Multi Party Privacy Prevention of Sensitive Markov Patterns by K-Anonymity and Perturbation

Mukesh Kumar Dangi, Prof. S. R. Yadav

P.G. Scholar, Head P.G.

Computer science & Engineering, Millennium Institute of Technology Bhopal

Abstract—As data contain different kind of information for the increase of efficiency and understanding of system. So data mining provide approaches for the same. Here Markov order based patterns are evaluate from the dataset. But this analysis has lead to generate some rules which harm the privacy of the people so suppressing of those rules is highly require. For this K-Anonymity and data perturbing techniques is used. Then sensitive information's or patterns are perturbed for hiding those patterns. Experiment is done on real dataset and comparison is done with previous work. Result shows that proposed work has maintained same level of information in the dataset while preserve sensitive information as well.

Keywords- Aprior Algorithm, Association Rules, Data mining, Markov order, Privacy Preserving, Perturbation.

I. INTRODUCTION

Data mining methodology can help associating knowledge gaps in human understanding. Such as analysis of any student dataset gives a better student model yields better instruction, which leads to improved learning. More accurate skill diagnosis leads to better prediction of what a student knows which provides better assessment. Better assessment leads to more efficient learning overall. The main objectives of data mining in practice tend to be prediction and description [4, 5]. Predicting performance involves variables, IAT marks and assignment grades etc. in the student database to predict the unknown values. Data mining is the core process of knowledge discovery in databases. It is the process of extracting of useful patterns from the large database. In order to analyze large amount of information, the area of Knowledge Discovery in Databases (KDD) provides techniques by which the interesting patterns are extracted. Therefore, KDD utilizes methods at the cross point of machine learning, statistics and database systems.

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.

II. RELATED WORK

R.Agrawal and R.Srikant [1] utilizes ARM (Association Rule Mining) approach on large database. This paper present two algorithm based on association rule that discover relation between items. Although performance decreases with increase in database. One more point is that it does not consider item quantity information.

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 dataset remain same as previous, so it requires being perturbed as it contains 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.

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.

III. BACKGROUND

K- Anonymity: In the k-anonymity model, the quasi-identifier feature set consists of features in a table that potentially reveals private information, possibly by joining with other tables. In addition, the sensitive feature is a feature serves as the class label of each record. As shown in table. 1(b), the set of three features {Zip, Gender, Age} is the quasi-identifier feature set, while the feature {Diagnosis} is the sensitive feature. For each record in this table, its feature values in the quasi-identifier feature set are generalized as capsule feature values, while its value of sensitive feature are not generalized. Through generalization, an equivalence class is the set composed of records in the table which has the same values on all features in the quasi-identifier feature set.

Zip	Gender	Age	Diagnosis
47918	Male	35	Cancer
47906	Male	33	HIV+
47918	Male	36	Flu
47916	Female	39	Obesity
47907	Male	33	Cancer
47906	Female	33	Flu

Table. 1. Patient diagnosis records in a hospital

The 1st, 3rd and 4th records in table. 1(b) are assembled to form one equivalence class, while the 2nd, 5th and 6th records are assembled to form another equivalence class. The number of records in each equivalence class must be not less than k, which is called as the k-anonymity requirement. The value of k is specified by users according to the purpose of their applications. The records in table. 1(b) satisfy 3-anonymity requirement since the numbers of records in its two equivalence classes are both equal to three.

Zip	Gender	Age	Diagnosis
4791*	Person	[35-39]	Cancer
4790*	Person	[30-34]	HIV+
4791*	Person	[35-39]	Flu
4791*	Person	[35-39]	Obesity
4790*	Person	[30-34]	Cancer
4790*	Person	[30-34]	Flu

Table 2. The k-anonymity protected table when k= 3.

IV. PROPOSED WORK

a. 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.

b. K-Anonymity:

Here some specific data like age, salary, postal code, etc. are to be hidden which directly specify the user relation with the transaction. This is done by creating the range of particular values and replacing that value with that range, so that individual privacy of the user is also taken care of in this work. For generating the range, random function is used that generates number in fix range then replace original information with this range.

Input: DS (Pre-process Dataset)

Output: PDS (Perturb Dataset)

1. Loop 1:n // n number of rows in DS
2. Loop 1:m // m number of sensitive attributes
3. Range←Randi(m) // Randi : Gaussian random function
4. PDS(n,m)←Range
5. EndLoop
6. *Kth Order Markov Modal:*

Let D, be a set of database transactions where each transaction T is a set of items, called Tid. Let I= {I1, I2,..., Im} be a set of items. An item set contains k items is a k item set. If a k item set satisfies minimum support (Min_sup) then it is a frequent k item set, denoted by kth markov modal. Firstly markov modal generated a set of candidates, which is candidate k-item sets, denoted by Ck. If the candidate item set satisfies minimum support then it is frequent item pattern. In order to hide the sensitive patterns set it need to specify the pattern set which is required to hide and minimum support values. So there are

two parameter which need to check that as it effect only those rules which contain sensitive patterns sets only.

