Volume 4, Issue 12, December 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Multilevel Privacy Preservation Technique for Outsourced Database

Asst. Prof. J. P Maurya
C.S Dept, IES, Bhopal, India

Asst. prof Sandeep Kumar
C.S Dept, IES, Bhopal, India

Sushanki Nikhade
Mtech Scholar IES, Bhopal, India

Abstract— In order to get the hidden knowledge from the dataset one has to apply the mining algorithms. But for hiding the knowledge these information the miner apply Privacy Preserving Data Mining algorithms are used. These techniques prevent the sensitive information by applying different methods of preservation. This paper focus on the enhancement of preserving the frequent association rules by providing the multilevel perturbation in the dataset and provide different level of separate copy for different server in order to increase the confusion of the dataset and those copy is irreversible. Here some items are add to dataset which are not the actual part of it so that perturbation of those are done without adding extra transactions. As the dataset is outsourced so the data need to be deperturbed whenever required so one of the copy should is so designed that it can deperturbed apply reverse technique. Using proposed algorithm this paper reduce time, space complexity.

Keywords:-Privacy Preserving Mining, Association Rule Mining, Data Perturbation.

I. INTRODUCTION

ATA 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 scarifying the privacy of the individuals and without learning underlying data values. Association rule mining is a technique in data mining that identifies the regularities found in large volume of data. Due to this technique 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.

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 [2,3].

Mainly three protocols govern privacy for building a privacy-preserving data mining system. The three protocols entities are shown below [2].

•*Data collection*, manages privacy during data transmission between the data providers to the data ware-house server.
• *Inference control*, protects privacy between the data warehouse server and data mining servers.
• *Information sharing, gives the* control on information shared among the data mining servers in different systems.

II. RELATED WORK

In [3] An Efficient Algorithm for Privacy Preserving Data Mining Using Heuristic Approach 1M.Mahendran, Dr. R. Sugumar. Whole work is divide into three steps first is for K-Anonymity because there are many transaction that give direct information of the customer such as the salary of the customer is the information which one need to be hide, then gender and age are also column of the customer. So out of many approaches of hiding this valuable data of the customer one can easily control by making multiple copy of the same data for increasing the confusion and no one get direct information of the customer.

In [4] Enabling Multilevel Trust in Privacy Preserving Data Mining Yaping Li, Minghua Chen, Qiwei Li, and Wei Zhang. They focus on the multilevel party trust by the use of the data perturbation where they create different dataset for the different users of different trust level. They focus on the problem that if multiple users having different copy of the trust level combine their copies then the original dataset can be regenerate [4]. So in order to overcome this problem they have given new concept of perturbing the dataset of lower trust from the higher perturbed dataset. In this way if all the lower trust level will combine their dataset then they cannot make the as original dataset as of the just higher level one.

Corner-wave Property states that fr M perturbed copies, the privacy goal is achieved if the noise covariance matrix KZZ has the corner-wave pattern as shown in [7]. Specifically, we say that an M xM square matrix has the corner-wave property if, for every from 1 to M, the following entries have the same value as the (i, j) entry. All entries to the right of the (i, j) entry in row i. 2. All entries below the ði; iÞth entry in column i.

In this paper main focus is for the privacy of the outsourced data is done which should get reverse but as well as protected so following issues are covered.

Data: In [4] they only cover the numeric data perturbation while text remain same with different levels so this paper over come with numeric as well text data.

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.

III. PROPOSED WORK

With the ease available of data at the server most of the miners would take advantage of these things so security measures need to be taken that people may get data and it should look original but the security need to be maintained for this. Keeping this goal in mind work focus for providing the privacy of the dataset.

As most of the cooperative dataset are put on the server which may contain different information which is valuable as well as contain information. So to protect it from the unauthorized person privacy need to be develop and proper de-perturbation algorithm also need to be develop for the reuse of the data and lossless recovery. Whole work is divide into two module first is of perturbation other of de-perturbation.

Module 1:

Pre-Processing: Dataset Pre-processing is needed in order to work on it, as the obtain dataset is in different arrangements, many unnecessary information is present which is need to be removed from the dataset.

