Implementation of Jointly Gaussian function with Association Rule for Privacy Preserving Data Mining

Miss. Priti N. Parikh[1], Prof. Shrikant Lade [2], Prof. Mahesh Malviya [3]

[1] *RKDF IST, BHOPAL, M.P, India*
[2] *RKDF IST, BHOPAL, M.P, India*
[3] *RKDF IST, BHOPAL, M.P, India*

***Abstract*— Data mining technology which reveals patterns in large databases could compromise the information that an individual or an organization regards as private. The aim of privacy-preserving data mining is to find the right balance between maximizing analysis results and keeping the inferences that disclose private information about organizations or individuals at a minimum. As there are many methods making the privacy of the dataset but perturbing both the text and numeric data. One new approach of generating the perturbed data then regenerate it back from the perturbed is emerging which is highly vulnerable for protection concern need to be expand. So a secure method is developing in this work which maintains both the security for individual privacy and regeneration of the perturbed dataset. By the use of Association rule and adding of fake transaction at jointly Gaussian position fruitful results are obtain that fulfill both the requirement in very less time.**

***Keywords*—Association Rule Mining, Data Perturbation, De-Perturbation, Fake Transaction, Privacy Preserving Data Mining.**

I. INTRODUCTION

Huge data mining is to extract information from large databases. Data mining is the knowledge discovery process of finding the useful information and patterns from the large database. The management is to obtain hidden information which used for any decision in recent data mining. While dealing with protection of sensitive information it becomes very important to protect data against unauthorized access [1].

A key problem faced to balance the confidentiality of the disclosed data with the legitimate needs of the data users. Data mining requires data preparation which can uncover information or patterns which may compromise confidentiality and privacy. This process becomes necessary to modify the support and pattern (Association Rules). Obtaining a true data between the disclosure and hiding is a tricky process [3]. This can be achieved largely by implementing hiding of rules that expose the sensitive part of the data. Normally hiding of association rule is a method to hide the pattern because association among the data is what is understood by most of the data users. Data perturbation is considered relatively easy and effective techniques in for protecting sensitive data from unauthorized use.

Our interest in data security is based not on physical and technical access methods, but to protect information by using data perturbation techniques and maintain confidential data. The threat to an individual's privacy comes into play when the data has access by any user. A data perturbation technique involves adding random noise to text attributes and numerical attributes, thereby protecting the original data set. Even while recover the original data, these methods allow users the ability to access important aggregate statistics (such as means, correlations and covariance, etc.) from the entire database, thus 'protecting' individual data set. As an example: sales data from store , the case of sales data, an employee may not be able to access what a particular individual purchased from a store on a given day, but the total sales volume for the store on the same day of employee could determine.

II. RELATED WORK

Most results in privacy-preserving data mining assume that the data is either horizontally partitioned (that is, each party to the protocol has some subset of the rows of an imaginary "global database"), or vertically partitioned (that is, each party has some subset of the columns of the "global database") [9].

Y-H Wu et al. [18] 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 [7] 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.

The authors in [11] presented five algorithms namely 1.a, 1.b, 2.a, 2.b, 2.c. All of these algorithms fall in the category of distortion based technique. Algorithms 1.a, 1.b, and 2.a were aimed towards hiding association rules. Algorithms 2.b, 2.c were related to hiding large itemsets. Metrics used in all of these five algorithms were efficiency and side effects. These algorithms were first of their kind in hiding association rules. Side effects of these algorithms were also high.

The author [1] concept in this paper is Privacy Preserving mining of frequent patterns on encrypted outsourced Transaction Database (TDB). They proposed a encryption scheme and adding fake transaction in the original dataset. Their method proposed a strategy for incremental appends and dropping of old transaction batches and decrypt dataset. They also analyse the crack probability for transactions and patterns. The Encryption/Decryption (E/D) module encrypts the TDB once which is sent to the server.

Mining is conducted repeatedly at the server side and decrypted every time by the E/D module [1]. Thus, we need to compare the decryption time with the time of directly executing a priori over the original database.

III. BACK GROUND

In recent years, with the explosive development in technologies of Internet and data processing technologies, the privacy preserving association rule mining has been important in business fields, market analysis, medical diagnostic etc.

