International Journal of Engineering Research and Development
ISSN: 2278-067X, Volume 1, Issue 8 (June 2012), PP.18-21
www.ijerd.com

Privacy Preserving Association Rule in Data Mining

Jayashree Patil[1], Y.C.Kulkarni[2]
[1,2]*BVDUCOE, Pune,India,*

Abstract—*Privacy is an important issue in Data mining. The privacy field has seen speedy advances in current years because ability to store data has increased. Precisely, current advances in the data mining field have led to increased concerns about privacy. Privacy-preserving data mining has been studied extensively, because of the wide proliferation of sensitive information on the internet. . Many methods have been brought out to solve this. As a result privacy becomes one of the prime anxieties in data mining research public. A new class of data mining approaches, known as privacy preserving data mining algorithms, has been developed by the research public working on security and knowledge discovery. The goal of these algorithms is the extraction of relevant knowledge from large amount of digital data and while protecting at the same time sensitive information. Several data mining techniques, incorporating privacy protection mechanisms, have been advanced that allow one to hide sensitive item sets or patterns, before the data mining process is executed. Association rule mining helps to preserves the confidentiality of each database. To find the association rule, each participant has to segment their own data. Thus, much privacy information may be transmitted or been illegal used. Association rule mining is one of the vital problems in data mining, privacy preserving classification methods, instead, prevent a miner from building a classifier which is able to predict sensitive data.*

Keywords—*Data mining, Data hiding, Knowledge hiding, Association Rule, Privacy preserving*

I. INTRODUCTION

Data mining technology has been developed with the goal of providing tools for automatically and intelligently transforming large amount of data in knowledge relevant to users. The extracted knowledge often expressed in form of association rules decision trees or clusters, allows one to find interesting patterns and regularities deeply buried in the data, which are meant to facilitate decision making processes. Such a knowledge discovery process however can also return sensitive information about individuals, compromising the individual's right to privacy. Moreover data mining techniques can reveal critical information about business transactions, compromising the free competition in a business setting. Thus, there is a strong need to prevent disclosure not only of confidential personal information, but also of knowledge which is considered sensitive in a given context. For this reason, recently much research effort has been devoted to addressing the problem of privacy preserving in data mining. As a result several data mining techniques, incorporating privacy protection mechanisms, have been developed based on different approaches. For instance, various sanitization techniques have been proposed for hiding sensitive items or patterns that are based on removing reserved information or inserting noise into data. Privacy preserving classification methods instead, prevent a miner from building a classifier able to predict sensitive data. Additionally privacy preserving clustering techniques have been recently proposed, which distort sensitive numerical attributes while preserving general features for clustering analysis.

Given the number of different privacy preserving association rule mining techniques that have been developed over the last years, there is an emerging need of moving toward standardization in this new research area. Because all the various techniques differ among each other with respect to a number of criteria like performance, data quality, privacy level, it is important to provide a systematic and comprehensive study for their evaluation. In many cases no technique is better than the other ones with respect to all criteria. Thus, one has to select the privacy preserving technique based on the criterion to be optimized. A framework like the one developed here is thus essential in order to select the privacy preserving technique which is more adequate based on the data and the requirements of the application domain. In evaluating privacy preserving association rule mining algorithm it is important to assess the quality of the transformed data. In evaluating privacy preserving association rule mining algorithm it is important to assess the quality of the transformed data To do this, we need methodologies for the assessment of the quality of data, intended as the state of the individual items in the database resulting from the application of a privacy preserving technique, as well as the quality of the information that is extracted from the modified data by using a given data mining method. This presents such a methodology which allows one to compare the various privacy preserving techniques on a common platform. The methodology consists of a number of evaluation criteria and a set of tools for data pre-processing and privacy preserving data mining algorithm evaluation [1].

