C Privacy Prevention of Discriminating Rules by Perturbing Sensitive Items

Deepak Patel
M. Tech. Scholar
Dept. of CSE
L.N.C.T. Bhopal

Vineet Richhariya, PhD
H.O.D. CSE
Dept. of CSE
L.N.C.T. Bhopal

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 community, gender, country, etc. 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 originality, reduce execution time, reduce data loss, at last suppress rules while other rules are remain unaffected.

General Terms

Privacy Preserving Mining

Keywords

Keywords are your own designated keywords which can be used for easy location of the manuscript using any search engines.

1. 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.

2. 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.

3. BACKGROUND

The data set is a combination of items and their attribute. Let the original dataset has an item attribute along with its value, e.g. Gender = {Male, Female}. So a combination of item is term as set of item such as e.g. {country, employee, salary}. Here discrimination rules is express as X→ C, where C has value of binary class such as true or false. Other is X item set such as {foreign, worker → False}.

Support(s) can be define as total percentage of any rule present in dataset. Such as rule X→C percentage in dataset is obtain by

Support (X→C) =(XUC) / D

Where D is total number of session present in dataset while XUC is number of session where X→C is present.

In similar fashion confidence of association rule is obtained by

Confidence (X→C) =(XUC) / X

Where X is total number of session where X present in dataset while XUC is number of session where X→C is present.

Elift: Pedreschi et al. [12] generate the new method of evaluation of rules from the measure that is elift which is the ratio of the confidence of the rule to the confidence of the non discriminatory items in that rule.

Elfit = Conf(AUB→C) / Conf(B→C)

4. PROPOSED WORK

4.1 Pre-Processing

As the dataset is obtain from the above steps contain many unnecessary information which one need to be removed for making proper operation. Here data need to be read as per the algorithm such as the arrangement of the data in form of matrix is required.

4.2 Generate Rules

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. There are many approaches of pattern finding in the dataset which are most frequent one of the most popular is aprior algorithm is use in this work.

4.3 Separate Direct and Indirect Rules

Now from the generate rule step one can get bunch of rules then it is required to separate those rules from the collection into direct and indirect rule set. Those rules which contain dicriminant items are identified as the direct rules which those not contain are indirect rules. This can be understand as the Let A, B →C where A is set of discriminant item then this rule is direct rule, where B, C are non discriminant items. If D, B→ C is a rule and D is the non discriminate item set the this rule is not direct rule.

Now all direct rules are need to find that either it is ά discriminate rule or not for this first find the elift value of the rule then those rules whose elift value are more than the ά value is term as the ά discriminatory rule.

4.4 Perturb Transaction

Now next step is to calculate number of transaction that need to be modified in order to hide the rules. For this follow steps:

$$Session = \frac{ABC \times B - \alpha \times BC \times AB}{B - \alpha \times BC}$$

Where ABC, AB, BC, B is number of session those items are present in dataset.

4.5 Perturbation

In order to hide that rule many approaches has been done that is mention in the table below

Table 1. Rule protection different approach

	Original Rule	Perturb Rule
Previous	A, B→C	A, B→C'
Proposed	A, B→C	A', B→C

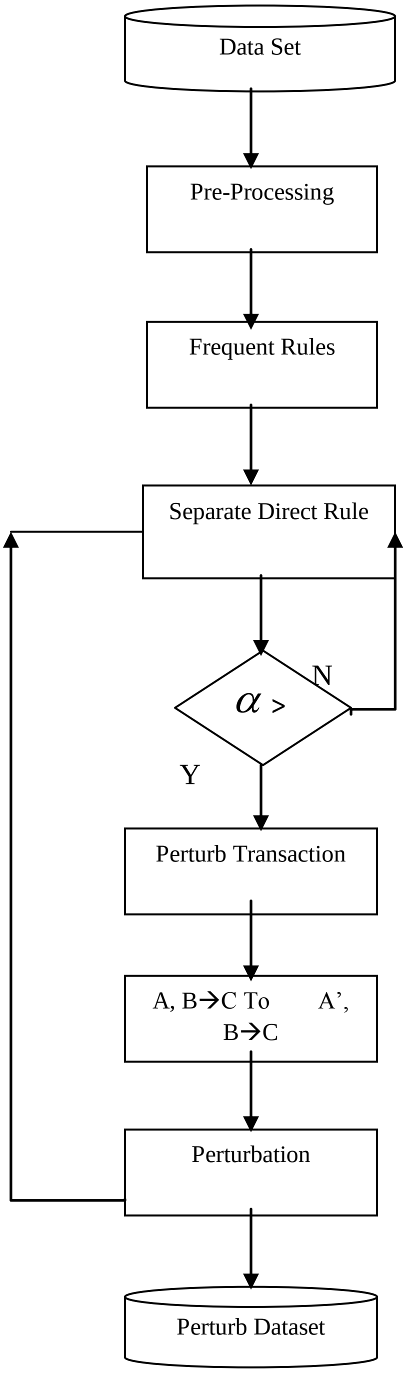

Fig 1: Represent Block diagram of proposed work

By change in the discriminate item directly from A to A' where A' is the opposite of A. Another advantage of this is it is not required to suppress the indirect rule separately as the indirect rule which is obtain from the combination of the discriminating items. So by suppressing the discriminating item only all kind of rule get hide.

