ACSIJ Advances in Computer Science: an International Journal, Vol. 4, Issue 4, No.16 , July 2015
ISSN : 2322-5157
www.ACSIJ.org

ACSIJ
WWW.ACSIJ.ORG

A Survey on the Privacy Preserving Algorithm and techniques of Association Rule Mining

Maryam Fouladfar[1], Mohammad Naderi Dehkordi[2]

Faculty of Computer Engineering, Najafabad Branch, Islamic Azad University, Najafabad, Isfahan, Iran

[1]*maryamfouladfar@sco.iaun.ac.ir*, [2]*naderi@iaun.ac.ir*

Abstract

In recent years, data mining is a popular analysis tool to extract knowledge from collection of large amount of data. One of the great challenges of data mining is finding hidden patterns without revealing sensitive information. Privacy preservation data mining (PPDM) is answer to such challenges. It is a major research area for protecting sensitive data or knowledge while data mining techniques can still be applied efficiently. Association rule hiding is one of the techniques of PPDM to protect the association rules generated by association rule mining. In this paper, we provide a survey of association rule hiding methods for privacy preservation. Various algorithms have been designed for it in recent years. In this paper, we summarize them and survey current existing techniques for association rule hiding.

Keywords: Association Rule Hiding, Data Mining, Privacy Preservation Data Mining.

1. Motivation

computers have promised us a fountain of wisdom but delivered a deluge of information. this huge amount of data makes it crucial to develop tools to discover what is called hidden knowledge. these tools are called data mining tools. so, data mining promises to discover what is hidden, but what if that hidden knowledge is sensitive and owners would not be happy if this knowledge were exposed to the public or to adversaries? this problem motivates for write this paper.

2. Introduction

The problem of privacy preserving data mining has become more important in recent years because of the increasing ability to store personal data about users and the increasing sophistication of data mining algorithm to leverage this information. A number of data mining techniques have been suggested in recent years in order to perform privacy preserving Data mining techniques have been developed successfully to extracts knowledge in order to support a variety of domains marketing, weather forecasting, medical diagnosis, and national security. But it is still a challenge to mine certain kinds of data without violating the data owners 'privacy .For example, how to mine patients private data is an ongoing problem in health care applications .As data mining become more pervasive, privacy concerns are increasing. Commercial concerns are also concerned with the privacy issue. Most organizations collect information about individuals for their own specific needs. Very frequently, however, different units within an organization themselves may find it necessary to share information. In such cases, each organization or unit must be sure that the privacy of the individual is not violated or that sensitive business information is not revealed .Consider, for example, a government, or more appropriately, one of its security branches interested in developing a system for determining, from passengers whose baggage has been checked, those who must be subjected to additional security measures. The data indicating the necessity for further examination derives from a wide variety of sources such as police records; airports; banks; general government statistics; and passenger information records that generally include personal information; demographic data; flight information; and expenditure data. In most countries, this information is regarded as private and to avoid intentionally or unintentionally exposing confidential information about an individual, it is against the law to make such information freely available. While various means of preserving individual information have been developed, there are ways for circumventing these methods. In our example, in order to preserve privacy, passenger information records can be de-identified before the records are shared with anyone who is not permitted directly to access the relevant data. This can be accomplished by deleting from the dataset unique identity fields. However, even if this information is

Copyright (c) 2015 Advances in Computer Science: an International Journal. All Rights Reserved.

deleted, there are still other kinds of information, personal or behavioral that, when linked with other available datasets, could potentially identify subjects. To avoid these types of violations, we need various data mining algorithm for privacy preserving. We review recent work on these topics. In this paper, it has been tried to focus on data -mining background in advance, while the important part of the paper has been focusing on introduction of different approaches of data-mining and algorithms of data mining privacy preserving for sanitizing sensitive knowledge in context of mining association rules or item sets with brief descriptions. It has been tried to concentrate on different classifications of data mining privacy preserving approaches.

