A Survey on Privacy Preserving Mining and Limitations

Priya Ranjan [#1] , Raj Kumar Paul [#2]
Computer Science and Engineering,
Vedica Institute of Technology,
Bhopal, MP,India
[#1] **Priyamsamrat01@gmail.com ,** [#2] **rajkumar.rkp@gmail.com**

***Abstract*— Now a days, with the increase in the data mining algorithm knowledge extraction from the large data is getting easy. But at the same time this lead to new problem of Privacy of the knowledge from the stored data at various servers. So it is required to provide privacy of the sensitive data from the data miners. This paper focus on various approaches implement by the miners for preserving of information at individual level, class level, etc. A detail description with limitation of different techniques of privacy preserving is explained. This paper explains different evaluation parameters for the analysis of the preserved dataset. A classification of privacy preserving techniques is presented and major algorithms in each class is surveyed.**

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

I. INTRODUCTION

There are new research areas in field of Data mining and knowledge discovery in databases**,** that investigate the automatic extraction of previously unknown patterns from large collections of data**,** From the Internet and other media, without rapidly new information well be documented , it reached to a point where coercion against in the common privacy on a daily basis and it deserve serious thinking.

In data mining and statistical databases are Privacy preserving data mining, and is also a novel research

direction where it analyzed data mining algorithm for the side effect in data privacy. There are two folds in privacy preserving data mining. Like identifiers, gender, religion, addresses are first sensitive raw data. and these like are changed or cut out from the original database, In Second, sensitive data, this fold using in

database. In database at a mining time can be using data mining algorithm that is also excluded. because such knowledge can equally well compromise data privacy. The main aim of privacy preserving data mining that is changing the original data to develop algorithms. That by the private data and private knowledge remain private even after the mining process. When confidential information may be derived from released data That time problem is arises and "database inference" problem is call by unauthorized users.

The Presently invention of discrimination methods to regard every one rule individually for measuring discrimination Another rules consider without rule or relation ,But this paper is relation and rules for

discrimination discovery and that is based on existence or nonexistence of discriminatory attributes. Discrimination prevention, In data mining have another antidiscrimination aim, which one patterns are including also. Decisions do not front to discriminatory even if the original training data sets are biased. Three approaches are conceivable:

A. Pre-processing
B. In-processing
C. Post-processing

II. Privacy Preserving Techniques

Data Swapping In this techniques is data maintains as a order basically data e evolve as a textual form,text data perturbation as a textual data form .textual data means addition new values and may not possible in all cases of textual datasets. so swapping technology is better option for the same In which most frequent values are observed and replace with the least or lesser frequent values so that original values or decision cannot be taken from the perturbed copy of the dataset.

In some case if the replacement of the single item is done for the most frequent item then detection of that hide technique can be easily breakable. So it is necessary to choose the item from a set randomly for replacing the frequent one.

Suppression

In some data set have some information ,that information is directly identify by the individuals person or individual class then those has to remove from the data set. So colums or items are delete from the original data set ,the is such types of sensitive data set, Suppression is used for protecting for information ,As Example: We have data set contain a driving licence number,the only one person can detectable and we cannot add or delete in driving licence. as format of that driving licence number is define. So such data is removed from the original dataset.

Noise Addition

In this approach data set change as a change in a numeric value where amount is change is a sequence of random value, that value reflected as a original values but not in original data set order. In [5] noise is generate by a Gaussian function that create number as a sequence form then add there sequence in the original value. so a kind of variation is develop over here for the privacy of the original one. While data can add a single value but it can be detect easily or observed also if intruder will present in data set.

There are different numeric category involving as : involving percentiles, sums, conditional means etc. Some noise addition techniques, Random Perturbation Technique, Probabilistic Perturbation Technique , etc.

Data Perturbation

In data Perturbation on data set is transformed in to perturbation and selecting random position data then add, subtraction the value into the original in order produce new value that is differ from the previous data. One is important information is here whatever you want add or subtraction delete from that value should not cross the limits of the original lets understand an age value is perturbed by adding or subtracting from original data but the resultant value or the perturbed value should not be less then 0 or

greater then a normal life of 120. In order to perform perturbation some kinds of random value that by original value change randomly. There are generate two approaches.

