IMPACT: International Journal of Research in Engineering & Technology (IMPACT: IJRET)
ISSN(E): 2321-8843; ISSN(P): 2347-4599
Vol. 2, Issue 6, Jun 2014, 51-58

A ROBUST TECHNIQUE FOR PRIVACY PRESERVATION OF OUTSOURCED TRANSACTION DATABASE

VINEET RICHHARIYA[1] & PRATEEK CHOUREY[2]

[1]HOD, Department of Computer Science & Engineering, LNCT, Bhopal, Madhya Pradesh, India

[2]Research Scholar, Department of Computer Science & Engineering, LNCT, Bhopal, Madhya Pradesh, India

ABSTRACT

Privacy Preserving Data Mining (PPDM) is used to extract relevant knowledge from large amount of data and at the same time protect the sensitive information from the data miners. The enhancement of data mining research will be the development of techniques that incorporate privacy concerns. This paper provides an enhanced technique for preserving privacy of association rules as well as private data of individuals in an outsourced business transaction database. As the importance of business transaction data has increased manifolds and the data has become an essential part of any business. This paper implement privacy by using a perturbation technique using jointly Gaussian Function that will not only maintain the privacy of association rules present in the dataset but also the sensitive attributes of individuals contained in it. Using this approach we are reducing time complexity, space complexity, and fake and false rules problems.

KEYWORDS: Privacy Preserving Mining, Association Rule Mining, Data Perturbation

INTRODUCTION

Data mining is to extract information from large databases. Data mining is the process of discovering new patterns from large data sets which gives advantages for research, marketing analysis, medical diagnosis, atmosphere forecast etc. Data mining is under attack from privacy advocates because of a misunderstanding about what it actually is and a valid concern about how it's generally done. This has caused concerns that personal data may be used for a variety of intrusive or malicious purposes. Privacy preserving data mining help to achieve data mining goals without sacrificing the privacy of the individuals and without allowing learning underlying original data values.

Association rule mining is a technique in data mining that identifies the regularities found in large volume of data [1, 2]. This technique could be compromised when allowing third party to identify and reveal hidden information that is private for an individual or organization. Privacy-preserving data mining using association rule refers to the area of data mining that seeks to safeguard sensitive information from unsolicited or unsanctioned disclosure.

As with the advancement of technology and worldwide connectivity through internet the privacy of dataset stored at different stations, whether they are stored in a centralized server for ease of access, has become important. The privacy of individual data or the dataset as whole that might be used for data mining has become so important and hence increasing the need for extensive research towards their privacy that could be done in different ways.

A company (data owner) lacking in expertise or computational resources can outsource its mining needs to a third party service provider (through server). However, both the items and the association rules of the outsourced database are considered private property of the corporation (data owner). To protect corporate privacy of business transaction

database, the data owner transforms its data and ships it to the server. Normally the dataset is in table format. Adversaries can use that data for deducing any relations or any sensitive data from it by applying linking attacks on quasi identifiers and sensitive attributes.

Protecting sensitive information in the context of our research encompasses two important goals: knowledge protection and privacy preservation. The former is related to privacy preserving association rule mining, while the latter refers to privacy-preserving clustering. An interesting aspect between knowledge protection and privacy preservation is that they have a common characteristic. For instance, in knowledge protection, an organization is the owner of the data so it must protect the sensitive knowledge discovered from such data, while in privacy preservation individuals are the owner of their personal information.

On the other hand, knowledge protection and privacy preservation also have a unique characteristic. Privacy preservation is related to the protection of explicit data (e.g., salary), while knowledge protection is concerned with the protection of implicit data, i.e., patterns discovered from the data. One limitation with the approach of knowledge protection is that the sensitive knowledge should be known in advance by the data owners. In this case, data owners have to mine their databases and use interestingness measures (e.g., support and confidence) with the purpose of finding the valuable patterns, i.e. sensitive knowledge. Subsequently, data owners hide the sensitive knowledge by using the algorithms. The released database is then shared for mining. Another limitation of the approach of knowledge protection is that we do not focus on protecting against correlations between variables, such as salary and age. Rather, we protect specific binary rules (e.g., X->Y), where X and Y represent items purchased in a store or attributes with specific values. Again, these rules are private to the company or organization owning the data and must be protected since they can provide competitive advantage in the business world.

RELATED WORK

Privacy preserving data mining has become a hot spot in data mining research. The main reasons behind it are the importance of private data, enhanced technology, allowing ease of storage, access, transfer, manipulation of centralized and distributed data. To save it from unauthorized access and attacks to get the knowledge many perturbation techniques have been used by various researchers. The attacker can have basic information regarding the dataset. The important distinction between our scenario and others is that, in ours, the results as well as sensitive attributes are not intended to be open to others like that in [3]. There are many techniques that are prevalent for privacy preserving data mining. The literature in [1] has given a gist of the methods that could be used for privacy preserving data mining and extensive work has been done on every one of them, out of that major emphasis has been on anonymization like in [15], perturbation [4] and many others.