3. Privacy Preserving Data Mining Concepts

Today as the usage of data mining technology has been increasing, the importance of securing information against disclosure of unauthorized access is one of the most important issues in securing of privacy of data mining [1]. The state or condition of being isolated from the view or presence of others is privacy [2] which is associated with data mining so that we are ab le to conceal sensitive information from revelation to public [1]. Therefore to protect the sensitive rule from unauthorized publishing, privacy preserving data mining (PPDM) has focused on data mining and database security field [3].

3.1 Association Rule Mining Strategy

Association rules are an important class of regularities within data which have been extensively studied by the data mining community. The problem of mining association rules can be stated as follows: Given I = $\{i_1 , i_2 , \dots , i_m \}$ is a set of items, T = $\{t_1, t_2 , \dots , t_n\}$ is a set of transactions, each of which contains items of the itemset I . Each transaction ti is a set of items such that $t_i \subseteq I$. An association rule is an implication of the form: $X \rightarrow Y$, where $X \subset I$, $Y \subset I$ and $X \cap Y = \emptyset$. X (or Y) is a set of items, called itemset. In therule $X \rightarrow Y$, X is called the antecedent, Y is the consequent. It is obvious that the value of the antecedent implies the value of the consequent. The antecedent, also called the “left handside” of a rule, can consist either of a single item or of a whole set of items. This applies for the consequent, also called the “right hand side”, as well. Often, a compromise has to be made between discovering all itemsets and computation time. Generally, only those item sets that fulfill a certain support requirement are taken into consideration. Support and confidence are the two most important quality measures for evaluating the interestingness of a rule. The support of the rule $X \rightarrow Y$ is the percentage of transactions in T that contain $X \cap Y$. It determines how frequent the rule is applicable to the transaction set T . The support of a rule is represented by the formula **(1)**:

$$Support(X,Y) = \frac{|X \cup Y|}{|D|} \quad (1)$$

where $|X \cap Y|$ is the number of transactions that contain all the items of the rule and n is the total number of transactions. The confidence of a rule describes the percentage of transactions containing X which also contain Y . It is given by **(2)**:

$$Confidance(X \rightarrow Y) = \frac{|X \cup Y|}{|X|} \quad (2)$$

Confidence is a very important measure to determine whether a rule is interesting or not. The process of mining association rules consists of twomain steps. The first step is, identifying all the itemsets contained in the data that are adequate for mining association rules. These combinations have to show at least a certain frequency and are thus called frequent itemsets. The second step generates rules out of the discovered frequent itemsets. All rules that has confidence greater than minimum confidence are regarded as interesting.

3.2 Side Effects

As it is presented in (Fig. 1), R is denoted as all association rules in the database D, as well as SR for the sensitive rules, the none sensitive rules ~SR, discovered rules R’ in sanitized database D’. The circles with the numbers of 1, 2, and 3 are possible problems that respectively represent the sensitive association rules that were failed to be censored, the legitimate rules accidentally missed, and the artificial association rules created by the sanitization process.

Copyright (c) 2015 Advances in Computer Science: an International Journal. All Rights Reserved.

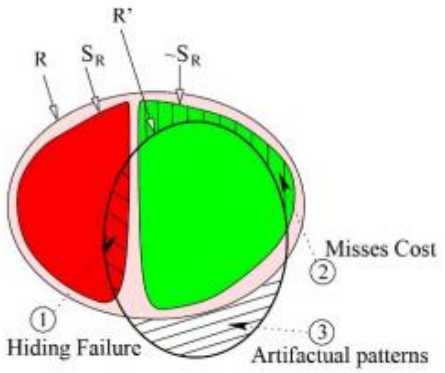

Fig. 1 Side Effects

The percentage of sensitive information that is still discovered, after the data has been sanitized, gives an estimate of the hiding failure parameter. Most of the developed privacy preserving algorithms are designed with the goal of obtaining zero hiding failure. Thus, they hide all the patterns considered sensitive. However, it is well known that the more sensitive information we hide, the more non-sensitive information we miss. Thus, some PPDM algorithms have been recently developed which allow one to choose the amount of sensitive data that should be hidden in order to find a balance between privacy and knowledge discovery. For example, in [4], Oliveira and Zaiane define the hiding failure (HF) as the percentage of restrictive patterns that are discovered from the sanitized database. It is measured as (3):