II. BACKGROUND

As we know, the data processed in data mining may be obtained from many sources in which different data types may be used. However, no algorithm can be applied to all applications due to the difficulty for fitting data types of the algorithm, so the selection of an appropriate mining algorithm is based on not only the goal of application, but also the data fit ability. Therefore, to transform the non-fitting data type into target one and protecting private data is also an important work in data mining, but the work is often tedious or complex since a lot of data types exist in real world. Merging the similar data types of a given selected mining algorithm into a generalized data type seems to be a good approach to reduce the transformation complexity. Recent research in the area of privacy preserving data mining has devoted much effort to

determine a trade-off between the right to privacy and the need of knowledge discovery which is crucial in order to improve decision-making processes and other human activities. Such research has resulted in several approaches to the evaluation of privacy preserving techniques.

Verykios et al. (2004) analyze the state-of-the-art in the area of Privacy Preserving Data Mining (PPDM), classifying the proposed privacy preservation techniques according to five different dimensions:

- Data distribution (centralized or distributed).
- The modification applied to the data (encryption, perturbation, generalization, and so on) in order to sanitize them.
- The data mining algorithm which the privacy preservation technique is designed for.
- The data type (single data items or complex data correlations) that need to be protected from disclosure.
- The approach adopted for preserving privacy (heuristic, reconstruction or cryptography based approaches).

Oliveira and Zaiane (2002) propose a heuristic-based framework for preserving privacy in mining frequent item sets. They focus on hiding a set of frequent patterns, containing highly sensitive knowledge. They propose a set of sanitized algorithms that only remove information from a transactional database, also known in the SDC area as non-perturbative algorithms, unlike those algorithms, that modify the existing information by inserting noise into the data, referred to as perturbative algorithms. The algorithms proposed by Oliveira and Zaiane rely on a item-restriction approach, in order to avoid the addition of noise to the data and limit the removal of real data. In the evaluation of the proposed algorithms they introduce some measures quantifying the effectiveness and the efficiency of their algorithms. The first parameter is evaluated in terms of: Hiding Failure, that is, the percentage of restrictive patterns that are discovered from the sanitized database Misses Cost, that is, the percentage of non-restrictive patterns that are hidden after the sanitization process; Artifactual Pattern, measured in terms of the percentage of discovered patterns that are artifacts[4].

The specific algorithms proposed by Oliveira and Zaiane do not introduce any artifactual patterns, whereas with respect to the hiding failure and misses cost parameters, it turns out that the more restrictive patterns are hidden, the more legitimate patterns are missed. The specification of a disclosure threshold Φ, representing the percentage of sensitive transactions that are not sanitized, allows one to find a balance between the hiding failure and the number of misses. The efficiency of the algorithms is measured in terms of CPU time, by first keeping constant both the size of the database and the set of restrictive patterns, and then by increasing the size of the input data in order to assess the algorithm scalability. Moreover, Oliveira and Zaiane propose three different methods to measure the dissimilarity between the original and sanitized databases. The first method is based on the difference between the frequency histograms of the original and the sanitized databases. The second method is based on computing the difference between the sizes of the sanitized database and the original one. The third method is based on a comparison between the contents of two databases.

In Sweeney (2002) Sweeney proposes a heuristic-based approach for protecting raw data through generalization and suppression techniques. The methods she proposes provide K-Anonymity. Roughly speaking a database is K-anonymous with respect to some attributes if there exist at least k transactions in the database for each combination of the attribute values. A database A can be converted into a new database A1 that guarantees the K Anonymity property for a sensible attribute by performing some generalizations on the values of the target attributes. As result, such attributes are susceptible to cell distortion due to the different level of generalization applied in order to achieve K Anonymity techniques for hiding sensitive information [5].

