ISSN 2277 – 5528
Impact Factor- 4.015

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & MANAGEMENT

PRIVACY PREVENTION OF TEXTUAL AND NUMERIC SENSITIVE INFORMATION BY K-ANONYMITY AND PERTURBATION TECHNIQUE

Priya Gupta, Sini Shibu, Prof. S.C.Kapoor
* M.Tech Research Scholar of Computer Science and Engineering Department, NRI-IST, Bhopal (M.P.) India
**Guide & HOD, Computer Science and Engineering Department, NRI-IST, Bhopal (M.P.) India
***Director, Computer Science and Engineering Department, NRI-IST, Bhopal (M.P.)

ABSTRACT

With the increase of digital data on servers different approach of data mining is done. This lead to important issue of proving privacy to the unfair information against any person, place, community etc. So Privacy preserving mining come in existence. This paper provide privacy for sensitive rule that discriminate data on the basis of frequency. So finding of those rules and suppression is done. Perturbation technique is use for the hiding sensitive rules. Experiment is done on real adult dataset for different ratio. Results shows that proposed work is better in maintaining the Perturbation Percentage, Individual Privacy at last suppress rules while other rules are remain unaffected.

Keywords- Data mining, Data Perturbation, Multiparty Privacy Preserving.

INTRODUCTION

As the number of digital data users are increasing day by day, so extracting information from this rough data is done by data mining. Different approach of mining is done for different type of data such as textual, image, video, etc. Information extraction is done in digital for resolving many issues. But some time this data contain information that is not fruitful for an organization, country, raise, etc. So before extraction such kind of information is remove. By doing this privacy for such unfair information is done. This is very useful for the security of data which contain some kind of medical information about the individual, financial information of family or any class. As this make some changes on the dataset, so present information in the dataset get modify and make it general for all class or rearrange so that miner not reach to concern person.

So privacy preserving mining consist of many approaches for preserving the information at various level form the individual to the class of items [3, 4]. But vision is to find the information from the dataset by observing repeated pattern present in the fields or data which can provide information of the individual, then perturb it by different methods such as suppression, association rules, swapping, etc.

Mostly when data is place on the server then miner can get the access of the whole information, so many researchers are working for the access of the data. If data is successfully achieved then it is possible for miner to get all kind of information present in it. Considering this problem people are working for providing security against large number of privacy attacks. Here before placing the data on the public server it get perturb so that unfavourable information or negative data is suppress. This lead to put same data with some modification on the server and it will not affect the overall privacy [5]. So it is hard to require that protection of data is done in prior steps by hiding important information like name of person, address, mobile number, date of birth, etc. But this kind of protection is not sufficient for many cases where data mining algorithm is apply as it directly or indirectly fetch information from the raw data. Although utilization of same for the ethical purpose is very helpful in all the data privacy measure. So data mining implies on data where terrorism activity can be involve.

RELATED WORK

In [1] perturbation of dataset is done for providing security of the data on server. As some of cooperative store data is store on server for regular updating in price, category, etc. Dataset need protection from unauthorized user. So proper solution for this problem is develop in this paper by perturbing the data before uploading it on server. Then proper algorithm is develop for the de-perturbing the uploaded perturbed copy as if authorized user again read data then it should get original copy. Here by the use of association rule sensitive information or pattern of items is obtained. Now those rule which are above the threshold of minimum support are perturbed by adding fake transaction in the dataset so that overall support get reduce and dataset get perturb by these fake transaction. Placement of these transactions is done by modulus table. As this modulus remember the fake position in the dataset. In order to increase perturbation Items are replace by chipper text where each text will specify one item in the original dataset. In [14] similar work for outsourcing is done but the algorithm is calculation is unknown to the client and server.

In [6] k-anonymity technique is use as it give direct protection for the individual before releasing the data. This can be understand as let a person having salary then

that is replace by the range of salary from ten thousand to twenty thousand. In the similar fasion age of a person is replace by range. So by this overall confusion of the data is increase while rest of value remain same. So they simply give range to the age, income. Let age = 24 then its range is 20-30. Then this paper find hidden information from the data with the help of Association Frequent rules. As for finding the pattern of purchasing of item from the transaction frequent pattern need to be generate with the help of association rules.
In [8] multilevel privacy is provide by the author, basic concept develop in this paper is separate perturbed copy of the dataset for different user. Here user are divide into there trust level so base on the trust level dataset is perturbation percentage get increase. Here paper resolve one issue of database reconstruction by combing the different level perturbed copy then regenerate into single original database. So to overcome this problem perturbation of next level is done in perturbed copy of previous one. In this way if lower trust user get combine and try to regenerate original dataset then only one higher perturbed copy can be regenerate. The distribution of the entries in such a matrix looks like corner-waves originated from the lower right corner.

