ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 8, August 2015

Privacy Preserving using Association Rule in Data Mining Techniques

Mrs. S. Vasanthi[1], Ms. S. Nandhini[2]

Assistant Professor, Dept. of Computer Science., P.S.G College of Arts and Science, India[1]

Research Scholar, Dept. of Computer Science, P.S.G College of Arts and Science, India[2]

Abstract: This paper describes the problem of Privacy Preserving Data Mining (PPDM). It describes some of the common cryptographic tools and constructs used in several PPDM techniques. The paper describes an overview of some of the well known PPDM algorithms, - ID3 for decision tree, association rule mining, EM clustering, frequency mining and Naïve Bayes. Most of these algorithms are usually a modification of a well known data mining algorithm along with some privacy preserving techniques. The paper finally describes the problem of using a model without knowing the model rules on context of passenger classification at the airlines security checkpoint by homeland security. This paper is intended to be a summary and a high level overview of PPDM.

Keywords: Anonymization, Data Mining, Sensitive Information, Privacy preserving data Mining, Provenance.

I. INTRODUCTION

Data mining refers to the techniques of extracting rules and patterns from data. It is also commonly known as KDD (Knowledge Discovery from Data). Traditional data mining operates on the data warehouse model of gathering all data into a central site and then running an algorithm against that warehouse. This model works well when the entire data is owned by a single custodian who generates and uses a data mining model without disclosing the results to any third party. Privacy preserving data mining (PPDM) is a novel research direction in Data Mining (DM), where DM Algorithms are analysed for the side-effects they incur in data privacy. The main objective of PPDM is to develop.

Algorithms for modifying the original data in some way, so that the private data and private knowledge remain Private even after the mining process. In DM, the users are provided with the data and not the association rules and are free to use their own tools; so, the restriction for privacy has to be applied on the data itself before the mining phase. For this reason, we need to develop mechanisms that can lead to new privacy control systems to convert a given database into a new one in such a way to preserve the general rules mined from the original database. The procedure of transforming the source database into a new database that hides some sensitive patterns or rules is called the sanitization process.

II. ASSOCIATION RULE MINING

Let I = {i1… in} be a set of items. Let D be a database which contains set of transactions. Each transaction t _ D is an item set such that t is a proper subset of I. As transaction t supports X, a set of items in I,
if X is a proper subset of t. Assume that the items in a transaction or an item set are sorted in lexicographic order. An association rule is an implication of the form X_Y, where X and Y are subsets of I and X_Y= Ø. The support of rule X_Y can be calculated by the following equation: Support(X_Y) = |X_Y| / |D|, where |X_Y| denotes the number of transactions containing the item set XY in the database, |D| denotes the number of the transactions in the database D. The confidence of rule is computed by Confidence (X_Y) = |X_Y|/|X|, where |X| is number of transactions in database D that contains item set X. A rule X_Y is strong if support (X_Y) _ min_support and confidence (X_Y) _ min_confidence, where min_support and min_confidence are two given minimum thresholds.

A. User Role-based Methodology:

Current models and algorithms proposed for PPDM mainly focus on how to hide that sensitive information from certain mining operations. However, as depicted in Fig. 1a, the whole KDD processes involve multi-phase operations. In this paper, we investigate the privacy aspects of data mining by considering the whole knowledge-discovery process.
We present an overview of the many approaches which can help to make proper use of sensitive data and protect the security of sensitive information discovered by data mining. We use the term "sensitive information" to refer to privileged or proprietary information that only certain people are allowed to see and that is therefore not accessible to everyone. If sensitive information is lost or used in any way other than intended, the result can be severe damage to the person or organization to which that information belongs.

The term "sensitive data" refers to data from which sensitive information can be extracted. Throughout the paper, we consider the two terms "privacy" and "sensitive information" are interchangeable. In this paper, we develop a user-role based methodology to conduct the review of related studies.

DOI 10.17148/IJARCCE.2015.48106

III. ASSOCIATION RULE HIDING ALGORITHMS

Association rule hiding algorithms can be divided into three distinct approaches. They are heuristic approaches, border-revision approaches and exact approaches.

A. Heuristic Approach:

Heuristic approaches can be further categorized into distortion based schemes and blocking based schemes. To hide sensitive item sets, distortion based scheme changes certain items in selected transactions from 1's to 0's and vice versa. Blocking based scheme replaces certain items in selected transactions with unknowns. These approaches have been getting focus of attention for majority of the researchers due to their efficiency, scalability and quick responses.

B. Border Revision Approach:

Border revision approach modifies borders in the lattice of the frequent and infrequent item sets to hide sensitive association rules. This approach tracks the border of the non sensitive frequent item sets and greedily applies data modification that may have minimal impact on the quality to accommodate the hiding sensitive rules. Researchers proposed many border revision approach algorithms such as BBA (Border Based Approach), Max– Min1 and Max-Min2 to hide sensitive association rules. The algorithms uses different techniques such as deleting specific sensitive items and also attempt to minimize the number of non sensitive item sets that may be lost while sanitization is performed over the original database in order to protect sensitive rules.

