Privacy Preserving of Outsourced Information by Eliminating Sensitive Items

Asst. Prof. Ravindra Tiwari[1], Dr. Priti Maheshwary[2]

Abstract— Data mining is really the procedure or strategy of finding designs in enormous databases This fundamental objective of the information mining process leads to issue of finding sensitive or secured items which harm individuals, organization, etc. Here set of rules are perturbed by finding relevant item session and eliminate sensitive item only. This type of elimination reduce the risk of the dataset as compared to other methods such as robfrugal. Analysis was done on genuine dataset. Results demonstrates that proposed work is better as contrast with different past methodologies on the premise of assessment parameters.

Keywords: Distributed Data, Data Mining, Encryption, Effective Pruning, Super class substitution.

I. Introduction

The data mining can be used in the cases of privacy, legality, and ethics. For the purpose of public sector security, national security and legal matters security there is Total Information Awareness Program that is ADVISE, has worked for privacy concerns. This paper need to do data preparation to initiate data mining that uncovers information, data or patterns which may result in confidentiality and privacy obligations. The most common process is data aggregation, it involves the process of combining the data together ,obtained from variety of sources. But this is not actually data mining, is the result of the preparation of data previously – and further for analysis purpose. The fear of individual's privacy comes into the scene when the data is compiled and then, causes the data miner, or any other individual can haves access to the new compiled data set, and can identify specific person, especially when the data is basically unspecified.

Requirement of Privacy Preserving: This is the main problem that arises in any huge databases. But sometimes it is needed because of law such as medical databases or for other business interests. However, in some situations the data sharing can lead for a mutual benefit .The key utility of huge databases today is research, whether it can be scientific, economic and market based. In the medical field this work can gain by collecting data for research and even competitive businesses with mutual gains.

Multidimensional Dataset: The aim of this research use multidimensional data set of students that are admitted in M.P. for engineering or professional courses like pharmacy, MBA, MCA from all over the India. The dimensions that are used for student are location and time. Location dimension will also refine from its multiple level like city<state.

In current study work assume that domain of each dimension attribute is ordinal and can be divided in to equal size intervals. For instance time can be divided in to day<week<month<quarter<year… and location in to city<state<country… etc. Stream again refine in multi level hierarchy like for engineering it may be for civil, mechanical, computer science, electronics communication, electrical engineering , information technology and many more. From this multidimensional design, get that what is the support and confidence of those student that are admitted in any particular branch from any particular state in a year.

With a specific end goal to analyze proposed algorithm, it need the dataset. So school affirmation dataset is utilize that has following property {branch, gender, course, pin-code, etc.}. Here understudy data are pin-code, branch, gender while sensitive things are vital for the confirmation dataset proprietor. So for the privacy preserving the two things require cover up. So keeping in mind the end goal to give security against the private information of the client one idea of

supermodularity has been incorporate which make various duplicate of a similar client with various values. Then for hiding the useful or sensitive data transaction, in other words the most frequent item set association rules are searched and hided. This work can provide privacy to those datasets only which have the pattern generation values in the transactions. In this dataset it contain different sets. This data set consists of 20,000 records. The data set has 10 attributes (without class attribute).

II. RELATED WORK

T.Calders and S.Verwer [2] utilizes Naive Bayes approach for classification of large database. Here author classifies dataset on the basis of frequent sensitive item sets. Here discrimination is done on the basis of gender, race, etc. which is natural class of the people. So separation done on this basis is against law, which needs to be suppressing in the dataset. Although numeric values present in the datset is remain same as previous, so it required to be perturbed as it contain many sensitive relations.

F.Kamiran and T.Calders [3] present a new approach of classification of database on the basis of non discriminating item sets. So presence of discriminating item in dataset for classification is not required. Here direct removal of sensitive information is performing. This is possible by sampling in the dataset, here sampling make data free from discrimination. Here discriminating models are not taken for evaluation that no information is mined from operated data. But doing classification base on non discriminating items is ethical view.

