ISSN 2278-8808 SJIF 2015: 5.403
An International Peer Reviewed & Referred
SCHOLARLY RESEARCH JOURNAL
FOR INTERDISCIPLINARY STUDIES

SAFETY ISSUES IN DATA MINING

Patki Ravi Prakashrao[1] & Piyush Pandey[2], Ph. D

Abstract

The development in data mining technology brings serious threat to the individual information. The objective of privacy preserving data mining (PPDM) is to safeguard the sensitive information contained in the data. The unwanted disclosure of the sensitive information may happen during the process of data mining results. In this study we identify four different types of users involved in mining application i.e. data source provider, data receiver, data explorer and determiner decision maker.We differentiate each type of user's responsibilities and privacy concerns with respect to sensitive information. We'd like to provide useful insights into the study of privacy preserving data mining. This paper presents a comprehensive noise addition technique for protecting individual privacy in a data set used for classification, while maintaining the data quality. We add noise to all attributes, both numerical and categorical, and both to class and non-class, in such a way so that the original patterns are preserved in a perturbed data set. Our technique is also capable of incorporating previously proposed noise addition techniques that maintain the statistical parameters of the data set, including correlations among attributes. Thus the perturbed data set may be used not only for classification but also for statistical analysis.

Keywords: *Data Mining, Security, Issues & Remedies, Privacy, Preservation, development, technology, information, process, etc.*

INTRODUCTION:

Data mining is often defined as the process of discovering meaningful, new correlation patterns and trends through non-trivial extraction of implicit, previously unknown information from large amount of data stored in repositories using pattern recognition as well as statistical and mathematical techniques.

A Structured Query Language (SQL) is usually stated or written to access a specific data while data miners might not even be exactly sure of what they need. So, the result of a SQL query is usually a part of the database; whereas the result of a data mining query is an analysis of full contents of the database.Data mining tasks can be classified as follows:

- Association rule mining or market basket analysis
- Classification and prediction
- Cluster analysis and outlier analysis

- Web Data mining and search engines'.
- Evolution analysis

The main focus of this thesis is to obtain secure Clustering results. Achieving accurate clustering results by providing privacy to sensitive data is trivial task. This thesis proposes two approaches for achieving the privacy for sensitive attributes during data mining [1].

Data Mining: Data mining also called as knowledge discovery in databases (KDD). It is defined as the process of evaluating interesting, useful and hidden patterns from large volumes of data stores and identifies the relationships among the patterns [2-4]. Data mining task requires utilities fir statistical data and Artificial Intelligence systems (AI). AI systems includes neural networks and machine learning sometimes one can combine them with database management system for evaluating or analyzing the huge volumes of digital data, which is the derived form of data sets..

Data mining has many applications; those have been listed in the above section. They can broadly categorized in to three area's one is business (insurance company, banking corporation, retail sector), second is science research (astronomy, medicine), and government security (detection of criminals and terrorists).

The large number of organizations, government and private data bases aims to ensure that the individual records are accurate and secure from unauthorized access. Data mining tasks are targeted toward extracting hidden predictive knowledge about a group rather than the individual.

Figure 1shows the Data mining process. First, data is collected from various sources in Data selection step. Next, Data will be pre-processed by dealing with null values and unformatted values. Then, Data will be transformed to proper format which is suitable for data mining operation [5]. Now, Knowledge will be extracted from data store which is nothing but data mining. Finally, evaluation of patterns for decision making takes place.

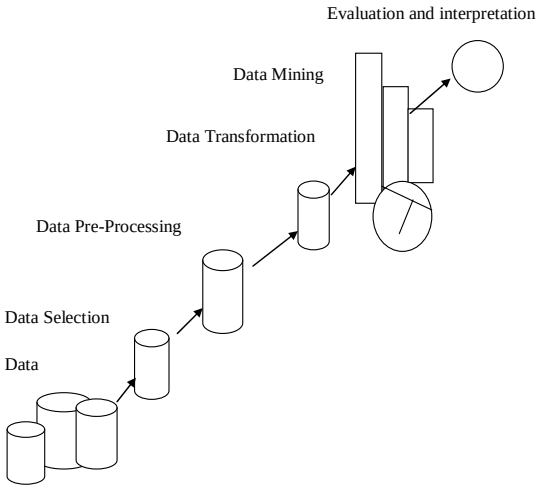

Figure 1.1 Data Mining Process

The specific goal of data mining process is to pull out the hidden information from a data set and change it into a good understandable structure for future use.