A. *Association Rule*

Association rule mining is the process of discovering sets of Items that frequently co-occur in a transactional database to produce significant association rules that hold for the data. Mostly all the existing algorithms for association rules rely on the support-confidence framework. Formally, association rules are defined as follows: Let I = { i_1,i_2… i_m} be a set of items. Let D the data set for relevant data, be a set of data set transactions where each transaction T is a set of items such that T ⊆ I. TID is an identifier for each transaction which is associated. Let A be a set of items. A transaction T is said to contain A if and only if A⊆ T [4]. An association rule is an implication of the form A =>B, where A⊂I, B⊂I and A⋂ B = Φ. The rule A => B holds in the transaction set D with support 's'. The percentage of transaction in D that contains A ∪B is 's' as support. The rule A => B has confidence c in the transaction set D if c is the percentage of transactions in D containing A which also contain B. 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 minimum support and sup (X∪Y) ≥ minsup, sup

$(X \cup Y)/\sup(X) \geq \text{minconf}$. Hence, first find all itemset I s.t. $\sup(I) \geq \text{minsup}$. Then for every frequent I: Split I in all possible ways $X \cup Y$ and Test if $\sup(X \cup Y)/\sup(X) \geq \text{minconf}$.

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.

B. *Jointly Gaussian*.

Let G_1 through G_L 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 [10]. Equivalently, G_1 through G_L 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 $G = |G_1, \ldots; G_L|^T$, 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 μ_G and K_G are the mean vector and covariance matrix of G, respectively.

Note that not all Gaussian random variables are jointly Gaussian. For example, let G_1 be a zero mean Gaussian random variable with a positive variance, and define G_2 as

$$G_2 = \begin{cases} G_1, & \text{if } |G_1| \leq 1; \\ -G_1, & \text{otherwise,} \end{cases}$$

where $|G_1|$ is the absolute value of G_1. It is straightforward to verify that G_2 is Gaussian, but $G_1 + G_2$ are not. Therefore, G_1 and G_2 are not jointly Gaussian.

If multiple random variables are jointly Gaussian [10], then conditional on a subset of them and the remaining variables are still jointly Gaussian. Here as the perturbation is done by adding noise generate by the Jointly Gaussian formula. The actual dataset add with the fake transaction on different position.

IV. PROBLEM FORMULATION

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. 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 the perturbed data is not the correct set for the user to read it, so a successful reading of the authorized user can be possible by a lossless recoverable method. For this method need for perturbing and remove that perturbation from the dataset. As transaction is a collection of item set that is figure out to proper co-relation during the perturbation.

In order to hide the frequent pattern or rule from the dataset the fake transaction of the less frequent rule are added to make it perturb dataset. Here each transactions (Fake Transaction) is adding by some value or replace with the existing combination of the fishy set. In order to over remembrance of which set is replace with this is done by Jointly Gaussian, as it generate values which are constant for fix variance. But it has to remember that where the fake transactions are added in the dataset as it needs to recover or make it de-perturb again. For this random position are generate by the jointly Gaussian formula which are dependent on the mean and co-variance. So the substitution of data is considered as chiper texts which replace the original text. So with the correct knowledge of the mean and covariance one can find the fake transaction position, then these are remove from it.

V. METHODOLOGY

The privacy preserving data mining (PPDM) has aim to preserving customer privacy by different techniques. In this technique, loss of information versus preservation of privacy is always a trade off. The question is, how much are the users willing to compromise their privacy and which method adopted for the security.

C. *Perturbation.*

Data perturbation refers to a data transformation process typically performed by the owners before publishing their data. The data owners want to change the data in a certain way in order to disguise the sensitive information contained in the published datasets. And on the other hand the data owners want the transformation to best preserve those domain specific data properties that are critical for building meaningful data mining models, thus marinating mining task specific data utility of the published datasets.

D. *De-Perturbation*

Here as the server get request of the dataset then it pass minimum support value for calculation of original dataset recovery from the perturbed dataset copy. As many chiper texts are replaced original groups of item then replace those with the original one. Now this support will specify the item set number to be present in the original dataset and on the basis of this it will remove the fake transaction as the position is finding by the Jointly Gaussian.

E. *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. This 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 the perturbed data is not the correct set for the user to read it, so a successful reading of the authorized user can be possible by a lossless recoverable method. For this method need for perturbing and remove that perturbation from the dataset.

