A Survey Paper on Data Restoration and Privacy Preserving of Data Using C4.5 Algorithm

Miss. Sharmila A. Harale
Dr. D. Y. Patil College of Engineering, Ambi, Talegaon.
University of Pune, Pune, India.

Abstract

In recent years, advances in hardware technology have led to an increase in the capability to store and record personal data about consumers and individuals. This has led to concerns that the personal data may be misused for a variety of purposes. To improve these concerns, a number of techniques have recently been proposed in order to perform the data mining tasks in a privacy-preserving way. Here, we introduce a privacy preserving approach that can be applied to decision tree learning, without associated loss of accuracy with the help of decision tree algorithm c4.5 as it provides more reliable and accurate results over previously implemented ID3 decision tree algorithm and data mining methods with mixed discretely and continuously valued attributes. It describes an approach to the preservation of the privacy of collected data samples in cases where information from the sample database has been partially lost. This approach converts the original sample data sets into a group of unreal data sets, from which the original samples cannot be reconstructed without the entire group of unreal data sets. Meanwhile, an accurate decision tree can be built directly from those unreal data sets. This approach can be applied directly to the data storage as soon as the first sample is collected.

Keywords—Data mining, Privacy preserving Data Mining (PPDM), Decision Tree, Decision Tree Learning, ID3 algorithm, C4.5 algorithm, discretely valued attributes, continuously valued attributes.

1. Introduction

1.1 What is Data Mining?

Data mining is the process of discovering interesting knowledge, such as patterns, associations, changes, and anomalies and significant structures, from large amounts of data stored in databases, data warehouses, or other information repositories. Due to the wide availability of huge amounts of data in electronic forms, and the imminent need for turning such data into useful information and knowledge for broad applications including market analysis, business management, and decision support, data mining has attracted a great deal of attention in information industry in recent years. Data mining has been popularly treated as a synonym of knowledge discovery in databases, although some researchers view data mining as an essential step of knowledge discovery. In general, a knowledge discovery process consists of an iterative sequence of the following steps:

Data Cleaning, this handles noisy, erroneous, missing, or irrelevant data,
Data Integration, where multiple, heterogeneous data sources may be integrated into one,
Data Selection, where data relevant to the analysis task are retrieved from the database,
Data Transformation, where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations,
Data Mining, which is an essential process where intelligent methods are applied in order to extract data patterns,
Pattern Evaluation, which is to identify the truly interesting patterns representing knowledge based on some interestingness measures, and
Knowledge Presentation, where visualization and knowledge representation techniques are used to present the mined knowledge to the user.

With the widely available relational database systems and data warehouses, the first four processes: data cleaning, data integration, data selection, and data transformation, can be performed by constructing data warehouses and performing some OLAP operations on

the constructed data warehouses. The data mining, pattern evaluation and knowledge presentation processes are sometimes integrated into one process, referred as data mining.

1.2 Privacy Preserving Data Mining

Data mining is a recently emerging field. Data mining is the process of extracting knowledge or pattern from large amount of data. It is widely used by researchers for science and business process. Data collected from information providers are important for pattern reorganization and decision making. The data collection process takes time and efforts hence sample datasets are sometime stored for reuse. However attacks are attempted to steal these sample datasets and private information may be leaked from these stolen datasets. Therefore privacy preserving data mining are developed to convert sensitive datasets into sanitized version in which private or sensitive information is hidden from unauthorized retrievers.

Privacy preserving data mining refers to the area of data mining that aims to protect sensitive information from illegal or unwanted disclosure. Privacy Preservation Data Mining was introduced to preserve the privacy during mining process to enable conventional data mining technique. Many privacy preservation approaches were developed to protect private information of sample dataset.

Contemporary research in privacy preserving data mining mainly falls into one of two categories: 1) perturbation and randomization-based approaches, and 2) secure multiparty computation (SMC)-based approaches. SMC approaches employ cryptographic tools for collaborative data mining computation by multiple parties. Samples are distributed among different parties and they take part in the information computation and communication process. SMC research focuses on protocol development for protecting privacy among the involved parties or computation efficiency; however, centralized processing of samples and storage privacy is out of the scope of SMC.

We introduce a new perturbation and randomization based approach that protects centralized sample data sets utilized for decision tree data mining. Privacy preservation is applied to sanitize the samples prior to their release to third parties in order to mitigate the threat of their inadvertent disclosure or theft. In contrast to other sanitization methods, our approach does not affect the accuracy of data mining results. The decision tree can be built directly from the sanitized data sets, such that the originals do not need to be reconstructed. Moreover, this approach can be applied at any time during the data collection process so that privacy protection can be in effect even while samples are still being collected.