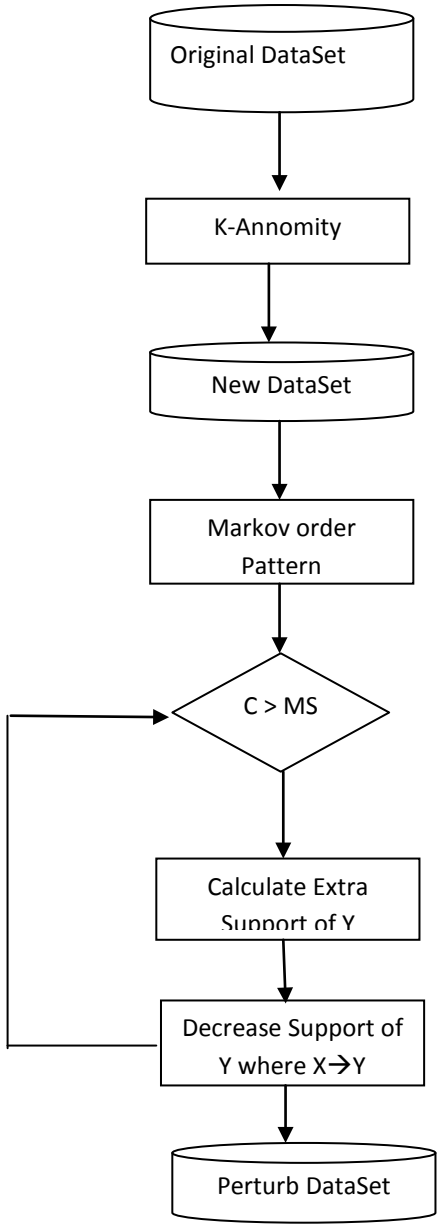

Fig. 1. Represent Block diagram of proposed work.

c. Hide Sensitive Pattern:

So in order to hide an pattern, {X, Y), it can decrease its support to be smaller than user-specified minimum support transaction (MST). To decrease the support of a rule, there is a approach: Decrease the support of the item set {X ,Y}. For this case, by only decrease the support of Y, the right hand side of the rule, it would reduce the support faster than simply reducing the support of {X , Y}.

Here it only reduce the RHS item Y of the pattern correspondingly. So for the pattern {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_support – Minimum_ support) * X_support* Total_transaction)

Above formula specify the number of transaction where one can modify and overall support of that hiding pattern is lower then the minimum support.

d. Multilevel Data Hierarchy

With the help of above calculation one can generate data copy for single party, by this the purpose of perturbation is not fulfill. As if one get few perturbed copy of the original dataset then producing of the original is not a big task. So distributing perturbed copy of single level perturbation is not sufficient.

So instead of doing single level perturbation, multilevel perturbation is more fruit full as different level copy is distribute to the different user of different trust. This can be understand as the data owner decide the priority of the user for distributing the perturbed data copy. Now steps to improve the perturbation of the original copy is simple Let original copy is X which is perturbed to Y.

$$Y = \text{perturb}(X)$$

Now for the lower level trust user new copy, is generate from the original data, then it will be not improve perturbation from the prior and the if higher level user can access the perturbed copy from the lower then chance of producing the original copy is more. So in order to reduce this probability of producing the original from the existing perturbed copy, perturbation for the new level is not generate from the original but it can be generate from the perturbed copy of the previous level.

V. 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 [9] Sara et. al. has used Adult dataset where it contain different discriminating item set such as country, Gender, Race, 1996. This data set consists of 48,842 records. The data set has 14 attributes (without class attribute).

b. Evaluation Parameters

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

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

Missed Pattern: 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.

c. Results

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

Table. 3. Represent comparison of proposed and previous work on the basis of Lost Patterns.

From table 3 it is obtained that proposed work has not affect non sensitive patterns in the dataset. While previous work do not apply any approach for pattern preservation so no affect on those patterns are present after previous approach.

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

Table. 4. Represent comparison of proposed and previous work on the basis of False Patterns.

From table 4 it is obtained that proposed work has not generate any sensitive as well non sensitive patterns in the dataset. While previous work do not apply any approach for pattern preservation so no affect on those patterns are present after previous approach.

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

Table. 5. Represent comparison of proposed and previous work on the basis of Missed Patterns.

From table 5 it is obtained that proposed work has not preserve all sensitive patterns in the dataset. While previous work do not apply any approach for pattern preservation so no affect on those patterns are present after previous approach. Here all sensitive information is hide in proposed work.

VI. CONCLUSION

In this work, a set of algorithms and techniques were proposed to solve privacy-preserving data mining problems. The experiments showed that the proposed algorithms perform well on large databases. It work better as the Maximum lost pattern percentage is zero a certain value of support. Then this work shows that false patterns value is zero. Comparison with the other algorithm it is obtained that including the K-Anonymity concept directly hide the sensitive information. It is shown in the results that accuracy of the perturbed dataset is preserved for low support values as well. Here Proposed work has resolve the multi party data distribution problem as well as different level trust party get different level of perturbed dataset copy.

REFERENCES

[1] R..Agrawal and R..Srikant, "Fast Algorithms for Mining Association Rules in Large Databases," *Proc. 20th Int'l Conf. Very Large Data Bases*, pp. 487-499, 1994.

[2] T. Calders and S. Verwer, "Three Naive Bayes Approaches for Discrimination-Free Classification," *Data Mining and Knowledge Discovery, vol. 21, no. 2*, pp. 277-292, 2010.

[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] Shibbir Ahmed, Rajshakhar Paul, and Abu Sayed Md. Latiful Hoque "Knowledge Discovery from Academic Data using Association Rule Mining". 2014 17th International Conference on Computer and Information Technology (ICCIT)

[5] N. Thai-Nghe (2011A), "Personalized Forecasting Student Performance", *Proceedings of the 11th IEEE International Conference on Advanced Learning Technologies*, Pp. 412–414.

[6]. 1M.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/, 1996.