International Journal of Scientific Research & Engineering Trends
Volume 4, Issue 5, Sept-Oct-2018, ISSN (Online): 2395-566X

# A Survey on Frequent Pattern Rules Techniques for Privacy Preserving mining

**M.Tech.Scholar Shivani Pandey**
Dept. of Computer Science Engg.
Adina Institute of Science and Technology
Sagar, MadhyaPradesh, India

**Asst.Prof. Monali Sahoo**
Dept. of Computer Science Engg.
Adina Institute of Science and Technology
Sagar, MadhyaPradesh, India

***Abstract-*** **Information sharing among the organizations is a general movement in a few zones like business advancement and showcasing. As portion of the sensitive rules that should be kept private might be revealed and such revelation of sensitive patterns may impacts the benefits of the organization that possess the information. Consequently the principles which are sensitive must be covered before sharing the information. In this paper to provide secure data sharing sensitive rules are perturbed first which was found various techniques are discussed. Here techniques of privacy data mining was also detailed.**

***Keyword s-*** **Distributed Data, Data Mining, Encryption, Effective Pruning, substitution etc.**

## I. INTRODUCTION

The requirement for information mining with security conservation has developed as an interest for trading sensitive data previously discharging information over the system. Additionally, the suspicious methodologies and refusal of the information providers towards the assurance of data. Internet Phishing is an ill-conceived approach to acquire private data, for example, usernames, passwords, and charge card points of interest by disguising as a dependable substance in an electronic correspondence. In this manner, expanded online assurance against phishing attacks is a region of colossal intrigue. As these attacks are advanced in nature, they represent a few difficulties as far as shirking techniques.

Internet phishing prompted a few security and financial strikes on the clients and undertakings around the world. Web payment gateways of internet banking have suffered and prompted generous money related misfortune [1, 2]. Consequently, enhanced information mining techniques with security are the need of great importance for secure data trade over the system.

These days, putting away clients' data has an obligation with the end goal that their security isn't damaged. Among a few existing calculation, the Data Mining with protection produces outstanding outcomes identified with the inside perception of privacy preserving with information mining. The security should be consolidated onto all mining components including clustering, association control, and order [1, 3].

Distributed computing enabled the business collaborators to store the information for the advantages of all partners. This has prompted gather clients' individual information and nourished into information mining plans which ought to guarantee that there is no loss of protection. Furthermore, the elements like usage, order of protection regarding its benefits and negative marks are not been audited legitimately. A few protection safeguarding plans in information mining exists which incorporate K-secrecy, cryptography, buildup, L-diversity, randomization, techniques [8, 9].

The PPDM strategies secure the information by concealing some unique data with the goal that private data isn't uncovered. The design is to adjust an exchange off amongst secrecy and productivity. The utilization cryptographic strategies dependably have computational expenses to avoid data spillage [4, 6].

## II. RELATED WORK

N. Muthu Lakshmi and K. Sandhya Rani [9] proposed a model to discover association rules for vertically divided databases considering the protection imperatives with 'n' number of sites alongside information data miner. This model compromises diverse cryptography strategies, for example, encryption, decoding and scalar item system to discover association runs productively and safely for vertically parceled databases.

F. Giannotti et al. [10] proposed an answer which depends on k-anonymity frequency. To counter frequency investigation intruder, the information proprietor embeds fake exchanges in the database to reduce the object frequency. Objects in the database are encoded with the 1-1 substitution words. In the wake of embeddings the fake exchanges, any object in the perturbed database will have a similar frequency with in any event $k - 1$ different objects. At that point dada proprietors outsource their database to the server for the mining assignment. The server runs visit item set mining calculation and returns the came about regular item sets

and their backings to the information proprietor. The information proprietor modifies these itemsets' backings by subtracting them with item sets' relating event check in the fake exchanges separately.

At that point, the information proprietor decodes the got item sets with the amended backings higher than the frequency limit and produces association rules in view of the incessant item sets. In these setting, information proprietor requires including item set events fake exchanges to counteract fake exchanges. Utilizing this strategy for the vertically parceled database, information proprietors can't perform such computations.

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.