Y-H Wu et al. [19] proposed method to reduce the side effects in sanitized database, which are produced by other approaches. They present a novel approach that strategically modifies a few transactions in the transaction database to decrease the supports or confidences of sensitive rules without producing the side effects.

Authors [18] presents a survey of different association rule mining techniques for market basket analysis, highlighting strengths of different association rule mining techniques. As well as challenging issues need to be addressed by an association rule mining technique. Which frequent pattern is utilized is known and it can be utilized for next decision. The results of this evaluation will help decision maker for making important decisions for association analysis.

Our work is mostly for the scenario as described in [3] and has also resemblance with the works like in [12] where there has been much emphasis given to the data quality and with our method we have also maintained data quality, its accuracy but with lesser calculations. Privacy preserving of transaction database has also been done in [17] by anatomy technique which has compromised with data quality and will not be suitable in our scenario. The elaborated work on perturbation in [4] has allowed us to understand various situations where different methods could be applicable on real time data including the use of guassian distribution for reconstruction in two-phase perturbation model.

Issue: The most important issue is of maintaining privacy of not only the individual but of the important association rules that a corporate or a company has through his transaction data base or warehouse which can help them transform their business and help in maintaining competitive edge on their competitors.

Space: In [3] the space complexity is high because of following reasons: They are using fake transactions that are increasing the data base size which is not useful as the data set is stored in server and needs to be used through internet. Maintaining table which stores data about the perturbation done on both the sides (from where the dataset is uploaded and where the data set is downloaded).

Time: Time complexity is also high because of the above mentioned scenarios.

PROPOSED WORK

As the privacy of dataset is important for storing it at different stations for ease of access, which is done in variety of ways but the attacker make the original dataset from the perturbed set. Here dataset is use for the privacy is taken from Cooperative customer expenditure, which has the item index, price, category, etc. In order to put this dataset on the server for different purpose it needs protection from unauthorized user who uses it for unfamiliar activities. As this dataset need to use by the authorized person as well, but such perturbed data is not the correct set for the user to read it or to gain knowledge, so a successful reading of the authorized user can be possible by a lossless recoverable method. For this, we need a method of perturbation which can preserve the privacy as well as it is also easy in removing the perturbation from the dataset. Process of perturbation starts from pre-processing of dataset, which removes those columns from the dataset which are not helpful in mining. Then separated columns can be processed in two modules.

Module 1

Association Rule

Association Rule The support is a measure of the frequency of a rule and the confidence is a measure of the strength of the relation between sets of items. Support(s) of an association rule is defined as the percentage/fraction of records that contain (A ∪ B) to the total number of records in the database.

$$\text{Support}\,(A \Rightarrow B) = \frac{\textit{Support count of } (A \cup B)}{\textit{Total number of transaction in } D}$$

Apriori is a breadth-first, level-wise algorithm is used to implement the association rule. This algorithm have a main steps follow : Exploits monotonicity as much as possible, Search Space is traversed bottom-up, level by level, Support of an itemset is only counted in the database if all its subsets were frequent.

Apriori algorithm approach is A rule X=>Y satisfies minsup an sup (X∩Y) / sup(X) >= minconf. Hence, first find all itemset I s.t. sup (I) >= minsup. Then for every frequent I: Split I in all possible ways X∩Y and Test if sup

(X∩Y) / sup(X) >= minconf. In privacy preserving data mining, association rules are useful for analyzing and predicting customer behavior and pattern of purchase. They play an important part in market analysis, data of basket shopping, product clustering, classification, and catalog design and store layout.

Similarly in this work Association rules are generated from the pre-processed dataset. These rules are generated by the Aprior Algorithm. Now, those rules whose support value is above the minimum support value are to be hidden. Here for hiding these rules, manipulation is done in transaction where other item is inserted into the transaction.

Jointly Gaussian Function

Let G1 through GL be L Gaussian random variables. They are said to be jointly Gaussian if and only if each of them is a linear combination of multiple independent Gaussian random variables. Equivalently, G1 through GL are jointly Gaussian if and only if any linear combination of them is also a Gaussian random variable. A vector formed by jointly Gaussian random variables is called a jointly Gaussian vector. For a jointly Gaussian vector.