4.6 Proposed Algorithm

Input: Org_DS (Original Dataset copy), α (elift threshold)

Output: Pert_DS (Perturbed Dataset copy)

```
1. DS ← Pre-Process(Org_DS)
2. Pert_DS = DS
3. FR[n] ← Aprior(DS) // n number Association rule
4. Loop 1:n
5. If FR[n]∩DI
6. DR[c]←FR[n] // c number of direct rules
7. Endif
8. End Loop
9. Loop 1:c
10. E←Elift(DR[c])
11. If E > $\alpha$
12. m = Perturb_Transaction(DR[c])
13. Loop 1:m
14. Pert_DS ←Perturbation(Org_DS [m])
15. EndLoop
16. EndIf
17. End Loop
```

In above proposed algorithm input is original dataset and output contain perturbed dataset. In whole algorithm first rules are generate which are above elift threshold value α . Then in-order to suppress those rules find number of sessions to perturb and perturb those session where that item set is present.

5. 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.

5.1 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 48,842 records, split into a "train" part with 32,561 records and a "test" part with 16,281 records. The data set has 14 attributes (without class attribute). For our experiments with the Adult data set, we set DI = {Gender=Female} and salary greater then the 50k$.

5.2 Evaluation Parameters

There are two approaches to evaluate the discriminating algorithm developed which can specify the quality of the work first is Discrimination Removal while second is data quality after the implementation of the algorithm. Normally balancing both is quit difficult as if data quality need to maintain then some of the rules will be unaffected and over all purpose will be not be solve while in case of maintaining discriminating rule less data [11], dataset the quality will definite degrade as it need to either change or remove from the dataset.

- **Direct Discrimination Prevention Degree (DDPD).** This measure quantifies the percentage of discriminatory rules that are no longer discriminatory in the transformed dataset [9].
- **Direct Discrimination Protection Preservation (DDPP).** This measure quantifies the percentage of the protective rules in the original dataset that remain protective in the transformed dataset [9].

- **Data Loss**: As proposed work provide privacy for the sensitive item set rules with minimum data loss. As in privacy data perturbation make data loss.
- **Originality**: As change in original data is the way to provide privacy in mining. So algorithm that will maintain maximum originality after perturbation is major expectation.
- **Execution time**: Third parameter is to evaluate execution time time of the algorithm that is time taken by the proposed method for execution. Algorithm time is expect after the evaluation of the direct and indirect rules.

5.3 Results

Table 2. Represent data Originality percentage at different α values

Elift Threshold (α)	Originality Percentage	
	Proposed Work	DIDP [9]
1.2	99.9803	99.9785
1.1	99.9741	99.9719
1	99.9669	99.9581

Table 3. Represent data Data Loss percentage at different α values

Elift Threshold (α)	Data Loss Percentage	
	Proposed Work	DIDP [9]
1.2	0.0197	0.0215
1.1	0.0259	0.0291
1	0.0331	0.0419

From table 2 it is obtained that with the decrease in ά value originality of the perturbed dataset also decrease. This is because as the decrease of ά will increase the number of session of those rules. One more observation is that proposed work originality is high as compare to previous work in [9].

From table 3 it is obtained that with the decrease in ά value data loss of the perturbed dataset get increase. This is because as the decrease of ά will increase the number of session of those rules and to hide those new session algorithm need to perturb more sessions. One more observation is that proposed work data loss percentage is lower as compare to previous work in [9].

Table 4. Represent data Execution time in Second at different α values

Elift Threshold (α)	Execution Time in Sec.	
	Proposed Work	DIDP [9]
1.2	5.9998	109.1003
1.1	15.6369	118.5923
1	22.4561	132.2482

Table 5. Represent data DDPP and DDPD percentage of both the algorithm

Elift Threshold (α)	Proposed Work & DIDP [9]	
	DDPP	DDPD
1.2	0	100
1.1	0	100
1	0	100

From table 4 it is obtained that with the decrease in ά value execution time for perturbing the dataset get increase. This is because as the decrease of ά will increase the number of session of those rules and to hide those new session algorithm need to perturb more sessions, so extra time is required for hiding those rules. One more observation is that proposed work data execution time is lower as compare to previous work in [9].

From table 5 it is obtained that with the decrease in ά value DDPP and DDPD values remain same. As per DDPP it shows that non-sensitive rules remain unaffected by the work. While DDPD shows that all the sensitive rules are hide successfully. So as per previous work our algorithm also have same level of accuracy on these parameters.

6. CONCLUSION

As data mining provide makes work easy for different organization. Preserving privacy mining of discriminate rules is done in this paper. Proposed work has generated rules by aprior algorithm where those which are above the elift threshold are consider as sensitive rules. For perturbing those rules sensitive item is suppressed by adopting perturbation where sensitive item value is convert into non-sensitive value of same category. 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.

7. REFERENCES

[1] Giannotti, Laks V. S. Lakshmanan, Anna Monreale, Dino Pedreschi, and Hui (Wendy) Wang, *In IEEE* Systems Journal, VOL. 7, NO. 3, SEPTEMBER 2013, pp. 385-395. “Privacy-Preserving Mining of Association Rules from Outsourced Transaction Databases*”*

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

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

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

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

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

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

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

[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. IEEE transaction on knowledge data engineering VOL. 26, NO. 7, JULY 2014. A Supermodularity-Based Differential "Privacy Preserving Algorithm for Data Anonymization".

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

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

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

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

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