D. Pedreschi, S. Ruggieri, and F. Turini [4] Proposed discrimination on the basis of social items or attributes for example majority, minority. So work try to remove decision laws obtain from the information. Here classification tool is use for resolving classification of sensitive information. So overall a good quality of discrimination prevention done by the author.

S. Hajian, J. Domingo-Ferrer, and A. Martinez-Balleste[6] proposed a antidiscrimination rule by adding or removing sensitive information in the data. This addition and deletion is data modification or transformation algorithm so prevention of discriminatory information is done. This paper has consider data quality analysis as well. Here propose work is not performing on real dataset and indirect information is still present in the data.

Faisal Kamiran, Toon Calders and Mykola Pechenizkiy [7] propose a decision making algorithm of data mining, for analyzing discrimination awareness. Here decision are generate by utilizing some of background information of the database. Although necessary measures are taken so that decision rules are not affected. Here two approach were develop first named as Dependency Aware Tree and other is Leaf Relabeling for decision making.

J. Lai et al. [11] proposed a protection saving outsourced association pattern mining arrangement. This arrangement is powerless against frequency examination attacks. Applying this answer for vertically apportioned databases will bring about the leakage of the correct backings to information proprietors.

T. Tassa [12] proposed for secure mining of association runs in on a level plane disseminated databases. The proposed convention depends on the quick conveyed calculation, which is an unsecured dispersed variant of Apriori calculation. The convention registers the union (or crossing point) of private subsets that each of the intriguing site hold. Likewise, the convention tests the incorporation of a component hold by one site in subset held by another. In any case, this arrangement is appropriate for level dividing, not for vertical apportioning.

III. Proposed Work

Whole work is a combination of two steps where first include site creation while second include distribution of columns on various sites. While transferring whole row emcryption was performed on the them to save on the sites. Explanation of whole work is shown in fig. 1.

Pre-Processing

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.

Pattern Generation

With the help of aprior algorithm proposed work has find all set of rules present in the dataset which are frequent. In this work support is the parameter which identifies the frequent rule presence in the dataset. So calculation of this value was done as:

Support: Support of an association rule is characterized like the rate/portion of information that comprise of A U B to the aggregate number of records in the database, Let D be the measure of datasets or records in the database.

Support (A➔B) = (AUB) / D

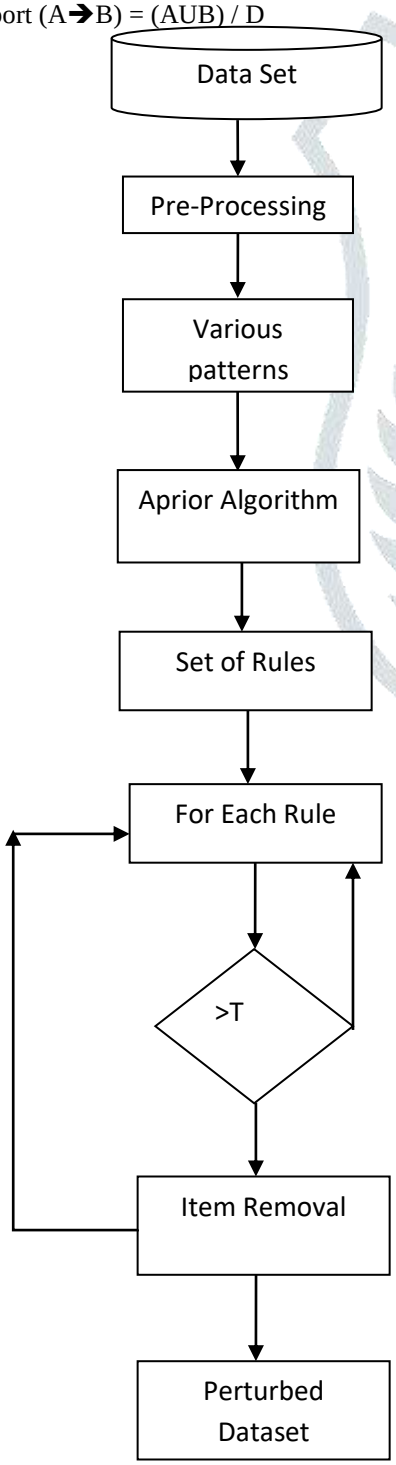

Fig. 1 Block diagram of proposed work rule generation.