Perturbation:

1) Cluster of Items: As transaction is a collection of item set that is figure out to Make proper co-relation during the perturbation.
2) Frequent Item: Here by association rule frequent item pattern are find and replace only those by the less frequent set and to remember these transaction one has to generate a sequence that can be obtain by Gaussian function. So special table need to maintain for remembrance at sender, receiver side.
3) Fake transactions: By this no fake transaction need to add in the dataset.

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.

1) 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 reset the transaction as the position is find by the Gaussian function.
2) As many chipper text is replace original groups of item then replace those with the original one.

The main feature of this is in previous work no fake transaction position are add in the dataset as well as no need to 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 perturb transaction are identify.

(Detail structure of the module is explain in the fig. 1.)

A. Association Rule

I. 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 = { i1,i2… im} 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 = i. 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.

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

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**.**

II. Apriori algorithm approach is a rule x=>y satisfies minsup and sup(xuy)/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 xuy and test if sup(xuy)/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.

Perturb Transaction

Although above step start the perturbing the dataset but if one know the replacement words then it will not save any information. So main knowledge in the dataset is the association rule that is generate. In order to hide those rules few fake transaction is added that are of those rules which have low support in the dataset and if lower support rules also have the similar support than knowledge generate from that perturb set is not fruitful.

This include find number of fake transaction for each rule then random position in the dataset where these rules can be placed.

1. Sort all rules according to the support value.
2. Now Find support that is lacking from the maximum support
   Noise = (Max_support – Rule_support)
3. Find fake transaction number for each rule by
   Transaction = (Noise x Dataset_size)/100

Once it come to know that how many fake transaction for a single rule need to generate then next step is to find at which transaction position it can be insert.By using Multivariate normal cumulative distribution functionrandom position are generate for the perturbation of the transaction so overall support will be reduce.

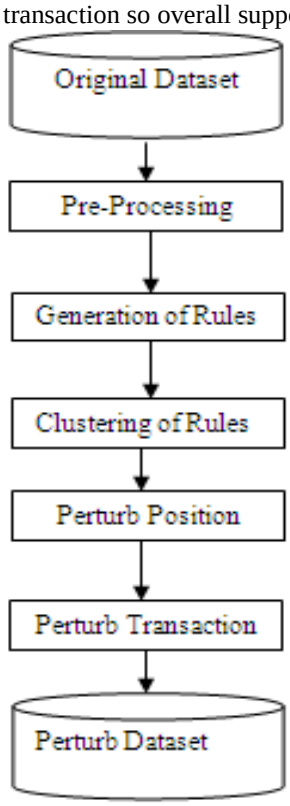

Fig-1. Perturbation Steps

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

Proposed Perturbation Algorithm

Input: DS (Original Dataset), MS (Minimum Support)
Output: PDS (Perturb Dataset)

1. DS ← Pre-Process(DS)
2. PDS = DS
3. AR[n] ← Aprior(DS) // n number Association rule

4. Loop 1:n
5. If AR[n] > MS
6. FR[m] ← AR[n] // Frequent Rule FR with Mini Supp
7. Endif
8. End Loop
9. Fakepos[s] ← MVNCDF() // Generate Random pos
10. Loop 1:s
11. PDS(Fake_pos) ← Perturb_session (FR, n) // This will reduce the support value
12. End Loop

Deperturbation: At the receiving side deperturbation of the dataset is done under following steps. Which simple pre-processes the dataset as done in perturbation. Then Process whole data in two modules.

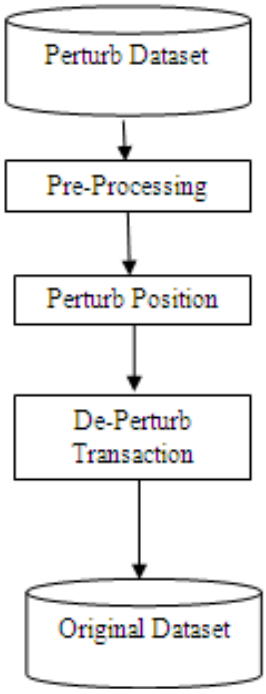

Fig.2 De-Perturbation Steps

Module 1:
Here it generate the random position by MVNCDF function, then de-perturb those transaction by replacing the original items at the specified positions.