1.3 Decision Tree Learning

A decision tree is a tree in which each branch node represents a choice between a number of alternatives, and each leaf node represents a decision. Decision tree are commonly used for gaining information for the purpose of decision -making. Decision tree starts with a root node on which it is for users to take actions. From this node, users split each node recursively according to decision tree learning algorithm. The final result is a decision tree in which each branch represents a possible scenario of decision and its outcome.

Decision trees classify instances by traverse from root node to leaf node. We start from root node of decision tree, testing the attribute specified by this node, and then moving down the tree branch according to the attribute value in the given set. This process is the repeated at the sub-tree level.

Decision tree learning is a method for approximating discrete valued target functions, in which the learned function is represented by a decision tree. Learned trees can also be represented as sets of if-then rules to improve human readability.

2. Literature Survey

a) P. K. Fong and J. H. Weber-Jahnke [1], in this paper a new method that we maintaining privacy preservation of data sing unrealized datasets. This paper introduces a privacy preserving approach that can be applied to decision tree learning, without concomitant loss of accuracy. It describes an approach to the preservation of the privacy of collected data samples in cases where information from the sample database has been partially lost. This approach converts the original sample data sets into a group of unreal data sets, from which the original samples cannot be reconstructed without the entire group of unreal data sets. Meanwhile, an accurate decision tree can be built directly from those unreal data sets. This novel approach can be applied directly to the data storage as soon as the first sample is collected. The approach is compatible with other privacy preserving approaches, such as cryptography, for extra protection.

b) J. Dowd et al. [2] proposed the several contributions towards privacy-preserving decision tree mining. The most important is that the framework introduced a new data perturbation technique based on random substitutions. This perturbation technique is similar to the randomization techniques used in the context of

statistical disclosure control but is based on a different privacy measure called ρ1-to-ρ2 privacy breaching and a special type of perturbation matrix called the γ-diagonal matrix.

c) In Privacy Preserving Data Mining: Models and Algorithms [3], Aggarwal and Yu classify privacy preserving data mining techniques, including data modification and cryptographic, statistical, query auditing and perturbation-based strategies. Statistical, query auditing and most cryptographic techniques are subjects beyond the focus of this paper. In section, we explore the privacy preservation techniques for storage privacy attacks. Data modification techniques maintain privacy by modifying attribute values of the sample data sets. Essentially, data sets are modified by eliminating or unifying uncommon elements among all data sets. These similar data sets act as masks for the others within the group because they cannot be distinguished from the others; every data set is loosely linked with a certain number of information providers. K-anonymity [7] is a data modification approach that aims to protect private information of the samples by generalizing attributes. K-anonymity trades privacy for utility. Further, this approach can be applied only after the entire data collection process has been completed.

d) L. Liu [3] proposed a new method that we build data mining models directly from the perturbed data without trying to solve the general data distribution reconstruction as an intermediate step. More precisely, proposed a modified C4.5 decision tree classifier that can deal with perturbed numeric continuous attributes. Privacy preserving decision tree C4.5 (PPDTC4.5) classifier uses perturbed training data, and builds a decision tree model, which could be used to classify the original or perturbed data sets. The experiments have shown that PPDTC4.5 classifier can obtain a high degree of accuracy when used to classify the original data set.

3. Existing System

In Previous work in privacy-preserving data mining has addressed two issues. In one, the aim is to preserve customer privacy by disturbing the data values. In this scheme random noise data is introduced to distort sensitive values, and the distribution of the random data is used to generate a new data distribution which is close to the original data distribution without revealing the original data values. The estimated original data distribution is used to reconstruct the data, and data mining techniques, such as classifiers and association rules are applied to the reconstructed data set.

The other approach uses cryptographic tools to build data mining models. The goal is to securely build an ID3 decision tree where the training set is distributed between two parties. Different solutions were given to address different data mining problems using cryptographic techniques.

3.1 Disadvantage of Existing System

Existing system covers the application of new privacy preserving approach with the ID3 decision tree learning algorithm and for discrete-valued attributes only.

4. Proposed Solution

In the proposed solution, privacy preserving decision tree learning using unrealized data sets technique is used. We will implement this technique using C4.5 decision tree learning algorithm and data mining methods with mixed discretely and continuously valued attributes and thus data restoration and preservation of privacy of data is done.

The following assumptions are made for the scope of this paper: first, as is the norm in data collection processes, a sufficiently large number of sample data sets have been collected to achieve significant data mining results covering the whole research target. Second, the number of data sets leaked to potential attackers constitutes a small portion of the entire sample database. Third, identity attributes (e.g., social insurance number) are not considered for the data mining process because such attributes are not meaningful for decision making. Fourth, all data collected are discretized; continuous values can be represented via ranged value attributes for decision tree data mining [1].