$$HF = \frac{\#SR(D')}{\#SR(D)} \quad (3)$$

where #RP (D) and #RP(D′) denote the number of restrictive patterns discovered from the original data base D and the sanitized database D′ respectively. Ideally, HF should be 0. In their framework, they give a specification of a disclosure threshold φ , representing the percentage of sensitive transactions that are not sanitized, which allows one to find a balance between the hiding failure and the number of misses. Note that φ does not control the hiding failure directly, but indirectly by controlling the proportion of sensitive transactions to be sanitized for each restrictive pattern.

When quantifying information loss in the context of the other data usages, it is useful to distinguish between: lost information representing the percentage of non-sensitive patterns (i.e., association, classification rules) which are hidden as side-effect of the hiding process; and the artifactual information representing the percentage of artifactual patterns created by the adopted privacy preserving technique. For example, in [4], Oliveira and Zaiane define two metrics misses cost and artifactual pattern which are corresponding to lost information and artifactual information respectively. In particular, misses cost measures the percentage of nonrestrictive patterns that are hidden after the sanitization process. This happens when some non-restrictive patterns lose support in the database due to the sanitization process. The misses cost (MC) is computed as (4):

$$MC = \frac{\#{\sim}SR(D) - \#{\sim}SR(D')}{\#SR(D)} \quad (4)$$

where # ∼ RP (D) and # ∼ RP(D′) denote the number of non-restrictive patterns discovered from the original database D and the sanitized database D′ respectively. In the best case, MC should be 0%. Notice that there is a compromise between the misses cost and the hiding failure in their approach. The more restrictive patterns they hide, the more legitimate patterns they miss. The other metric, artifactual pattern (AP), is measured in terms of the percentage of the discovered patterns that are artifacts. The formula is (5):

$$AP = \frac{|R'||R \cap R'|}{|R'|} \quad (5)$$

where |X | denotes the cardinality of X . According to their experiments, their approach does not have any artifactual patterns, i.e., AP is always 0. In case of association rules, the lost information can be modeled as the set of non-sensitive rules that are accidentally hidden, referred to as lost rules, by the privacy preservation technique, the artifactual information, instead, represents the set of new rules, also known as ghost rules, that can be extracted from the database after the application of a sanitization technique.

4. Different Approaches Sin PPDM

Many approaches have been proposed in PPDM in order to censor sensitive knowledge or sensitive association rules [5,6]. Two classifications in existing sanitizing algorithm of PPDM shown in (fig. 2).

Copyright (c) 2015 Advances in Computer Science: an International Journal. All Rights Reserved.

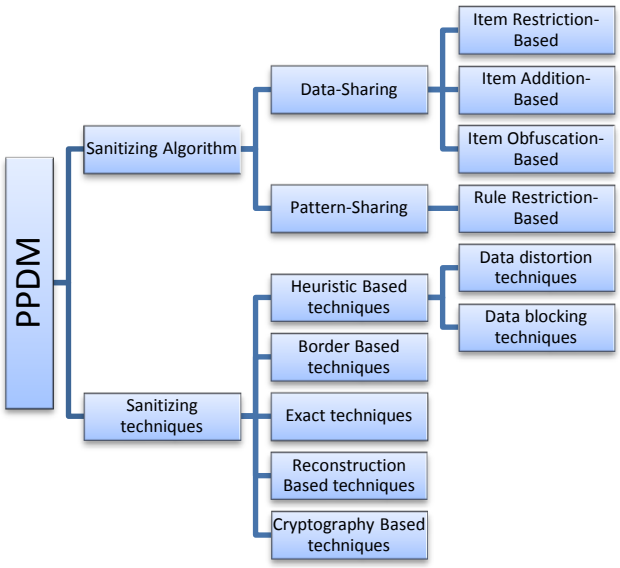

Fig. 2 Classification of Approaches

Sanitizing Alghorithm

data-sharing: In data-sharing technique, without analyzing or any statistical techniques, data will be communicated between parties. In this approach, the algorithms suppose change database by producing distorted data in the data base [6,7,8].

pattern-sharing: In pattern-sharing technique, the algorithm tries to sanitize the rules which are mined from the data set [6,8,9].