Proposed De-Perturbation Algorithm
Input: PDS (Perturb Dataset)
Output: DS (Original Dataset), MS (Minimum Support)

1. PDS ← Pre-Process(PDS)
2. Fakepos[s] ← MVNCDF // Generate Random pos
3. Loop 1:s
4. DS[s] ← PDS(s)
5. End Loop

The main feature of this is in previous work perturb transaction position 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 MVNCDF function that generate a fix sequence and at those place perturb transaction are introduce.

IV. 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.

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.

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:

Experiment start by passing the original dataset then minimum support for the rules to be in safe mode. With these input experiment start by increasing the size of the dataset one can find the different evaluation results. One more parameter for result variation is the support value for the rules for less support perturbation get automatically increases.

Table 1: Perturbation results From the [3] algorithm

Min. Supp	Perturbation		
	Execution Time	Originality	Dataset Size
19	32.114	93	16447
20	31.266	95	163088

Table 2: De-Perturbation results From the [3] algorithm

Min. Supp	Perturbation		
	Execution Time	Originality	Dataset Size
19	2.59	100	15000
20	2.205	100	15000

From above table 1 and 2 it can be conclude that with the increase pf support value perturbation of dataset is increase, while in case of the execution time similar raise of time is found as the perturb new transaction is add to the dataset so the dataset size after perturbation increase. While in case of the de-perturbation two step get reduce first is rob frugal and other is rules generation so execution time get reduce sharply.

Table 3: Perturbation results From the Proposed work.

Min. Supp= 20	Proposed Work Perturbation		
	Execution Time	Originality	Dataset Size
First level	1.2751	96.4867	15000
Second level	2.3606	85.9533	15000
Third level	2.4547	77.4867	15000

Table 4: Perturbation results From the Proposed work.

Min. Supp= 19	Proposed Work Perturbation		
	Execution Time	Originality	Dataset Size
First level	2.587	88.6867	15000
Second level	2.3417	79.94	15000
Third level	2.4053	73.76	15000

From above table 3 and 4 it can be conclude that with the increase of support value perturbation of dataset is increase, while in case of the execution time similar raise of time is found. As the perturb not include new transaction to the dataset so the dataset size after perturbation remain same in this way space complexity get reduce.With the removal of the hash table and replace it by *Multivariate normal cumulative distribution functionwhich directly reduce time complexity for the same minimum support values.*

Table 5: Perturbation results From the Proposed work.

Min. Supp= 19	Proposed Work De-Perturbation		
	Execution Time	Originality	Dataset Size
First level	1.7517	100	15000
Second level	1.7616	96	15000
Third level	2.0209	95	15000

Table 6: Perturbation results From the Proposed work.

Min. Supp= 19	Proposed Work De-Perturbation		
	Execution Time	Originality	Dataset Size
First level	1.7517	100	15000
Second level	1.7419	95	15000
Third level	1.83	92	15000

From above table 5 and 6 it can be conclude that in case of the de-perturbation two step get reduce first is rob frugal and other is rules generation so execution time get reduce sharply. So approx similar results obtain in both the proposed work, and [3] algorithm. This result also shows that the space required for this is same so proposed fight is better for space complexity as well.

It has been tested that by passing the same parameter for the second and third level perturbed copy one cannot generate the original copy of the dataset as it is generated by the first level of the perturbed copy. So the multilevel requirement is also fulfill.

V. CONCLUSION

Researchers are working in different field out of which Preserving privacy mining is one of the new and important era. This paper focus on two thing first is of provide preserving for outsourced data for which a perfect perturbation and de-perturbation technique is develop and other is of the multilevel data sharing out of which data cannot be deperturb again. If the intruder have same algorithm or even same parameters so this common perturbation algorithm work in both the requirement. In future, more work need for the same field for cloud storage 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] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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

[11] 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.

[12] Nikunj H. Domadiya, Udai Pratap Rao, "Hiding Sensitive Association Rules to Maintain Privacy and Data Quality in Database", in IEEE Journal, 2012.

[13] 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.

[14] 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.

[15] 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.

[16] Kevin Chiew, Shaowen Qin, "Analysis of Privacy-Preserving Mechanisms for Outsourcing Data Mining Tasks", in IEEE Journal, 2008.

[17] 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

[18] D.Narmadha, G.NaveenSundar and S.Geetha,"A Novel Approach to Prune Mined Association Rules in Large Databases‖", In IEEE, 2011 pp. 409413.

[19] 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.