Lichun Li et al. [13] proposed a security protecting association run digging answer for outsourced vertically divided databases. In such a situation, information proprietors wish to take in the association administers or regular itemsets from an aggregate informational index and unveil as meager data about their ( sensitive) crude information as conceivable to other information proprietors and outsiders. Symmetric homomorphic encryption procedure is utilized for calculation of help and certainty which guarantees the security of the information and mining result moreover.

## III.TECHNIQUES OF PRIVACY PRESERVING

**1.Apriori Algorithm** It is a basic technique for extracting frequent patterns by generating candidates. As the name implies, it requires the prior knowledge of frequent itemset properties. It is an incremental approach where frequent k-item set is used to generate frequent (k+1)-itemset. Initially, the database is scanned for finding count of all 1-itemsets. Then based upon the threshold value; frequent 1-itemsets are extracted. A cross join on the resultant is applied to get all possible 2-itemsets combinations. Again database is scanned for the counts of those itemsets and the process repeats until there is no new frequent itemset. To reduce the number of candidates, algorithm uses apriori property, also called downward closure property, says,"If an itemset is not frequent, its supersets will never be frequent". Hence, the algorithm works in two steps: joining(cross join is performed on k-itemsets to generate k+1 itemsets) and pruning(casting out infrequent itemsets based upon apriori property).The disadvantage of using this algorithm is that database is required to be scanned multiple time which increases the execution time. The generation of large number of candidates increases the space complexity1 .

**2.FP Growth-** It is a method that extracts frequent itemsets based on divide and conquer technique.FP-Growth works in two steps: Creating and Mining FP tree. While creating tree, the database items are scanned and arranged as a branch of tree in the descending order of their counts for each transaction. Items are marked along with these counts. Root is always NULL. If some sequence of a transaction isalready existing, then the remaining items are joined below it and the count of subset items is increased by one.

Tree is mined by constructing its conditional pattern which includes the paths to reach the node through root. A sub tree is constructed and patterns are generated by concatenating the item with its path. Search space is reduced due to the generation of conditional patterns. It gives good results for even long patterns. Since there is no need of candidate generation, space complexity is reduced .

**3.ECLAT-** It is an improvement of apriori algorithm. It uses vertical data format (item: transaction id set). It is similar to apriori, just the table is reversed. The item sets having count less than minimum support threshold will be eliminated. 2-itemsets will be generated by the intersection of transaction id sets of 1-itemsets. Cross join is performed to generate three item sets.

The 2-itemset subsets of 3-itemset are evaluated from previous table. From the downward closure property, 2-itemsets which are not frequent, their 3-itemset will also be infrequent. So, those 3-itemsets are casted out. Algorithm repeats till no new frequent itemset is generated. Due to this methodology, multiple scans of database are not required since transaction id set contains all the required information for counting supports. But length of TID-set requires large memory space. Computation time is also affected during intersection process.

**4. Rapid Association Rule Mining (RARM).** RARM [1] is another association rule mining methodology that

uses the tree structure to represent the initial information and avoids candidate generation method. RARM is claimed to be a lot of quicker than FP-Tree algorithmic program with the experiments result shown within the original paper. By exploitation the SOTrieIT structure RARM will generate giant 1- itemsets and 2-itemsets quickly while not scanning the information for the second time and candidate's generation. Just like the FP-Tree, each node of the SOTrieIT contains one item and also the corresponding support count.

## IV.EVALUATION PARAMETERS

**1.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.

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

**2. 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.

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

**3.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.

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

## V. TECHNIQUES OF PRIVACY PRESERVING MINING

Privacy preserving techniques can be classified based on the protection methods used by them.

**1. Data Perturbation** -Data is directly modified in this technique so it come under data modification category. It is a category of data modification approaches that protect the sensitive data from intruders. Here selected portion of the dataset is considered as the sensitive information which need to be hide by modifying those values or information.

So released data is contained inaccurate data where sensitive information is modified. While doing modification it is required to do perturbation in the information having same statistics, as different values get directly act as outliers. So perturbation divided into two main category first is probability distribution approach and the other is value distortion approach. The approach of probability distribution, replaces the data with same data from the distribution of value present in original.

**2.Noise Addition** -This technique is applied on numeric data only as noise can be produce by some noise producing function such as Gaussian function. Here data quality is maintained by the technique so it look like original, while privacy of the information is maintained [8].