Sanitizing techniques

Heuristic-Based: Heuristic-based techniques resolves how to select the appropriate data sets for data modification. Since the optimal selective data modification or sanitization is an NP-Hard problem, heuristics is used to address the complexity issues. The methods of Heuristic based modification include perturbation, which is accomplished by the alteration of an attribute value by a new value (i.e., changing a 1-value to a0- value, or adding noise), and blocking, which is the replacement of an existing attribute value with a "?" [10,11,12]. Some of the approaches used are as follows.

M. Atallah et al [13], tried to deal with the problem of limiting disclosure of sensitive rules. They attempt to selectively hide some frequent item sets from large databases with as little as possible impact on other, non-sensitive frequent item sets. They tried to hide sensitive rules by modifying given database so that the support of a given set of sensitive rules, mined from the database, decreases below the minimum support value.

N. Radadiya [14] proposed an algorithm called ADSRRC which tried to improve DSRRC algorithm. DSRRC could not hide association rules with multiple items in the antecedent (L.H.S) and consequent (R.H.S.), so it uses a count of items in consequence of the sensible rules and also modifies the minimum number of transactions to hide maximum sensitive rules and maintain data quality.

Y. Guo [15] proposed a framework with three phases: mining frequent set, performing sanitation algorithm over frequent item sets, and generate released database by using FP-tree-based inverse frequent set mining.

Border-based: In this approach by the concepts of borders, the algorithm tries to preprocess the sensitive rules, so the minimum number of them will be censored. Afterward, Database quality will maintain as well while side effects will be minimized [14,9]. One of the approaches used are as follows.

Y. Jain et al [16] proposed two algorithms called ISL (Increase Support of Left hand side) and DSR (Decrease Support of Right hand side) to hide useful association rule from transaction data. In ISL method, confidence of a rule is decreased by increasing the support value of Left Hand Side (L.H.S.) of the rule, so the items from L.H.S. of a rule are chosen for modification. In DSR method, confi dence of a rule is decreased by decreasing the support value of Right Hand Side (R.H.S.) of a rule, so items from R.H.S. of a rule are chosen for modification. Their algorithm prunes number of hidden rules with the same number of transactions scanned, less CPU time and modification.

Exact: In this approach it tries to formulate the hiding problem to a constraint satisfactory problem (CSP). The solution of CSP will provide the minimum number of transactions that have to be sanitized from the original database. Then solve it by helping binary integer

Copyright (c) 2015 Advances in Computer Science: an International Journal. All Rights Reserved.

programming (BIP), such as ILOG CPLEX, GNU GLPK or XPRESS-MP [14, 9]. Although this approach presents a better solution among other approaches, high time complexity to CSP is a major problem. Gkoulalas and Verykios proposed an approach in finding an optimal solution for rule hiding problems [17].

Reconstruction-Based: A number of recently proposed techniques address the issue of privacy preservation by perturbing the data and reconstructing the distributions at an aggregate level in order to perform the association rules mining. That is, these algorithms are implemented by perturbing the data first and then reconstructing the distributions. According to different methods of reconstructing the distributions and data types, the corresponding algorithm is not the same. Some of the approaches used are as follows.

Agrawal et al. [18] used Bayesian algorithm for distribution reconstruction in numerical data. Then, Agrawal et al.[19] proposed a uniform randomization approach on reconstruction-based association rule to deal with categorical data. Before sending a transaction to the server, the client takes each item and with probability p replaces it by a new item not originally present in this transaction. This process is called uniform randomization. It generalizes Warner's "randomized response" method. The authors of [20] improved the work over the Bayesian-based reconstruction procedure by using an EM algorithm for distributionreconstruction.