Filter Sensitive Rule

Now from the generated rule one can get bunch of rules then it is required to separate those rules from the collection into sensitive and non- sensitive rule set. Those rules which cross sensitive threshold are identified as the sensitive rules while those not containing are indirect rules. This can be understood as the Let A, B →C where this pattern cross minimum threshold value so this rule is sensitive rule. If D, B→ C is a rule and not cross sensitive or minimum threshold then this rule is not sensitive rule.

Sensitive Pattern Hiding:

In this method sensitive item in the dataset get removed by just replacing blank in the cell of the row. Here this reduces whole risk while utilization of the dataset also get highly reduced. In this work suppression of single row elements are removed while those session who have achieved two level of anonymity are remain same so algorithm not required to increase the dataset size.

So in order to hide pattern, {X, Y), this work can decrease its support to be lesser than user-provided minimum support transaction (MST). In order to decrease the support value the approach is to lessen the support of the item set {X, Y}.

((Rule_support – Minimum_support) * Total_transaction)/100**Input:** A source database D, A minimum support in Transaction (MST).

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. P[c] ← Pattern_generation(D) // s = support

2. Loop I = For each P

3. If Intersect(P[I], H) and P[I] > MST

4. New_transaction ← Find_transaction(P[I], MST)

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

IV. EXPERIMENT AND RESULT

Dataset

In order to analyze proposed algorithm, it is in need of the dataset. So college admission dataset is use that has following attribute {branch, course, gender, pincode, etc.}. Here student information are pincode, gender, branch while sensitive items are important for the admission dataset owner. So for the privacy preservation both things need hide. So in order to provide protection against the private data of the customer one concept of supermodularity has been include which make multiple copy of the same student with different values.

Evaluation Parameters

Risk:

In this parameter the sum of information is done where highest subclass get higher value of risk. Each set of attribute have different set of subclass so risk of sharing information vary as per value pass in the perturbed dataset.

$$R = \frac{R(i, j)}{j}$$

Originality:

This specifies the percentage of the privacy provide by the adopting technique. Here total number of cells are count which are originally pass without any changes.

$$Originality = \frac{\sum Same\_cell}{Total\_cell}$$

Utility:

In this parameter the sum of information is done where highest subclass get higher value of utility. Each set of attribute have different set of subclass so utility of sharing information vary as per value pass in the perturbed dataset.

$$U = \log \frac{U(i, j)}{j}$$

Results

Table 1 Comparison of proposed and previous work on the basis of Originality percentage.

Dataset Percentage	Originality percentage	
	Robfrugal [5]	Item Removal
20	90.489	95.0795
30	91.3961	95.2197
40	92.2546	95.3748
50	93.0477	95.5498

Table 2 Comparison of proposed and previous work on the basis of execution time.

Dataset Percentage	Execution Time (second) Algorithm	
	Robfrugal	Item Removal
20	16.8897	1.86967
30	71.8729	3.1203
40	85.8229	1.3784
50	119.3401	3.9243

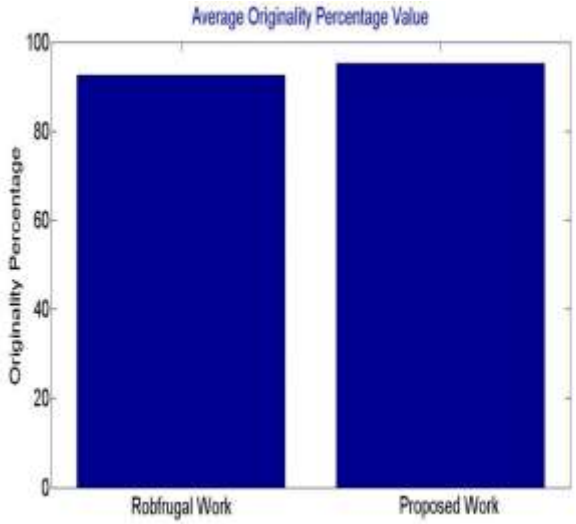

Figure 1 Comparison of average value of Originality percentage from various approaches.