A framework for mining association rules from transactions consisting of categorical items is proposed (Evfimievski et al. (2002) where the data has been randomized to preserve privacy of individual communications, ensuring at the same time that only true associations are mined. They also provide a formal definition of privacy cracks and a class of randomization operators that are effective in limiting cracks than uniform randomization. Definition accordingly (Evfimievski et al. (2002) an item set A results in a privacy break of level ρ if the likelihood that an item in A belongs to a non randomized transaction, given that A is included in a randomized transaction, is greater or equal to ρ . In some scenarios, being confident that item be not present in the original transaction may be considered a privacy breach. In order to evaluate the privacy cracks, the approach taken by Evfimievski et al. is to count the occurrences of an item set in a randomized transaction and in its sub-items in the corresponding non randomized transaction. Out of all sub-items of an item set, the item causing the worst privacy breach is chosen. Then, for each combination of transaction size and item set size, the worst and the average value of this breach level are computed over all frequent item sets. Finally, the item set size giving the worst value for each of these two values is selected.[2].

Clifton (2002) defined a Cryptography-based approach. Such approach reports the problem of association_ rule_ mining in vertically_ partitioned_ data. Aim is to determine the item frequency when transactions are split through different sites, without revealing the contents of individual transactions. A security and communication analysis is also presented. In particular, the security of the protocol for computing the scalar product is analysed. The total communication cost depends on the number of candidate item sets and can best be expressed as a constant multiple of the I/O cost of the aprior algorithm [3].

III. EVALUATION MODEL

The main goals of a privacy preserving association rule mining algorithm should enforce:

- A privacy preserving association rule mining algorithm should have to prevent the discovery of sensible information.
- It should be resistant to the various data mining techniques.
- It should not compromise the access and the use of non sensitive data.
- It should be usable on large amounts of data.
- It should not have an exponential computational complexity

Current privacy preserving association rule mining algorithms do not satisfy all these goals at the same time; for instance, only few of them satisfy the point (2). The above list of goals helps us to understand how to evaluate these algorithms in a common way. The framework is identified is based on the following evaluation dimensions:

- **Efficiency -** the ability of a privacy preserving algorithm to execute with good performance in terms of all the resources implied by the algorithm
- **Scalability -** which evaluates the efficiency trend of a privacy preserving data mining algorithm for increasing sizes of the data from which relevant information is mined while ensuring privacy
- **Data quality –** After the application of a privacy preserving technique, considered both as the quality of data themselves and the quality of the data mining results after the hiding strategy is applied
- **Hiding failure -** The portion of sensitive information that is not hidden by the application of a privacy preservation technique. Privacy level offered by a privacy preserving technique, which estimates the degree of uncertainty, according to which sensitive information, that has been hidden, can still be predicted.

IV. CURRENT METHODOLOGY

To measure the effective accuracy of the parameters some experimental evaluations on a set of data hiding algorithms have been carried out.. The data set and methodology which have been used are described for the evaluation process.

A. Rule hiding Problem:

A formal description of the rule hiding problem is given. Let D be a transactional database and I = {i1,. . . , in} be a set of literals, called items. Each transaction can be considered as an item set that is included in I. The items in a transaction or in an item set are sorted according to a lexicographic order. According to a bitmap notation, each transaction t in the database D is represented as a triple <TID ,values of items, size>, where TID is the identifier of the transaction t, and values of items is a list of values, one value for each item in I, associated with transaction t. An item is supported by a transaction t if its value in the values of items is 1, and it is not supported by t if its value in the values of items is 0. Size is the number of 1 value which appears in the values of items, that is, the number of items supported by the transaction. An association rule is an effect of the form X => Y between two disjoint item sets X and Y in I. Each rule is assigned both a support and a confidence value. The first one is a measure of a rule frequency, more precisely, it is the chance to find in the database transactions containing all the items in XȕY, whereas the confidence is a measure of the strength of the relation between the antecedent X and the consequent Y of the rule, that is, the probability to find transactions containing all the items in XȕY, once we know that they contain X. An association rule mining process consists of two steps: first step is the identification of all the frequent item sets, that is, all the item sets, whose supports are higher than a pre-determined minimum support threshold, minsupp; and second step is the generation of strong association rules from the frequent item sets, that is, those frequent rules whose confidence values are higher than a minimum confidence threshold, minconf. Along with confidence and support, a sensitivity level is assigned only to both frequent and strong rules. If a strong and frequent rule is above a certain sensitivity level, the hiding process should be applied in such a way that either the frequency or the strength of the rule is reduced below the min_supp and the min_conf correspondingly. The problem of association rule hiding can be stated as follows: given a database D, a set R of relevant rules that are mined from D and a subset Rh of those sensitive rules included in R, we want to transform D into a database D' in such a way that the rules in R can still be mined, except for the rules in Rh.