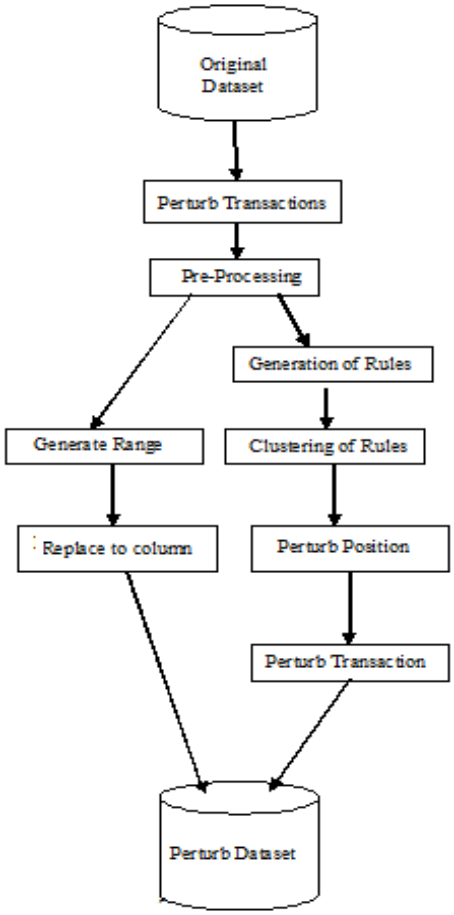

Figure 1: Perturbation Steps

G = |G1, . . .; GL|T, its probability density function (PDF) is as follows: for any real vector g.

$$f_{\mathbb{G}}(g) = \frac{1}{\sqrt{(2\pi)^L \det(K_{\mathbb{G}})}} e^{-(g-\mu_{\mathbb{G}})^T K_{\mathbb{G}}^{-1}(g-\mu_{\mathbb{G}})/2}$$

Where μG and KG are the mean vector and covariance matrix of G, respectively.

Decision of choosing the transaction number in the database where perturb transaction need to add is random & that is generate by the Gaussian function which take two parameter mean, Co-variance.

Finally, perturb the selected transaction which is mention by the Gaussian function for the particular rule to hide it and decrease the overall support value of those rules whose support is greater than minimum value.

Module 2

Here some specific data like age, salary, postal code, etc. are to be hidden which directly specify the user relation with the transaction. This is done by creating the range of particular values and replacing that value with that range, so that individual privacy of the user is also taken care of in this work. For generating the range, modulus function is used that generates remainder by dividing the mean and Gaussian then add that value to the value to range hence also increasing the randomization.

Proposed Perturbation Algorithm

Input: DS (Original Dataset), MS (Minimum Support)

Output: PDS (Perturb Dataset)

- DS ← Pre-Process(DS)
- PDS = DS
- AR[n] ← Aprior(DS) // n number Association rule
- Loop 1:n
- If AR[n] > MS
- FR[m] ← AR[n] // Frequent Rule FR with Mini Supp
- Endif End Loop
- Fakepos[s] ← JointlyGaussian // Generate Random pos
- Loop 1:s
- PDS(Fake_pos) ← Perturb_session (SR, n) // This will reduce the support value
- End Loop

Experiment and Result

This section presents the experimental evaluation of the proposed perturbation and de-perturbation technique for privacy prevention. To obtain AR this work used the Apriori algorithm [1], which is a common algorithm to extract frequent rules. All algorithms and utility measures were implemented using the MATLAB tool. The tests were performed on an 2.27 GHz Intel Core i3 machine, equipped with 4 GB of RAM, and running under Windows 7 Professional. Experiment done on the customer shopping dataset which have collection of items, cost, Total amount, etc. attributes.

Evaluation Parameter

Execution Time

As the work done on the important resource that is server so execution time should be less as possible. So this is a very important parameter to evaluate this work.

Fake Transaction

As the dataset is perturbed by adding the fake transaction in it, so the number of fake transactions one includes depends on the minimum support values of the rules. In order to make proper perturbation number of fake transactions are needed to be controlled, which is done by deciding the proper support value.

Data Set Size

Here size of dataset is analyzed after perturbation. As if the size increases then it require more space to store it on the server.

Originality

The amount of original data present in the dataset after perturbation.

Results

Perturbation is done in the original dataset before sending it to the server. When the Min. supp is increase the frequent pattern is decrease and the execution time is also decrease. This shows in the following table.

The following graph represents the execution time is reduce in the above method.

Table 1: From the [3] Algorithm

Min. Supp	Execution Time	Frequent Pattern	Dataset Size
18	39.1666	30	16958
19	34.8114	28	16447
12	0.9579	30	16698

Graph: 1 Minimum Support versus Frequent Pattern

Table 2: From the Proposed Work

Min. Supp	Execution Time	Frequent Pattern	Dataset Size
12	33.7740	30	15000
18	12.3557	30	15000
19	8.8780	28	15000

From above table we can observe that when the minimum support is increased the frequent pattern is decreased. It indicates that when more support, less rules are indentified so the execution time is less. This relation helps to control the fake transaction addition in the original dataset in [3]. But proposed work has showed that increase of the frequent rules or decrease of min support value will not affect the perturbed copy dataset as the dataset size is always constant. This result also shows that the space required for this is same so proposed fight is better for space complexity as well.