In [9, 12] paper cover a new issue for the direct indirect discrimination prevention in the dataset. Here it will collect discriminate item set which help in producing the association rule for identifying the direct or indirect rules. Then hide the rules which are above the threshold value by converting the X→Y to X→Y' where X is a set of discriminating item this tend to hide the information which will generate only those rules that not give any discriminating rule. Here Y is change to Y' means an opposite value is replace at few attributes.

In case of Pre-processing there are methods that can identify those rules or attributes in the database that is obtained from the source data then remove, modify those discriminatory rules or attributes biases contained in the original data so that no unfair decision rule can be mined from the transformed dataset by using any of the data mining algorithms. The pre-processing approaches of data transformation and hierarchy-based generalization can be adapted from the privacy preservation literature [5, 11].

One more category of discrimination prevention is In-processing approach where privacy prevention rules are apply in the algorithm which generate information. This can be understand as some non-discretionary constraints are apply on the decision tree of [10] so that generated information is discriminant free. Although it is found that in-processing discrimination prevention algorithms are depends on the special purpose data mining approaches as standard data mining algorithms cannot be used because they ought to be adapted to satisfy the non-discrimination requirement.

PROPOSED WORK

4.2Original Dataset:
This is the collection of the transaction which has all original values. Now as the original dataset contain most of the transaction that give fruitful analysis of the data. In order to reduce this mining of the data some perturbation can be add to it. This can be understand by an example that let an dataset has {items, name, gender, age, salary} then by generating association rules from this one can easily generate the pattern of the customers, these frequent rules are useful for any competitor. So hiding of these rules is done by perturbing the dataset transaction.

ISSN 2277 – 5528
Impact Factor- 4.015

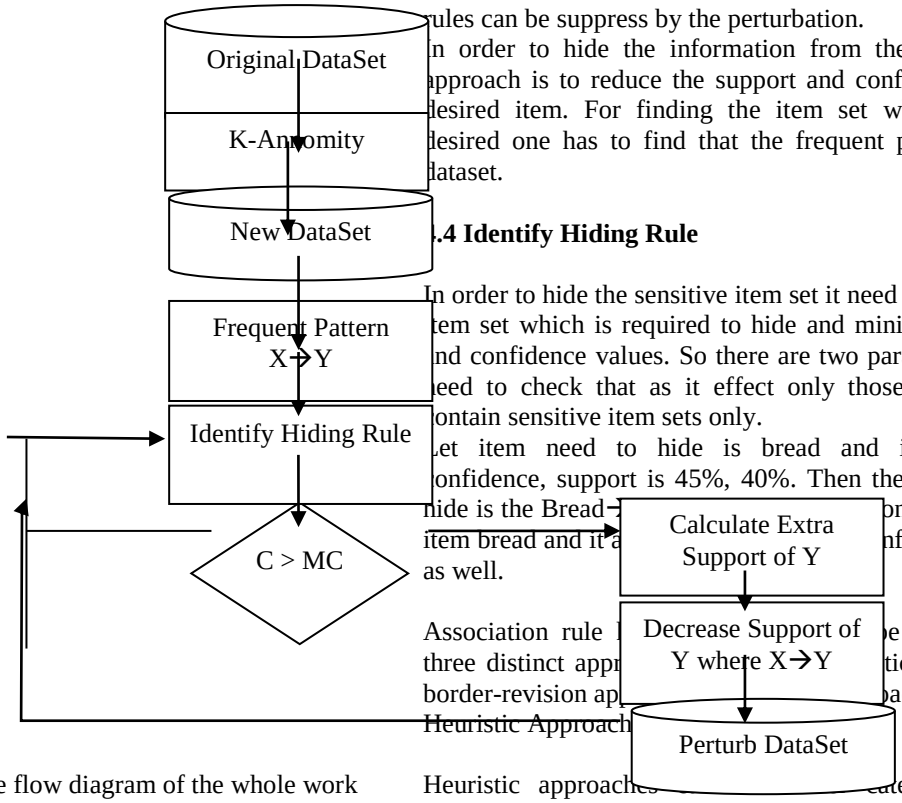

Fig. 4.1 Represent the flow diagram of the whole work

K Anonymity: As 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. Such as for the following original set one can increase the confusion by K-Anonymity where k = 2,3,…n.

4.3 Frequent Pattern:

In this step hidden information present in the dataset in form of sensitive rules are extract then suppress those information in the dataset using perturbation. Here by using Aprior Algorithm association rules are generate where pattern of each textual attribute is analyzed as the item. This can be understood as the T1 and T2 are two textual values of different attributes. Now find support of this pattern or rule in the dataset. Let T1→T2 is present in 300 sessions and dataset contain 1000 session where T2 is present then confidence is 0.3 or 30 percent. Which is calculate by

$$Support = \frac{T1 \rightarrow T2}{D}$$

So those rules which cross minimum threshold value is consider as the sensitive rules. Now sensitivity of those rules can be suppress by the perturbation.