REVIEW OF LITERATURE:

Privacy-preserving distributed data mining is a multidisciplinary field and requires close cooperation between researchers and practitioners from the fields of cryptography, data mining, public policy and law. Now, the question is how to compute the results without pooling the data in a way that reveals nothing but the final results of the data mining computation [6]. This question of privacy-preserving data mining is actually a special case of a long-studied problem in cryptography called secure multiparty computation. This problem deals with a setting where a set of parties with private inputs wishes to jointly compute some function of their inputs. This joint computation should have the property that the parties learn the correct output and nothing else, even if some of the parties maliciously collude to obtain more information [7]. Clearly, a protocol is needed to solve privacy-preserving distributed data mining problems.

Database keeps growing rapidly because of the availability of powerful and affordable database systems. This explosive growth in data and databases has generated an urgent need for new techniques and tools that can intelligently and automatically transform the processed data into useful information and knowledge. Consequently, data mining has become a research area with increasing importance [8]. To design an effective data mining technique several issues to be taken into account such as types of data, efficiency and scalability of data mining algorithms, usefulness, different sources of data, protection of privacy & data security and so on.

Advantages of Data Mining: Data mining deals with extracting inherent, historical, and hypothetically critical information from huge databases. Data mining is a very challenging task since it involves building and providing software that will manage, explore, summarize, model, analyze and interpret large datasets in order to evaluate patterns and abnormalities. The methods or techniques of data mining are widely used at a higher rate in various forms of applications. Some of the important and critical applications are fraud prevention, detecting tax avoidance, catching drug smugglers, reducing customer churn and learning more about customers' behavior.

Misuses of Data Mining: There are also some (miss) uses of data mining that have little to do with any of these applications. For example, a number of news studys in early 2005 have reported results of analyzing associations between the political party that a person votes for and the car the person drives. The statistics of various branded cars used by the two key political parties of USA was analyzed. In the wake of 9/11 terrorist attacks, considerable use of personal information, provided by individuals for other purposes as well as information collected by governments including intercepted emails and telephone conversations, is being made in the belief that such information processing (including data mining) can assist in identifying persons who are likely to be involved in terrorist networks or individuals who might be in contact with such persons or other individuals involved in illegal activities (e.g. drug smuggling). Under legislation enacted since 9/11, many governments are able to demand access to most private sector data. This data can include records on travel, shopping, housing, utilities, credit, telecommunications and so on. Such data can then be mined in the belief that patterns can be found that will help in identifying terrorists or drug smugglers.

Data Mining for Healthcare: Data mining applications have an enormous potential and advantage in the healthcare industry. However, the quality and potential of data mining usage depends on the quality of data available in healthcare. So, keeping this in respect, the healthcare industry has the necessity to ensure quality data is captured, stored, managed, and placed. The benefit area is majorly standardization of clinical tasks and sharing the medical data among the medical organizations to enhance mining.

Data Mining for Market Analysis: Data Mining can also be used in market analysis. For instance, when a customer visits a store to buy certain products, then data mining helps us to identify the associated various items that the customer picks from the store. Identifying such data helps market analysis and to promote business. Such different customers and their buying patterns help identify the needs of the customers [9]. This technique helps improve the profits and to the customers to find their associate products better. So, Data mining

unveils the data, which is hidden in the database, but owners will not be happy if that hidden data is confidential, and they feel very uncomfortable if this data was submitted to the public. This Problem motivates and enhances the interest in doing the research to invent different types of algorithms and protocols for preserving privacy. These algorithms assure data owners that privacy is maintained while fulfilling accuracy and privacy.