First is probability distribution approach and Second is Value distortion approach

- probability distribution approach :-The approach of probability distribution,In this approach data replace with another sample from the same (estimated) distribution or by the distribution itself.
- Value distortion approach:- The approach of Value distribution, perturbed the value of data and elements or directly by adding or multiplicative some noise before releasing of the data.

III. Related work

This paper addresses [10] secure mining of association rules over horizontally partitioned data. The methods incorporate cryptographic techniques to minimize the information shared, while adding little overhead to the mining task. Privacy concerns may prevent the parties from directly sharing the data, and some types of information about the data. That allow parties to choose their desired level of security are needed, allowing efficient solutions that maintain the desired security.

Tzung Pei et al presented Evolutionary privacy preserving in data mining [4]. Collection of data, dissemination and mining from large datasets introduced threats to the privacy of the data. Some sensitive or private information about the individuals and businesses or organizations had to be masked before it is disclosed to users of data mining. An evolutionary privacy preserving data mining method was proposed to find about what transactions were to be hidden from a database. Based on the reference and sensitivity of the individuals data in the database different weights were assigned to the attributes of the individuals. The concept of pre large item sets was used to minimize the cost of rescanning the entire database and speed up the evaluation process of chromosomes. The proposed approach [4] was used to make a good tradeoff between privacy preserving and running time of the data mining algorithms.

This authors [3] presents a survey of different association rule mining techniques for market basket analysis, highlighting strengths of different association rule mining techniques. As well as challenging issues need to be addressed by an association rule mining technique. The results of this evaluation will help decision maker for making important decisions for association analysis.

Y-H Wu et al. [11] proposed method to reduce the side effects in sanitized database, which are produced by other approaches. They present a novel approach that strategically modifies a few transactions in the transaction database to decrease the supports or confidences of sensitive rules without producing the side effects.

A classification of privacy preserving techniques is presented and major algorithms in each class is surveyed. The merits and demerits of different

techniques were pointed out [2]. The algorithms for hiding sensitive association rules like privacy preserving rule mining using genetic algorithm.

Chung-Min Chen, [8] present dithered B-tree, a B-tree index structure that can serve as a building block for realizing efficient system implementations in the area of secure and private database outsourcing. The dithered tree insert algorithm [8] can be further optimized to incur

only one traversal from the root to the leaf, instead of two. The index structure from learning whether or not the search term (i.e., key) is present in the database and check the data for secure and private database outsourcing.

In Privacy Preserving Data Mining, data perturbation is a data security technique that adds 'noise' to databases to allow individual record confidentiality. This technique [9] allows users to ascertain key summary information about the data while preventing a security breach. Four bias types have been proposed which assess the effectiveness of such a technique. However, these biases deal with simple aggregate concepts (averages, etc.) found in the database. The author propose a fifth type of bias that may be added by perturbation techniques (Data mining Bias), and empirically test for its existence. In e-commerce applications, organizations are interested in applying data mining approaches to databases to discover additional knowledge about customers.

The author concept in this paper is Privacy Preserving mining of frequent patterns on encrypted outsourced Transaction Database (TDB) [1]. They proposed a encryption scheme and adding fake transaction in the original dataset. Their method proposed a strategy for incremental appends and dropping of old transaction batches and decrypt dataset. They also analyze the crack probability for transactions and patterns. The Encryption/Decryption (E/D) module encrypts the TDB once which is sent to the server. Mining is conducted repeatedly at the server side and decrypted every time by the E/D [1] module. Thus, we need to compare the decryption time with the time of directly executing a priori over the original database.

IV. 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 [6, 7], dataset the quality will definite degrade as it need to either change or remove from the dataset.