Chen et. al. [21] first proposed a Constraint-based Inverse Itemset Lattice Mining procedure (CIILM) for hiding sensitive frequent itemsets. Their data reconstruction is based on itemset lattice. Another emerging privacy preserving data sharing method related with inverse frequent itemset mining is inferring original data from the given frequent itemsets. This idea was first proposed by Mielikainen [22]. He showed finding a dataset compatible with a given collection of frequent itemsets is NPcomplete.

A FP-tree based method is presented in [23] for inverse frequent set mining which is based on reconstruction technique. The whole approach is divided into three phases: The first phase uses frequent itemset mining algorithm to generate all frequent itemsets with their supports and support counts from original database D. The second phase runs sanitization algorithm over

frequent itemset FS and get the sanitized frequent itemsets of FS'. The third phase is to generate released database D' from FS' by using inverse frequent set mining algorithm. But this algorithm is very complex as it involves generation of modified dataset from frequent set.

Cryptography-Based: In many cases, multiple parties may wish to share aggregate private data, without leaking any sensitive information at their end. This requires secure and cryptographic protocols for sharing the information across the different parties[24,25,26,27]. one of the approache used are as follows. The paper proposed by Assaf Schuster et al.[28] presents a cryptographic privacy-preserving association rule mining algorithm in which all of the cryptographic primitives involve only pairs of participants. The advantage of this algorithm isits scalability and the disadvantage is that, a rule cannot be found correct before the algorithm gathers information from k resources. Thus, candidate generation occurs more slowly, and hence the delay in the convergence of the recall. The amount of manager consultation messages is also high.

5. Conclusion

We present a classification and an extended description and clustering of various algorithms of association rule mining. The work presents in here, which indicates the ever increasing interest of researchers in the area of securing sensitive data and knowledge from malicious users. At present, privacy preserving is at the stage of development. Many privacy preserving algorithms of association rule mining are proposed, however, privacy preserving technology needs to be further researched because of the complexity of the privacy problem.

References

[1] S.R.M. Oliveira, O.R. Zaıane, Y. Saygin, "Secure association rule sharing, advances in knowledge discovery and data mining, in: Proceedings of the 8th Pacific-Asia Conference (PAKDD2004), Sydney, Australia, 2004, pp.74–85.

[2] Elena Dasseni, Vassilios S. Verykios, Ahmed K.Elmagarmid, and Elisa Bertino, "Hiding Association

Copyright (c) 2015 Advances in Computer Science: an International Journal. All Rights Reserved.

Rules by using Confidence and Support," In Proceedings of the 4th Information Hiding Workshop (2001), pp.369–383.

[3] Verykios, V.S., Elmagarmid, A., Bertino, E., Saygin, Y., and Dasseni, E. Association rule hiding. IEEE Transactions on Knowledge and Data Engineering, 2004, 16(4):434-447

[4] Oliveira, S.R.M., Zaiane, O.R.: Privacy preserving frequent itemset mining. In: IEEE icdm Workshop on Privacy, Security and Data Mining, vol. 14, pp. 43–54 (2002).

[5] Oliveira SRM, Zaiane OR (2006) A unified framework for protecting sensitive association rules in business collaboration. Int J Bus Intell Data Min 1:247–287.

[6] HajYasien A (2007) Preserving privacy in association rule mining. Ph. D Thesis, University of Griffith.

[7] Oliveira SRM, Za OR, Zaiane OR, Saygin Y (2004) Secure association rule sharing. Adv. Knowl. Discov. Data Min. Springer, pp 74–85.

[8] Verykios VS, Gkoulalas-Divanis A (2008) Chapter 11 A Survey of Association Rule Hiding Methods for Privacy. Privacy-Preserving Data Min 267–289.

[9] Gkoulalas-Divanis A, Verykios VS (2010) Association rule hiding for data mining. Springer.