Privacy Preserving Data Mining: Privacy preserving data mining aims to provide valid data mining results by not revealing the underlined sensitive information. Figure 2 shows the Architecture of privacy preserving data mining [10]. Data mining techniques extracts valuable information from data stores. When the techniques are applied, it not only extracts useful data, may also reveal sensitive information. So as to provide protection for sensitive information some privacy preserving techniques can be applied on original data then mining can be performed.

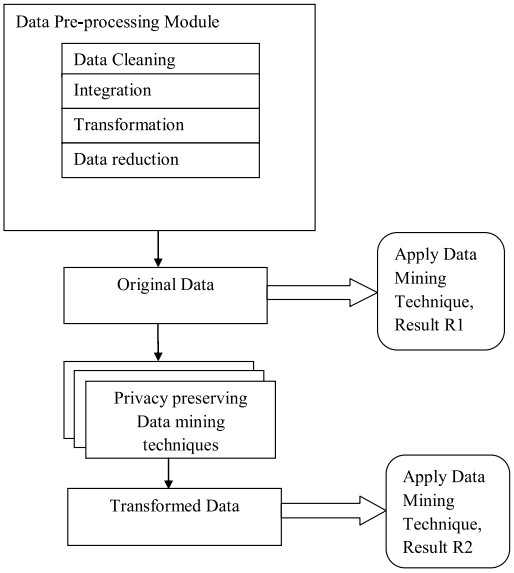

Figure: 2: Privacy Preserving Data mining in proposed approach

Privacy preserving data mining has much significance because of following reasons:

1. Data mining causes and ethical problem, because it reveals data, which should requires privacy?

2. Privacy preserving data mining provides security to private data against unauthorized access is a long term achievement for data mining security research community and for the government agencies.

3. Hence, the security issue is one of the emerging areas that became valuable research area in data mining.

Figure: 3 explore the sequence of steps to be followed for achieving secure data mining results. This work is proposed to perform clustering task on both original data called as R1 and on transformed data called as R2. Finally R1 and R2 will be observed and analyzed for evaluating the performance of proposed approach. This work provides one of the solutions for privacy preserving data mining (central data warehouse not on distributed databases) and the performance is measured in terms of accuracy of data mining result and privacy of sensitive data.

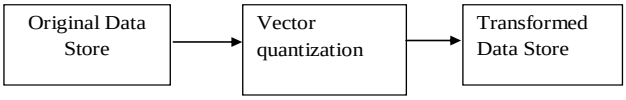

Figure: 3 Proposed Approach.

This thesis presents Vector Quantization technique with two approaches, which transforms underlined sensitive data with the help of codebook, in such a way that the patterns from the original data set are maintained more securely in transformed data set.

The experimental results are explored and analyzed; it is been observed and evaluated that the cluster objects can be extracted securely using the proposed approach. i.e., clusters obtained from the original data set and transformed data set are very similar in terms of accuracy with sufficient protection to sensitive data. It is been proposed to maintain a high quality of transformed data with privacy constraints.

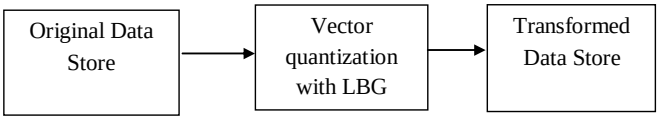

Figure: 4 Proposed Approaches 1.

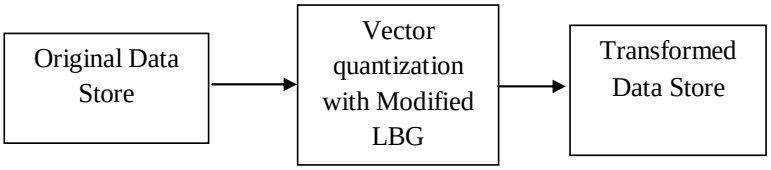

Figure: 5 Proposed Approaches 2.

