Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth,
Padhraic., From Data Mining to Knowledge Discovery in
Databases, 1996
[5]. Günnemann, Stephan; Kremer, Hardy; Seidl, Thomas, An
extension of the PMML standard to subspace clustering
models, 2011
[6]. Hastie, Trevor, Tibshirani, Robert Friedman, Jerome, The
Elements of Statistical Learning: Data Mining, Inference,
and Prediction, 2009
[7]. Kantardzic, Mehmed, Data Mining: Concepts, Models,
Methods, and Algorithms, 2003
[8]. Lukasz Kurgan and Petr Musilek, A survey of Knowledge
Discovery and Data Mining process models, 2006
[9]. Mena, Jesús, Machine Learning Forensics for Law
Enforcement, Security, and Intelligence, 2011
[10]. Óscar Marbán, Gonzalo Mariscal and Javier Segovia, A Data
Mining & Knowledge Discovery Process Model. 2009
[11]. Piatetsky-Shapiro, Gregory; Parker, Lesson: Data Mining,
and Knowledge Discovery: An Introduction, 2011
[12]. Witten, Ian H., Frank, Eibe, Hall, Mark A, Data Mining:
Practical Machine Learning Tools and Techniques (3 ed.),
201