[10]Oliveira SRM, Zaiane OR (2002) Privacy preserving frequent itemset mining. Proc. IEEE Int. Conf. Privacy, Secur. data mining-Volume 14. pp 43–54

[11] Verykios VS, Pontikakis ED, Theodoridis Y, Chang L (2007) Efficient algorithms for distortion and blocking techniques in association rule hiding. Distrib Parallel Databases 22:85–104. doi: 10.1007/s10619-007-7013-0

[12] Saygin Y, Verykios VS, Clifton C, Saygm Y (2001) Using unknowns to prevent discovery of association rules. ACM SIGMOD Rec 30:45–54.

[13] Atallah M, Bertino E, Elmagarmid a., et al. (1999) Disclosure limitation of sensitive rules. Proc. 1999 Work. Knowl. Data Eng. Exch. (Cat. No.PR00453)

[14] Radadiya NR, Prajapati NB, Shah KH (2013) Privacy Preserving in Association Rule mining. 2:208–213.

[15] Guo Y (2007) Reconstruction-based association rule hiding. Proc. SIGMOD2007 Ph. D. Work. Innov. Database Res. pp 51–56

[16] Jain YK, Yadav VK, Panday GS (2011) An Efficient Association Rule Hiding Algorithm for Privacy Preserving Data Mining. Int J Comput Sci Eng 3:2792–2798

[17] Gkoulalas-Divanis A, Verykios VS (2006) An integer programming approach for frequent itemset hiding. Proc. 15th ACM Int. Conf. Inf. Knowl. Manag. ACM Press, New York, New York, USA, pp 748–757

[18] Chris Clifton, Murat Kantarcioglou, XiadongLin and Michaed Y.Zhu, "Tools for privacy preserving distributed data mining," SIGKDD Explorations 4, no. 2, 2002.

[19] Alexandre Evfimievski, Ramakrishnan Srikant, Rakesh Agrawal, Johannes Gehrke. Privacy Preserving Mining of Association Rules. SIGKDD 2002, Edmonton, Alberta Canada.

[20] D. Agrawal and C. C. Aggarwal, "On the design and quantification of privacy preserving data mining algorithms", In Proceedings of the 20th Symposium on Principles of Database Systems, Santa Barbara, California, USA, May, 2001.

[21] Chen, X., Orlowska, M., and Li, X., "A new framework for privacy preserving data sharing.", In: Proc. of the 4th IEEE ICDM Workshop: Privacy and Security Aspects of Data Mining. IEEE Computer Society, 2004. 47-56.

[22] Mielikainen, T. "On inverse frequent set mining". In: Proc. of the 3rd IEEE ICDM Workshop on Privacy Preserving Data Mining. IEEE Computer Society, 2003. 18-23.

[23] ZongBo Shang; Hamerlinck, J.D., "Secure Logistic Regression of Horizontally and Vertically Partitioned Distributed Databases," Data Mining Workshops, ICDM Workshops 2007. Seventh IEEE International Conference on 28-31 Oct. 2007, pp.723–728.

[24] DuW., AtallahM.: SecureMulti-party Computation: A Review and Open Problems.CERIAS Tech. Report 2001-51, Purdue University, 2001.

[25] Ioannidis, I.; Grama, A, Atallah, M., "A secure protocol for computing dot-products in clustered and distributed environments," Proceedings of International Conference on Parallel Processing, 18-21 Aug. 2002, pp.379–384.

[26] A. Sanil, A. Karr, X. Lin, and J. Reiter, "Privacy preserving analysis of vertically partitioned data using secure matrix products," Journal of Official Statistics, 2007.

[27] M. Kantarcioglu, C. Clifton, "Privacy-preserving distributed mining of association rules on horizontally partitioned data," The ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD'02). ACM SIGMOD'2002 [C]. Madison, Wisconsin, 2002, pp.24–31.

[28]Assaf Schuster, Ran Wolff, Bobi Gilburd," Privacy-Preserving Association Rule Mining in LargeScale Distributed Systems", fourth IEEE symposium on Cluster Computing and Grid, 2004.

Copyright (c) 2015 Advances in Computer Science: an International Journal. All Rights Reserved.