B. Rule hiding algorithms based on data fuzziffication:

According to the data fuzzification method , a sequence of symbol in the new alphabet of an item {0,1,?} is related with each transaction where one symbol is associated with each item in the set of items I. The ith value in the list of values is 1 if the transaction supports the ith item, it is 0 otherwise. The novelty of this approach is the insertion of an uncertainty symbol, e.g. a question mark, in a given position of the list of values which means that there is no information on whether the transaction supports the corresponding item. In this case, the confidence and the support of an association rule may be not uniquely determined, but they can range between a minimum and a maximum value. The least support of an item set is defined as the percentage of transactions that certainly support the item set, while the extreme support represents the percentage of transactions that support or could support the item set. The minimum confidence of a rule is obviously the minimum level of confidence that the rule can accept based on the support value, and similarly for the maximum confidence.

A rule hiding process takes place according to two different strategies: decreasing its support or its confidence. The adopted alternative strategies aim at introducing uncertainty in the frequency or the importance of the rules to hide.

The two strategies reduce the minimum support and the minimum confidence of the item sets generating these rules below the minimum support threshold (MST) and minimum confidence threshold (MCT) correspondingly by a certain safety margin (SM) fixed by the user. In order to reduce the support of the large item set generating a sensitive rule, Algorithm 1 replaces 1's by " ? " for the items in transactions associate the item set until its minimum support goes below the minimum support threshold MST by the fixed safety margin SM. Algorithms CR1 and CR2 operate by reducing the minimum sureness value of sensitive rules. The first one decreases the minimum support of the generating item set of a sensitive rule by replacing items of the rule consequent with unknown values. The second one, instead, increases the supreme support value of the antecedent of the rule to hide via placing question marks in the place of the zero values of items in the antecedent. All the fuzzification algorithms hide a sensitive rule with an uncertainty level by decreasing the minimum support or confidence values below the resulting thresholds, MST-SM and MCT-SM.

V. CONCLUSION

In this paper, a methodology for evaluating privacy preserving association rule mining algorithms is proposed. Such methodology allows one to assess the different features of a privacy preserving algorithm according to a variety of evaluation criteria. Parameters like level of privacy, data quality, data types fitability and hiding failure have been defined and the evaluations of such parameters over a set of association rule hiding algorithms have been presented, which can help users to select suitable data mining algorithms for their data sets. The proposed evaluation methodologies can be applied in new set of privacy preservation like cryptography-based and web-based algorithms.

REFERENCES

[1]. Agrawal, R. and Srikant, R. 2000. *Privacy preserving datamining.*In Proceeedings of the ACM SIGMOD conferenceof management of data, ACM, pp. 439.450.

[2]. J Evfimievski, A. 2002. *Randomization in privacy preservingdata mining*. SIGKDD Explor.Newsl., 4(2):43.48J. Breckling, Ed., *The Analysis of Directional Time Series: Applications to Wind Speed and Direction*, ser. Lecture Notes in Statistics. Berlin, Germany: Springer, 1989, vol. 61.

[3]. Kantarcioglu, M. and Clifton, C. 2002. *Privacy preserving distributed mining of association rules on horizontallypartitioned data*. In ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, pp. 24.31.

[4]. Oliveira, S.R.M. and Zaiane, O.R. 2002. *Privacy preservingfrequent itemset mining*. In IEEE icdm Workshop onPrivacy, Security and Data Mining, Vol. 14, pp. 43.54.

[5]. Sweeney, L. 2002. *Achieving k-anonymity privacy protectionusing generalization and suppression*. International Journalof Uncertainty, Fuzziness and Knowledge Based Systems,10(5):571.588.