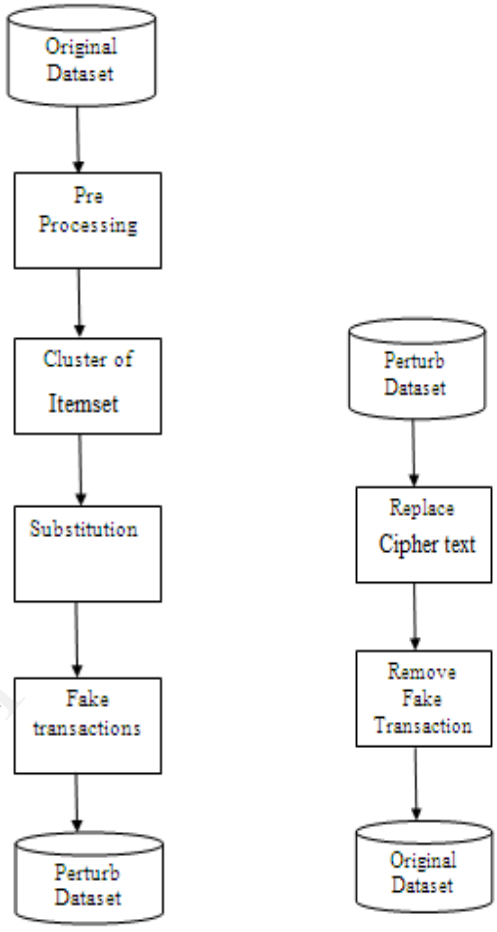


Figure 1: Represent Perturbation and De-perturbation steps.

The above steps have to be flowed for perturbation and de-perturbation of dataset. This represent the flow of process to be performed for these techniques is implemented.

Proposed Perturbation Algorithm

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

Output: PDS (Perturb Dataset)

1. DS ← Pre-Process(DS)
2. AR[n] ← Apriori (DS) /n number Association rule
3. Loop 1:n
4. If AR[n] > MS

5. FR[m] ← AR[n]
   // Frequent Rule FR with Minimum Support.
6. End if
7. End Loop
8. SR[s] ← RobFrugal(FR)
   //Rules set SR & s = total set
9. Fake_pos ← Jointly Gaussian
   // Generate Random position.
10. Loop 1:s
11. PDS(Fake_pos) ← Fake_session(SR, n)
11. End Loop

Proposed De- Perturbation Algorithm

Input: PDS (Perturb Dataset)

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

1. PDS ← Pre-Process(PDS)
2. Fake_pos ← Jointly Gaussian
   // Generate Random position.
3. Loop 1:s
4. DS ← PDS(Fake_pos)
5. End Loop

The main feature of this is in previous work fake transaction positions are store in the table which take memory as well as time and it is constant for all the perturbed copy as well but if it is replace with the Gaussian function that generate a fix sequence and at those place fake transaction are identify.

VI. 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 a 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.

F. *Evaluation Parameter*

Execution time:

As the work done on the important resources that is server so execution time should be less as possible. So the perturbation and de-perturbation take less time. This is a very important parameter to evaluate this work.

Fake Transaction:

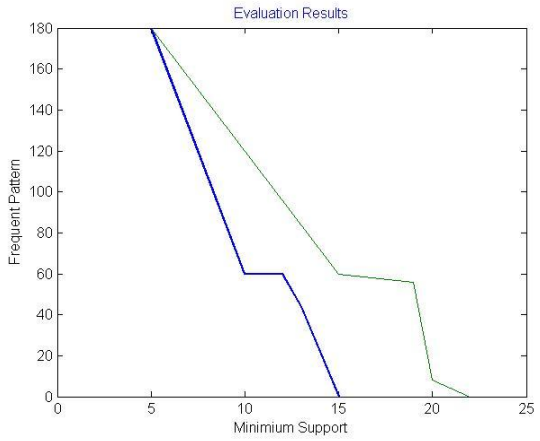
As the dataset is perturbed by adding the fake transaction in it, so the number of fake transaction one include is depend on the minimum support value of the rules. In order to make proper perturbation number of fake transaction are need to be control that is done by deciding the proper support value.

Results:

Perturbation done in the original dataset before sending 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.

.No of Data Item	Minimum Support	Frequent Pattern
10000	1	180
	5	180
	10	60
	12	60
	13	44
	15	0
15000	1	180
	5	180
	15	60
	19	56
	20	8

Table1: Different dataset and Frequent Pattern

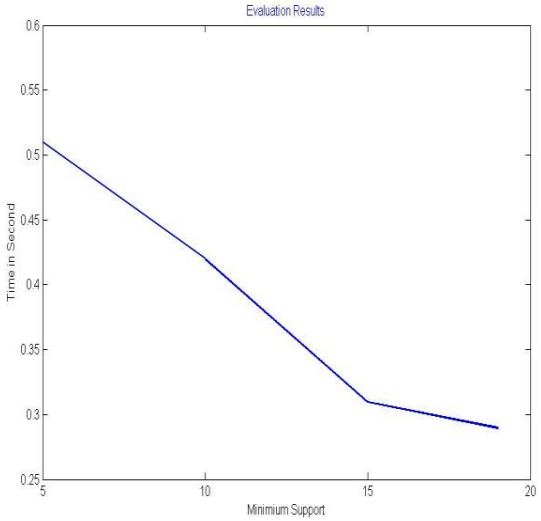


Graph:1 Minimum Support versus Frequent Pattern.