The underlying distributions of a perturbed data set can be unpredictable if the distributions of the corresponding original data set and/or the distributions of the added noise is not multivariate normal. In such a case responses to queries involving percentiles, sums, conditional means etc. Some noise addition techniques, Probabilistic Perturbation Technique, Random Perturbation Technique, All leaves probabilistic perturbation technique.

**3.Data Swapping**- Data swapping techniques mainly applied on the datasets where it was desired to keep all original values in the data set, at the same time the record re-identification is very difficult [1]. Data swapping means replaces the original data set by another one. Here some original values belonging to a sensitive attribute are exchanged between them. This swapping can be done in a way so that the t-order statistics of the original data set are preserved. A t-order statistic is a statistic that can be generated from exactly t attributes. A new concept called approximate data swap was introduced for practical data swapping.

It computes the t-order frequency table from the original data set, and finds a new data set with approximately the same t-order frequency. The elements of the new data set are generated one at a time from a probability distribution constructed through the frequency table. The frequency of already created elements and a possible new element is used in the construction of the probability distribution. Inspired by existing data swapping techniques used for statistical databases a new data swapping technique has been introduced for privacy preserving data mining, where the requirement of preserving t-order statistics has been relaxed.

**4. Suppression**-In suppression technique, sensitive data values are deleted or suppressed prior to the release of a data [4]. This technique is used to protect an individual privacy from intruder's attempts to accurately predict a suppressed value. A Sensitive value is predicted by an intruder through various approaches. For example, a

built classifier on a released data set can be used in an attempt to predict a suppressed attribute value. Therefore sufficient number of attribute values should be suppressed in order to protect privacy. However, suppression of attribute values results in information loss. An important issue in suppression is to minimize the information loss by minimizing the number of values suppressed. For some applications like a medical diagnosis the suppression is preferred over noise addition in order to reduce the chance of having misleading patterns in the perturbed data set.

## IV.CONCLUSIONS

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 shown various approaches of different researchers proposed in their work. Here frequent pattern finding algorithm is discussed which can help in retrieving the information. Finally evaluation parameters with privacy preserving techniques are also explained in this paper for complete understanding of the field. As research is never end handle so in future one can embrace other method of privacy preserving for stored data on servers.

## 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]. Huhtala, Y., Karkkainen, J., Porkka, P., and Toivonen, H., (1999), TANE: An Efficient Algorithm for discovering Functional and Approximate Dependencies, The Computer Journal, V.42, No.20, pp.100-107.

[5]. Lichun Li, Rongxing Lu, Kim-Kwang Raymond Choo, Anwitaman Datta, and Jun Shao. "Privacy-Preserving-Outsourced Association Rule Mining on Vertically Partitioned Databases". IEEE Transactions On Information Forensics And Security, Vol. 11, No. 8, August 2016 1847.

[6]. Shyue-liang Wang, Jenn-Shing Tsai and Been-Chian Chien, "Mining Approximate Dependencies Using Partitions on Similarity-relation-based Fuzzy Databases", IEEE International Conference on Systems, Man and Cybernetics(SMC) 1999.

[7]. Yao, H., Hamilton, H., and Butz, C., FD_Mine: Discovering Functional dependencies in a Database Using Equivalences, Canada, IEEE ICDM 2002.

[8]. Wyss. C., Giannella, C., and Robertson, E. (2001), FastFDs: A Heuristic-Driven, Depth-First Algorithm for Mining Functional Dependencies from Relation Instances, Springer Berlin Heidelberg 2001.

[9]. N. V. Muthu Lakshmi1 & K. Sandhya Rani, "Privacy Preserving Association Rule Mining in Vertically Partitioned Databases," In IJCSA, vol. 39, no. 13, pp. 29-35, Feb. 2012.

[10]. F. Giannotti, L. V. S.Lakshmanan, A. Monreale, D. Pedreschi, and H. Wang, "Privacy-Preserving Mining of Association Rules from Outsourced Transaction Databases," IEEE Syst. J., vol. 7, no. 3, pp. 385- 395, Sep. 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.

[13]. L. Li, R. Lu, S. Member, K. R. Choo, and S. Member, "Privacy Preserving-Outsourced Association Rule Mining on Vertically Partitioned Databases," IEEE Trans. Info. Foren. Secur., vol. 11, no. 8, pp. 1847–1861, Aug. 2016.