Figure: 4 and Figure: 5 show the two proposed approaches. When a little amount of distortion is added to original data, one should take care about the accuracy of data mining tasks, such as classification, association rule mining and clustering. Adding the noise is not only the technique for preserving privacy, other methods like swapping, suppression, anonymization can be used.Data mining is a process to extract the information or knowledge automatically and intelligently from a huge amount of data, here in the process of data mining sensitive

information can be disclosed by compromising the individual's right to privacy. Increasing demand of Privacy preservation in data mining gives us direction to research about privacy preserving data mining.

Considering the rapid development in technology such as internet, data storage, data processing methods, we need to pay equal attention towards privacy preserving data mining. For secured public system we not only need to take care about the trimming of data but also the data inference.

Shu-Hsien Liao (2012) explained that increases in digital data have raised concerns about information privacy on a global basis. This particular research study is considered the seminal work in PPDM research. Their research laid the foundation for future research that addresses privacy issues within a data mining context. They explain that the Internet has made data collection and data storage much easier, but the potential for misuse has also risen significantly. Data mining results can show models of aggregate data, but the model's accuracy depends on the quality of data. The authors raise the concern that any changes to data affect the accuracy and output of data mining models. Their approach to this problem allows the consumer to provide a perturbed value for sensitive attributes. This allows consumers to participate in the process and hopefully gives the consumer a sense of control over his or her own information. A major drawback of this approach is that output accuracy is lost during data mining activities. However, the authors maintain that small drops in accuracy are an acceptable trade-off for privacy.

CONCLUSION:

This paper presented a privacy preserving technique that adds noise to each and every attribute, both numerical and categorical, of a data set. We added noise in such a way so that a high data quality is preserved in the perturbed data set. We measured data quality through the following quality indicators: degree of similarity between two decision trees obtained from an original and a perturbed data set, prediction accuracy of the decision trees, and correlation matrices of the original and the perturbed data set. Therefore, the perturbed data set can be used for classification, prediction and correlation analyzes. More- over, since we add a little amount of noise the perturbed data set can also be used for many other data analyzes. Since noise is added to all attributes, it makes record re-identification determining the confidential class values difficult. The presented techniques for adding noise to a sensitive class attribute. We added the same amount of noise in three class attribute perturbation techniques, namely the RPT, PPT and ALPT. We compared results of our experiments on all these techniques.

Our experimental results suggest that the RPT and PPT preserve the patterns better than the ALPT - although the same amount of noise has been added in the techniques.

REFERENCES:

S. R. M. Oliveria, (2005) Data Transformation for Privacy-Preserving Data Mining, Ph. D. thesis, University of Alberta.

Charu C, Aggarwal, Philip S and Yu, (2008) A General Survey of Privacy- Preserving Data Mining Models and Algorithms, Springer.

Shu-Hsien Liao, Pei-Hui Chu and Pei-Yuan Hsiao, (2012) "Data mining techniques and applications–A decade review from 2000 to 2011", Expert Systems with Applications.

Riccardo Bellazzi and Blaz Zupan, (2008) "Predictive data mining in clinical medicine: Current issues and guidelines", International Journal of Medical Informatics, pp. 81–97.

Wei-Yin Loh (2011) "Classification and regression trees", WIREs Data Mining and Knowledge Discovery, Volume 1.

Cristóbal Romero, Sebastián Ventura and Enrique García, (2007) Data mining in course management systems: Moodle case study and Tutorial, Computers & Education.

Yi Peng, Yong Zhang, Yu Tang and Shiming Li, (2011) An incident information management framework based on data integration, data mining, and multi- criteria decision making, Decision Support Systems, pp.316–327.

Tak-chung Fu, (2011) "A Review on Time Series Data Mining", Engineering applications of Artificial Intelligence, pp.164-181.

MikalaiTsytsarau and Themis Palpanas, (2012) "Survey on mining subjective data on the web", Data Mining Knowledge Discovery, pp.478–514.

injun Qi and MingkuiZong, (2011) "An Overview of Privacy Preserving Data Mining", International Conference of Environment Sienes and engineering.