Sensitive Item Prevention Degree (SIPD): This measure quantifies the percentage of sensitive rules that are no longer discriminatory in the transformed dataset.Non Sensitive Item Protection Prevention Degree **(NSIPP)**. This measure quantifies the percentage of the

protective rules in the original dataset that remain protective in the transformed dataset. Since the above measures are used to evaluate the success of the proposed methods in direct and indirect discrimination prevention, ideally their value should be 100%.Data-Set Originality: As the privacy for the sensitive item is provide by hiding the sensitive item or replacing by other similar value but this lead to make dataset for perturbation. So work which maintain high data quality after prevention is better.

Execution time: Work need time for the effective result but algorithm that generate results in very sort duration of time then much better. So execution time is another evaluation parameter for the same.

Misses Cost (MC): This measure quantifies the percentage of rules among those extractable from the original dataset that cannot be extracted from the transformed dataset (side-effect of the transformation process). Ghost Cost (GC): This measure quantifies the percentage of the rules among those extractable from the transformed dataset that were not extractable from the original dataset (side-effect of the transformation process). MC and GC should ideally be 0%. However, MC and GC may not be 0% as a side-effect of the transformation process.

V. Conclusion

This paper addresses secure mining of association rules over horizontally partitioned data. The methods incorporate cryptographic techniques to minimize the information shared, while adding little overhead to the mining task. In Privacy Preserving Data Mining, data perturbation is a data security technique that adds 'noise' to databases to allow individual record confidentiality. Mining information from the data is the primary requirement of the data mining out of which privacy preserving mining is opening new emerging field which preserve knowledge from the data. Paper detailed various method like perturbing, swapping, etc. for privacy preserving, where each has its own importance. Researchers works find knowledge in dataset by Aprior and other mining algorithm then apply preserving technique on them. Hiding information at different level is also term as multi-level privacy which provide only numeric data hiding. While in few works both numeric and text data is hide but the time and space required for those algorithm is comparatively high. So a algorithm is still need to develop for the reduced time and space complexity without compromising time and space.

References

1. Pedreschi, D., Ruggieri, S. & Turini, F. (2008). Discrimination-aware data mining. Proc. of the 14th ACM International Conference on Knowledge Discovery and Data Mining (KDD 2008), pp. 560-568. ACM.

2. Hajian, S., Domingo-Ferrer, J. & Martinez-Ballesté, A. (2011a). Discrimination prevention in data mining for intrusion and crime detection. Proc. of the IEEE Symposium on Computational Intelligence in Cyber Security (CICS 2011), pp. 47-54. IEEE.

3. Verykios, V. & Gkoulalas-Divanis, A. (2008). A survey of association rule hiding methods for privacy. In C. C. Aggarwal and P. S. Yu (Eds.), Privacy- Preserving Data Mining: Models and Algorithms. Springer.

4. Meij, J. (2002) *Dealing with the data flood; mining data, text and multimedi*a, The Hague: STT Netherlands Study Centre for Technology Trends.

5. Calders, T., & Verwer, S. (2010). Three naive Bayes approaches for discrimination-free classification. Data Mining and Knowledge Discovery, 21(2):277-292.

6. Sara Hajian and Josep Domingo-Ferrer "A Methodology for Direct and Indirect Discrimination Prevention in Data Mining" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 7, JULY 2013

7. Pedreschi, D., Ruggieri, S. & Turini, F. (2009a). Measuring discrimination in socially-sensitive decision records. Proc. of the 9th SIAM Data Mining Conference (SDM 2009), pp. 581-592. SIAM

8. Hajian, S. & Domingo-Ferrer, J. (2012). A methodology for direct and indirect discrimination prevention in data mining. Manuscript.

9. C. Clifton. Privacy preserving data mining: How do we mine data when we aren't allowed to see it? *In Proc. of the ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD 2003)*, Tutorial, Washington, DC (USA), 2003.

10. D. Pedreschi, S. Ruggieri and F. Turini, "Discrimination-aware Data Mining," *Proc. 14th Conf. KDD* 2008, pp. 560-568. *ACM*, 2008.

11. D. Pedreschi, S. Ruggieri and F. Turini, "Measuring discrimination in socially-sensitive decision records," *SDM* 2009, pp. 581-592. *SIAM*, 2009.