In order to hide the information from the dataset one approach is to reduce the support and confidence of the desired item. For finding the item set which is most desired one has to find that the frequent pattern in the dataset.

.4 Identify Hiding Rule

In order to hide the sensitive item set it need to specify the tem set which is required to hide and minimum support nd confidence values. So there are two parameter which eed to check that as it effect only those rules which ontain sensitive item sets only.

Let item need to hide is bread and its minimum onfidence, support is 45%, 40%. Then the rule need to hide is the Bread→ ontain sensitive item bread and it a nfidence values as well.

Association rule e divided into three distinct appr tic approaches, border-revision ap paches. Heuristic Approach

Heuristic approaches categorized into distortion based schemes and blocking based schemes. To hide sensitive item sets, distortion based scheme changes certain items in selected transactions from present to absent and vice versa. Blocking based scheme replaces certain items in selected transactions with unknowns. These approaches have been getting focus of attention for majority of the researchers due to their efficiency, scalability and quick responses.

4.4.1 Hide Sensitive Rule:

So in order to hide an association rule, X → Y, we can either decrease its support or its confidence to be smaller than user-specified minimum support transaction (MST) and minimum confidence transaction (MCT). To decrease the confidence of a rule, there is two approach:

(1) Increase the support of X, the left hand side of the rule, but not support of X → Y.

(2) Decrease the support of the item set X →Y .For the second case, if we only decrease the support of Y, the right hand side of the rule, it would reduce the confidence faster than simply reducing the support of X → Y.

Here it only reduce the RHS item Y of the rule correspondingly. So for the rule Bread→ Milk can generate reduce the support of Y only. Now it need to find that for how many transaction this need to be done. So calculation of that number is done by

((Rule_confidence – Minimum_confidence) * X_support* Total_transaction)
Above formula specify the number of transaction where one can modify and overall support of that hiding rule is lower then the minimum confidence.
Proposed Algorithm:

For this algorithm t is a transaction, T is a set of transactions, R is used for rule, RHS (R) is Right Hand Side of rule R, LHS (R) is the left hand side of the rule R, Confidence (R) is the confidence of the rule R, a set of items H to be hidden.

4.5 K Anonymity Algorithm:

Input: Dataset D, K value {2, 3, …n}. Sensitive attribute A
OutPut: D with K-anonymity

1. Loop I = T is not empty
2. t ← T[I]
3. Loop J = I+1 is not empty
4. t' ← T[J]
5. If Equals (t[A], t'[A])
6. Count ← count +1
7. Mark(t')
8. EndIf
9. EndLoop
10. While count< K
11. T←Generate_transaction(S, t)
12. EndWhile
13. EndLoop

Hiding Rules:

Input: A source database D, A minimum support in_support (MST), a minimum confidence min_confidence (MCT), a set of hidden items X.

Output: The sanitized database D, where rules containing X on Left Hand Side (LHS) or Right Hand Side (RHS) will be hidden.

- **Steps of algorithm:**

1. R[c,s] ← Aprior(D, X) // c= confidence & s = support
2. Loop I = For each rule R
3. If Intersect(R[I], H) and R[I] > MCT
4. New_transaction ← Find_transaction(R[I], MCT)
5. While (T is not empty OR count = New_treansaction)
6. If t←T have XUY rule then
7. Remove Y from this transaction
8. End While
9. EndIf
10. End Loop

EXPERIMENT AND RESULT

This section present the experimental dataset and different evaluation parameter description. Here Results are shown and comparison of those result is also done.

a. Dataset

In [15] it has use Adult dataset where it contain different discriminating item set such as country, Gender, Race, 1996. This data set consists of 32,561. The data set has 14 attributes (without class attribute).

b. Evaluation Parameters

5.3 Evaluation Parameters

In order to compare our work one of the previous algorithm from [10] is utilize in which it increase the support of X and decrease the support of Y, for the confidence=(XUY)/X. In order to evaluate this work following are the few parameters of evaluation:

Lost Rules: Representing the number of non-sensitive patterns (i.e., association, classification rules) which are hidden as side-effect of the hiding process

False Rules: Representing the number of art factual patterns created by the adopted privacy preserving technique.

Missed Rule: Representing the number of Sensitive patterns still present in dataset even after applying adopted privacy preserving technique.

Privacy Percentage: This specify the percentage of the privacy provide by the adopting technique.

5.4 Results

ISSN 2277 – 5528
Impact Factor- 4.015

Support	Lost Rules Percentage	
	Previous work	Proposed Work
1	0	0
2	0	0
3	0	0
4	0	0
5	0	0

Table. 1. Represent comparison of proposed and previous work on the basis of Lost Rules.