C. Exact Approach:

Third class of approach is non heuristic algorithm called exact, which conceive hiding process as constraint satisfaction problem. These problems are solved by integer programming. This approach can be concerned as descendant of border based methodology.

D. Association Rule Mining

We describe the privacy preserving association rule mining technique for a horizontally partitioned data set across multiple sites. Let I = be a set of items and T = be a set of transactions where each.

A transaction contains an item set only if . An association rule implication is of the form ( ) with support s and confidence c if s% of the transactions in T contains and c% of transactions that contain X also contain.

IV. PROPOSED ALGORITHM

In order to hide an association rule, X → Y, we can either decrease its support or its confidence to be smaller than user-specified minimum support transaction (MST) and minimum confidence transaction (MCT).

To decrease the confidence of a rule, we can either (1) increase the support o of X, the left hand side of the rule, but not support of X → Y, or (2) decrease the support of the item set X →Y .For the second case, if we only decrease the support of Y, the right hand side of the rule, it would reduce the confidence faster than simply reducing the support of X → Y.

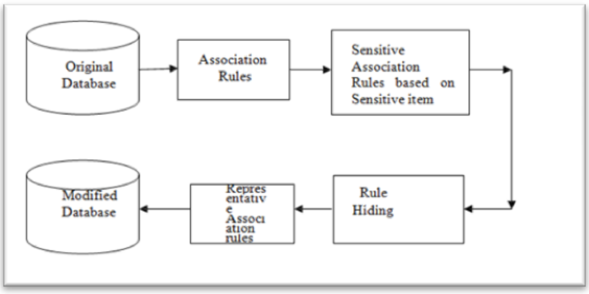
Original Database
Association Rules
Sensitive Association Rules based on Sensitive item
Rule Hiding
Repres entativ e Associ ation rules
Modified Database

Figure 1: Association Rule Hiding Framework

To decrease support of an item, we will modify one item at a time by changing from 1 to 0 or from 0 to 1 in a selected transaction. Based on these two concepts, we propose a new association rule hiding algorithm for hiding sensitive items in association rules. In our algorithm, a rule X → Y is hidden by decreasing the support value of X →Y and increasing the support value of X. That can increase and decrease the support of the LHS and RHS item of the rule correspondingly. This algorithm first tries to hide the rules in which item to be hidden i.e., X is in right hand side and then tries to hide the rules in which X is in left hand side. For this algorithm t is a transaction, T is a set of transactions, R is used for rule, RHS (R) is Right Hand Side of rule R, LHS (R) is the left hand side of the rule R, Confidence (R) is the confidence of the rule R, a set of items H to be hidden.

V. CONCLUSION AND FUTURE WORK

This paper, the database privacy problems are addressed and a new technique for privacy preservation is proposed. Association rule hiding techniques are used to hide sensitive association rules. A new heuristic method to hide the sensitive association rules is proposed. Data distortion technique is applied so that sensitive information cannot be discovered through data mining techniques. Confidence of the rules is represented as representative rules. Confidence of the rule is recomputed and compared with threshold level. The confidence of the sensitive rules might be reduced while maintaining the support. From the experimental results, it is observed that all the rules containing sensitive items are hidden. The algorithm is implemented and numerical example is shown. Further research is in progress to evolve a method which can avoid the computational overhead associated with confidence of the rules.

REFERENCES

[1] Alberto Trombetta and Wei Jiang (2011), 'Privacy-Preserving Updates to Anonymous and Confidential Databases', IEEE Transactions on Knowledge and Data Engineering, Vol. 22, pp. 578-568.

[2] Gayatri Nayak and Swagatika Devi (2011), 'A Survey On Privacy Preserving Data Mining: Approaches And Techniques', International Journal of Engineering Science and Technology, pp.2127-2133

[3] Guang Li and Yadong Wang (2011), 'Privacy-Preserving Data Mining Based on Sample Selection and Singular Value Decomposition', Proceedings of the IEEE International Conference on Internet Computing and Information Services , pp.298-301

[4]Jain Y.K. (2011), 'An Efficient Association Rule Hiding Algorithm for Privacy Preserving Data Mining', International Journal of Computer Science and Engineering, pp.96-104

[5]Tools for Privacy Preserving Data Mining, Chris Clifton, Murat Kantarcioglu and Jaideep Vaidya, Purdue University.

[6]. Privacy Presercing Classification of Customer Data without Loss of Accuracy, Zhiqiang Yang, Sheng Zhong, Rebecca N. Wright.

DOI 10.17148/IJARCCE.2015.48106