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

Minimum Support	Execution Time	Frequent Pattern
19	0.29	56
15	0.31	60
10	0.42	120
5	0.51	180
1	0.65	180

Table2:ExecutionTime for different Minimum Support.



Graph:2 Minimum Support versus Time in Second

The above graph 2 represents the execution time is reduce in the above method which shows in the table 2. In graph 2 when the minimum support is increases the frequent pattern is decrease. It indicates that when more support, less rules are indentified so the execution time is less.

VII. CONCLUSION

In this paper, we studied the problem of privacy-preserving mining of frequent patterns (from which association rules can easily be computed). We proposed an encryption scheme, called grouping, that is based on 1–1 substitution ciphers for items and adding fake transactions to make each cipher item share the same frequency as ≥ k−1 others. Our work considers the cipher text-only attack model, in which the attacker has access only to the encrypted items. Time complexity and space complexity is reduce as the time required for Jointly Gaussian is low as compared to hash table as well as the space is not required for the same to maintain the number of fake transaction of the dataset.

REFERENCES

[1] 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, pp. 385-395.

[2] N. Zhang and W. Zhao, "*Privacy Preserving Data Mining Systems "In* IEEE Computer society, 2007 pp. 52-58.

[3] W.K. Wong, D. W. Cheung, E. Hung, B. Kao, and N. Mamoulis, " *Security in outsourcing of association rule mining*" In Proc. Int. Conf.Very Large Data Bases, 2007 pp. 111- 122.

[4] K.Sathiyapriya and Dr. G.Sudha Sadasivam, *"A Survey on Privacy Preserving Association Rule Mining",* In IJKDP Vol.3 No 2– March-2013, pp. 119-131.

[5] R. Agrawal and R. Shrikant, "*Privacy-Preserving data mining,"* In Proc.ACM SIGMOD Int. Conf. Manage. Data, 2000, pp. 439–450.

[6] Y.Li,S.Zhu,L.Wang, and S.Jajodia "*A privacy enhanced micro aggregation method",* In Proc. of 2nd International Symposium on Foundations of Information and Knowledge systems. 2002 , pp.148-159.

[7] D.Narmadha, G.NaveenSundar and S.Geetha,"*A Novel Approach to Prune Mined Association Rules in Large Databases*‖", In IEEE, 2011 pp. 409-413.

[8] T zung -Pei, Hong Kuo-Tung Yang, Chun-Wei Lin and Shyue-Liang Wang, "*Evolutionary privacy preserving in data mining "* In IEEE World Automation Congress Conference, 2010, pp.1-7.

[9] Z. Yang and R. N. Wright. "*Privacy-preserving computation of bayesian networks on vertically partitioned data."*, In IEEE Trans. on Knowledge and Data Engineering , 2006, pp.1253–1264.

[10] Yaping Li, Minghua Chen, Qiwei Li, and Wei Zhang "*Enabling Multilevel Trust in Privacy Preserving Data Mining",* IEEE Transactions on Knowledge and Data Engineering, Vol. 24, No. 9, September 2012. pp.1598-1612.

[11] Stanley R. M. Oliveira and Osmar R. Zaiane, "*Privacy preserving frequent itemset mining",* In Proceedings of the IEEE ICDM Workshop on Privacy, Security and Data Mining (2002), pp.43–54.

[12] Jing Wang, Yongcheng Luo, Yan Zhao, Jiajin Le, "*A Survey on Privacy Preserving Data Mining",*] In IEEE Computer Society Washington, In IEEE Computer Society Washington, DC, USA - 2009 pp. 111-114.

[13] Pingshui Wang, "*Research on Privacy Preserving Association Rule Mining A Survey*", IEEE International Conference -2010. pp. 194-198.

[14] T. Dalenius and S. P. Reiss., "Data *Swapping; A technique for disclosure control"*, Journal of Statistical Planning and Interface. -1982. pp.73-85.

[15] Chung-Min Chen, Andrzej Cichocki, Allen McIntosh, Euthimios Panagos, "*Privacy-Protecting Index for Outsourced Databases",* In ICDE Workshops 2013. pp. 83-87.

[16] Wikipedia: . http://en.wikipedia.org/wiki/Data_mining.

[17] S. D. C. di Vimercati, S. Foresti, S. Paraboschi, G. Pelosi, and P. Samarati, "*Efficient and private access to outsourced data*", In ICDCS, 2011, pp. 710–719.

[18] 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.