From table 1 it is obtained that proposed work has not affect non sensitive rules in the dataset. While previous work do not apply any approach for rule preservation so no affect on those rules are present after previous approach.

Support	False Rules Percentage	
	Previous work	Proposed Work
1	0	0
2	0	0
3	0	0
4	0	0
5	0	0

Table. 2. Represent comparison of proposed and previous work on the basis of False Rules.

From table 2 it is obtained that proposed work has not generate any sensitive as well non sensitive rules in the dataset. While previous work do not apply any approach for rule preservation so no affect on those rules are present after previous approach.

Support	Missed Rules Percentage	
	Previous work	Proposed Work
1	100	0
2	100	0
3	100	0
4	100	0
5	100	0

Table. 3. Represent comparison of proposed and previous work on the basis of Missed Rules.

From table 3 it is obtained that proposed work has not preserve all sensitive rules in the dataset. While previous work do not apply any approach for rule preservation so no affect on those rules are present after previous approach. Here all sensitive information is hide in proposed work.

Support	Privacy Percentage	
	Previous work	Proposed Work
35	71.4286	93.7812
30	71.4286	91.2218
25	71.4286	87.9724
20	71.4286	87.0726

Table. 4. Represent comparison of proposed and previous work on the basis of privacy percentage.

From table it is obtained that with increase in support value number of rules crossing minimum support get decrease. This lead to one more evaluation as number of rule get decease so percentage of originality get increase.

CONCLUSION

As data mining provide makes work easy for different organization. Preserving privacy mining of discriminate rules as well as numerical values is done in this paper. Proposed work has generated rules by aprior algorithm where those which are above the support threshold are consider as sensitive rules. For perturbing those rules sensitive item is suppressed by adopting perturbation. Results shows that proposed work perform well in different evaluation parameter as compare to previous works. As research is the continuous process where, so in future different rule generation algorithm can be use which automatically identify sensitive items.

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]. C. Tai, P. S. Yu, and M. Chen, "K-support anonymity based on pseudo taxonomy for outsourcing of frequent itemset mining," in Proc. Int. Knowledge Discovery Data Mining, 2010, pp. 473–482.

[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. Srikant, "Privacy-preserving data mining," in Proc.ACM SIGMOD Int. Conf. Manage. Data, 2000, pp. 439–450.

[6]. M.Mahendran, 2Dr.R.Sugumar "An Efficient Algorithm for Privacy Preserving Data Mining Using Heuristic Approach" International Journal of Advanced Research in Computer and Communication Engineering. Vol. 1, Issue 9, November 2012

[7] 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.

[8] Yaping Li, Minghua Chen, Qiwei Li, and Wei Zhang. "Enabling Multilevel Trust in Privacy Preserving Data Mining" IEEE transaction on knowledge data engineering, VOL. 24, NO. 9, SEPTEMBER 2012.

[9]. Sara Hajian and Josep Domingo-Ferrer. "A Methodology for Direct and Indirect Discrimination Prevention in Data Mining". IEEE transaction on knowledge data engineering, VOL. 25, NO. 7, JULY 2013.

[10]. Mohamed R. Fouad, Khaled Elbassioni, and Elisa Bertino. A Supermodularity-Based Differential Privacy Preserving Algorithm for Data Anonymization. IEEE transaction on knowledge data engineering VOL. 26, NO. 7, JULY 2014

[11]. F. Kamiran, T. Calders, and M. Pechenizkiy, "Discrimination Aware Decision Tree Learning," Proc. IEEE Int'l Conf. Data Mining (ICDM '10), pp. 869-874, 2010.

[12]. D. Pedreschi, S. Ruggieri, and F. Turini, "Discrimination-Aware Data Mining," Proc. 14th ACM Int'l Conf. Knowledge Discovery and Data Mining (KDD '08), pp. 560-568, 2008.

[13]. D. Pedreschi, S. Ruggieri, and F. Turini, "Measuring Discrimination in Socially-Sensitive Decision Records," Proc. Ninth SIAMData Mining Conf. (SDM '09), pp. 581-592, 2009.

[14]. Yogachandran Rahulamathavan, Raphael C.-W. Phan, Suresh Veluru, Kanapathippillai Cumanan and Muttukrishnan Rajarajan."Privacy-Preserving Multi-Class Support Vector Machine for Outsourcing the Data Classification in Cloud ".IEEE IEEE transaction on dependable and secure computing, VOL. 11, NO. 5, September 2014.

[15]. R. Kohavi and B. Becker, "UCI Repository of Machine Learning Databases," http://archive.ics.uci.edu/ml/datasets/Adult, 1996.

[16]. Privacy-Preserving Data Publishing for Multiple Numerical Sensitive AttributesQinghai Liu, Hong Shen_, and Yingpeng Sang. TSINGHUA SCIENCE AND TECHNOLOGY Volume 20, Number 3, June 2015.