From above table it is also observed that the execution time in the proposed algorithm is less as compared to the [3]. So work done for privacy preserving is good in all sense as compared to the previous work done in [3] in all aspects.

CONCLUSIONS AND FUTURE WORK

Preserving privacy in data mining activities is a very important issue in many applications. Randomization-based techniques are likely to play an important role in this domain. In this paper, a new approach to solve the problem of privacy preserving data mining in the scenario of outsourced business transaction database has been solved successfully. This approach is efficient and better than many other perturbation and anonymization techniques. Proposed algorithm have reduced the time complexity, space complexity as well as false rules problems in effective manner from the previous work.

In future, we will try to make it more powerful for cloud and distributed databases.

REFERENCES

1. Majid Bashir Malik, M. Asger Ghazi, Rashid Ali, “Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects” in IEEE 2012 Third International Conference on Computer and Communication Technology.
2. R. Agrawal, T. Imielinski, and A. N. Swami. “Mining Association Rules between Sets of Items in Large Databases”. SIGMOD 1993.
3. Fosca Giannotti, Laks V. S. Lakshmanan, Anna Monreale, Dino Pedreschi, and Hui (Wendy) Wang, “Privacy-Preserving Mining of Association Rules From Outsourced Transaction Databases” in IEEE SYSTEMS JOURNAL, VOL. 7, NO. 3, SEPTEMBER 2013.
4. Li Liu *, Murat Kantarcioglu, Bhavani Thuraisingham, “The applicability of the perturbation based privacy preserving data mining for real-world data” in L. Liu et al. / Data & Knowledge Engineering 65 (2008) 5–21.
5. L. Qiu, Y. Li, and X. Wu, “Protecting business intelligence and customer privacy while outsourcing data mining tasks,” Knowledge Inform. Syst., vol. 17, no. 1, pp. 99–120, 2008.
6. F. Giannotti, L. V. Lakshmanan, A. Monreale, D. Pedreschi, and H. Wang, “Privacy-preserving data mining from outsourced databases,” in Proc. SPCC2010 Conjunction with CPDP, 2010, pp. 411–426.
7. Zhengyou Zhou, Liusheng Huang, Ye Yun, “Privacy Preserving Attribute Reduction Based on Rough Set”, in IEEE, 2009, Second International Workshop on Knowledge Discovery and Data Mining.
8. Xiaolin Zhang, Hongjing Bi, “Research on Privacy Preserving Classification Data Mining Based on Random Perturbation”, in, International Conference on Information, Networking and Automation (ICINA), 2010
9. F. Giannotti, L. V. Lakshmanan, A. Monreale, D. Pedreschi, and H. Wang, “Privacy-preserving outsourcing of association rule mining,” ISTI-CNR, Pisa, Italy, Tech Rep. 2009-TR-013, 2009.
10. Nikunj H. Domadiya, Udai Pratap Rao, “Hiding Sensitive Association Rules to Maintain Privacy and Data Quality in Database”, in IEEE Journal, 2012.

11. Li Liu, Murat Kantarcioglu and Bhavani Thuraisingham, “Privacy Preserving Decision Tree Mining from Perturbed Data”, in, Proc. of 42nd Hawaii International Conference on System Sciences – 2009.
12. Fang Lu, Wei-jun Zhong, Yu-lin Zhang, Shu-e Mei, “Privacy-preserving Association Rules Mining Using the Grouping Unrelated-question Model”, in IEEE Journal, 2007.
13. R. Mahesh, T. Meyyappan “Anonymization Technique through Record Elimination to Preserve Privacy of Published Data”, in Proc. 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering, Feb. 21-22.
14. Kevin Chiew, Shaowen Qin, “Analysis of Privacy-Preserving Mechanisms for Outsourcing Data Mining Tasks”, in IEEE Journal, 2008.
15. Yingjie Wu, Shangbin Liao, Xiaowen Ruan, Xiaodong Wang, “Privacy Preservation in Transaction Databases based on Anatomy technique”, in IEEE International Conference on Computer Science & Education, 2010
16. D. Narmadha, G. NaveenSundar and S. Geetha,” A Novel Approach to Prune Mined Association Rules in Large Databases”, In IEEE, 2011 pp. 409413.
17. Y. H. Wu, C.M. Chiang and A.L.P. Chen, “Hiding Sensitive Association Rules with Limited Side Effects,” IEEE Transactions on Knowledge and Data Engineering, vol. 19(1), Jan. 2007, pp. 29–42.