4.1 System Architecture

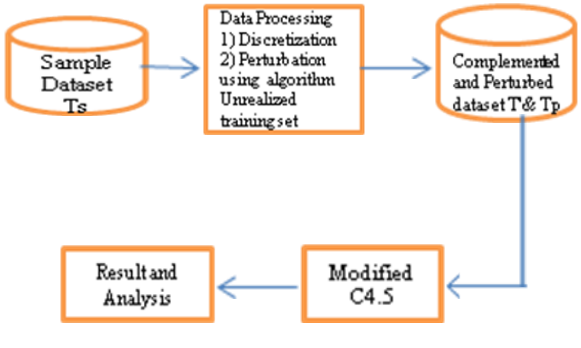

Figure 1. System Architecture

The system architecture of proposed system is shown in above figure. It consist of two main module first is data

preprocessing and second is decision tree generation. In data preprocessing module initially continuous value attribute dataset is converted into discrete value after that the dataset is converted into sanitized version by using algorithm unrealized training set. Then generated complemented dataset and perturbed dataset is given as an input to decision tree generation module in which decision tree is built by using C4.5 and result generated by algorithm is compared to analyze the algorithm.

5. System Requirement and Specifications

5.1 Software Requirements

Front End: Java (JDK 1.5)
Tools Used: Net Beans
Operating System: Windows 7
Database: MySQL Server

5.2 Hardware Requirements

Processor: Pentium IV– 500 MHz to 3.0 GHz
RAM: 1GB
Hard Disk: 20 GB.
Monitor: Any Color Display.
Keyboard: Standard Windows Keyboard
Mouse: 3 Button

6. Conclusion

Security of data and files are the most attentive area in the field of IT industry and for individual personnel. The World need a good quality of tools that providing security to prevent unauthorized access of network or system, but today's modern technologies provide attack launching tools globally without much more hurdle. Our objective is to develop an overall framework for defending attacks and threats to computer systems. Data generated from network traffic monitoring tends to have been very high dimensionality, volume and heterogeneity. Devising the performance of serial data mining algorithms insufferable for online analysis.
In this paper as we have seen in literature survey, the performance parameters of the two classification algorithms ID3 and C4.5, and the results was satisfactory of C4.5 [12]. Experimental evaluation on real world data shows that C4.5 can learn to identify users simply by what commands they use and how often, and such an identification can be used to detect intrusions in a network computer system. So we are using C4.5 algorithm for implementation of our proposed model to achieve effective data restoration and privacy preservation of data.

7. References

[1] Pui K. Fong And Jens H. Weber-Jahnke, "Privacy Preserving Decision Tree Learning Using Unrealized Data Sets" Proc. IEEE Transactions On Knowledge And Data Engineering, Vol. 24, No. 2, February 2012.
[2] J. Dowd, S. Xu, and W. Zhang, "Privacy-Preserving Decision Tree Mining Based on Random Substitions," Proc. Int'l Conf. Emerging Trends in Information and Comm. Security (ETRICS '06), pp. 145-159, 2006.
[3] C. Aggarwal and P. Yu, Privacy-Preserving Data Mining:, Models and Algorithms. Springer, 2008.
[4] L. Liu, M. Kantarcioglu, and B. Thuraisingham, "Privacy Preserving Decision Tree Mining from Perturbed Data," Proc. 42nd Hawaii Int'l Conf. System Sciences (HICSS '09), 2009.
[5] Y. Lindell and B. Pinkas "Privacy preserving data mining" In Advances in Cryptology, volume 1880 of Lecture Notes in Computer Science, pages 36–53. Springer-Verlag, 2000.
[6] P.K. Fong, "Privacy Preservation for Training Data Sets in Database: Application to Decision Tree Learning," master's thesis, Dept. of Computer Science, Univ. of Victoria, 2008.
[7] L. Sweeney, "k-Anonymity: A Model for Protecting Privacy," Int'l J. Uncertainty, Fuzziness and Knowledge-based Systems, vol. 10, pp. 557-570, May 2002.
[8] S. Ajmani, R. Morris, and B. Liskov, "A Trusted Third-Party Computation Service," Technical Report MIT-LCS-TR-847, MIT, 2001.
[9] S.L. Wang and A. Jafari, "Hiding Sensitive Predictive Association Rules," Proc. IEEE Int'l Conf. Systems, Man and Cybernetics, pp. 164- 169, 2005.
[10] R. Agrawal and R. Srikant, "Privacy Preserving Data Mining," Proc. ACM SIGMOD Conf. Management of Data (SIGMOD '00), pp. 439-450, May 2000.
[11] Q. Ma and P. Deng, "Secure Multi-Party Protocols for Privacy Preserving Data Mining," Proc. Third Int'l Conf. Wireless Algorithms, Systems, and Applications (WASA '08), pp. 526-537, 2008.
[12] Surbhi Hardikar, Ankur Shrivastava, Vijay Choudhary,"Comparison Between ID3 And C4.5 In Contrast To IDS", Proc. VSRD-IJCSIT, Vol. 2 (7), 2012, 659-667.