From the above figure and table it is obtained that proposed work has highly maintain the originality of the dataset after applying the perturbation algorithm. Here by change in the dataset value originality of the previous work is less than proposed work as originality maintained always above around 95%. Here pattern preservation has less affect on dataset while previous approach was having higher affect.

Table 3 Comparison of proposed and previous work on the basis of Risk values.

Dataset Percentage	Risk Value	
	Robfrugal [5]	Item Removal
20	43728.125	39582.5
30	65603.125	59947.5
40	87478.125	80787.5
50	109353.125	101950

From table 2 it is obtained that the risk value of the dataset is reduced after applying the proposed work. In other words previous work has reduced the risk value but to less extent. It was obtained that Item removal have reduce risk as compare to previous but not that much as done by super class substitution algorithm proposed in this work. Here proposed work replace less informative data so risk of the outsourced dataset was quit less.

V. Conclusion

As scientists are chipping away at various field out of which finding a powerful vertical examples is measure issue with this becoming advanced world. This paper has proposed an information distribution algorithm for various servers. Here legitimate vertical columns are produce with the assistance of aprior algorithm and item removal method. By the utilization of item removal security of the information at server side get upgrade too. Results demonstrates that proposed work execution time get decrease. As research is never end handle so in future one can embrace other example era method for enhancing the server execution.

References

[1] 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, pp.1445-1459, JULY 2013.

[2] 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,pp,. 737-744 November 2012

[3] F. Kamiran and T. Calders, "Classification with no Discrimination by Preferential Sampling," Proc. 19th Machine Learning Conf.Belgium and The Netherlands, pp 1-6, 2010.

[4] European Commission, "EU Directive 2006/54/EC on Anti- Discrimination," http://eurlex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2006:204:0023:0036:en:PDF,0.

[5] D. Pedreschi, S. Ruggieri, and F. Turini, "Integrating Induction and Deduction for Finding Evidence of Discrimination," Proc. 12th ACM Int'l Conf. Artificial Intelligence and Law (ICAIL '09), pp. 157-166, 2009.

[6] S. Hajian, J. Domingo-Ferrer, and A. Martı´nez-Balleste´, "Rule Protection for Indirect Discrimination Prevention in Data Mining," Proc. Eighth Int'l Conf. Modeling Decisions for Artificial Intelligence (MDAI '11), pp. 211-222, 2011.

[7] 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.

[8] Fosca Giannotti, Laks V. S. Lakshmanan, Anna Monreale, Dino Pedreschi, and Hui (Wendy) Wang. "Privacy-Preserving Mining of Association Rules From Outsourced Transaction Databases". IEEE SYSTEMS JOURNAL, VOL. 7, NO. 3, SEPTEMBER 2013 385

[9] Sara Hajian, Josep Domingo-Ferrer and Antoni Martnez-Balleste Universitat Rovira Virgili. "Discrimination Prevention in Data Mining for Intrusion and Crime Detection" pp. 1-8, 2011 IEEE.

[10] 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, pp. 385-395, SEPTEMBER 2013.

[11] J. Lai, Y. Li, R. H. Deng, J. Weng, C. Guan, and Q. Yan, "Towards Semantically Secure Outsourcing of Association Rule Mining on Categorical Data," Inf. Sci.., vol. 267, pp. 267-286, May 2014.

[12] T. Tassa, "Secure Mining of Association Rules in Horizontally Distributed Databases Scalable Algorithms for Association Mining," IEEE Trans.Knowl. Data Eng., vol. 26, no. 4, Apr. 2014.