International Journal of Advance Research in Science and Engineering
Volume No.07, Special Issue No. (03), January 2018
www.ijarse.com
IJARSE
ISSN: 2319-8354

An Robust Data Distribution Approach By Utilizing Frequent Pattern Mining Tree Rules

Avinash Sharma[1],Dr. N.K. Tiwari[2]

[1,2]Research scholar

Director Bansal Group of Institution

Bhopal(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 by frequent pattern tree. Here sensitive set of rules are perturbed by substitution. This type of substitution reduces the risk and increase the utility of the dataset as compared to other methods. Analysis is 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, subsitution.**

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 itemset mining calculation and returns the came about regular itemsets and their backings to the information proprietor. The information proprietor modifies these itemsets' backings by subtracting them with itemsets' relating event check in the fake exchanges separately. At that point, the information proprietor decodes the got itemsets with the amended backings higher than the frequency limit and produces association rules in view of the incessant itemsets. In these setting, information proprietor requires including itemset 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.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.

Modify Frequent Pattern Tree

In this step transaction comes in the dataset are pass in the tree such that various combination of the items in the transaction are count in this pass. Here inverse sequence is pass in the tree where items present in this transaction are count by various parent in the tree. This can be understand as:

A, B, D
A, C, D
C, B
B, D, A

Table 1 Represent transaction set of elements.

Let number of different items in the transaction sets are four, than tree has four child. Find number of different combination of the item set as per the set cardinality like cardinality 2 = {AB, AC, AD, BC, BD, CD}, cardinality 3 = {ABC, ACD, BCD}, cardinality 4 = {ABCD}. Construct tree other level as per the cardinality node shown in fig. 2.

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

International Journal of Advance Research in Science and Engineering
Volume No.07, Special Issue No. (03), January 2018
www.ijarse.com
IJARSE
ISSN: 2319-8354

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.

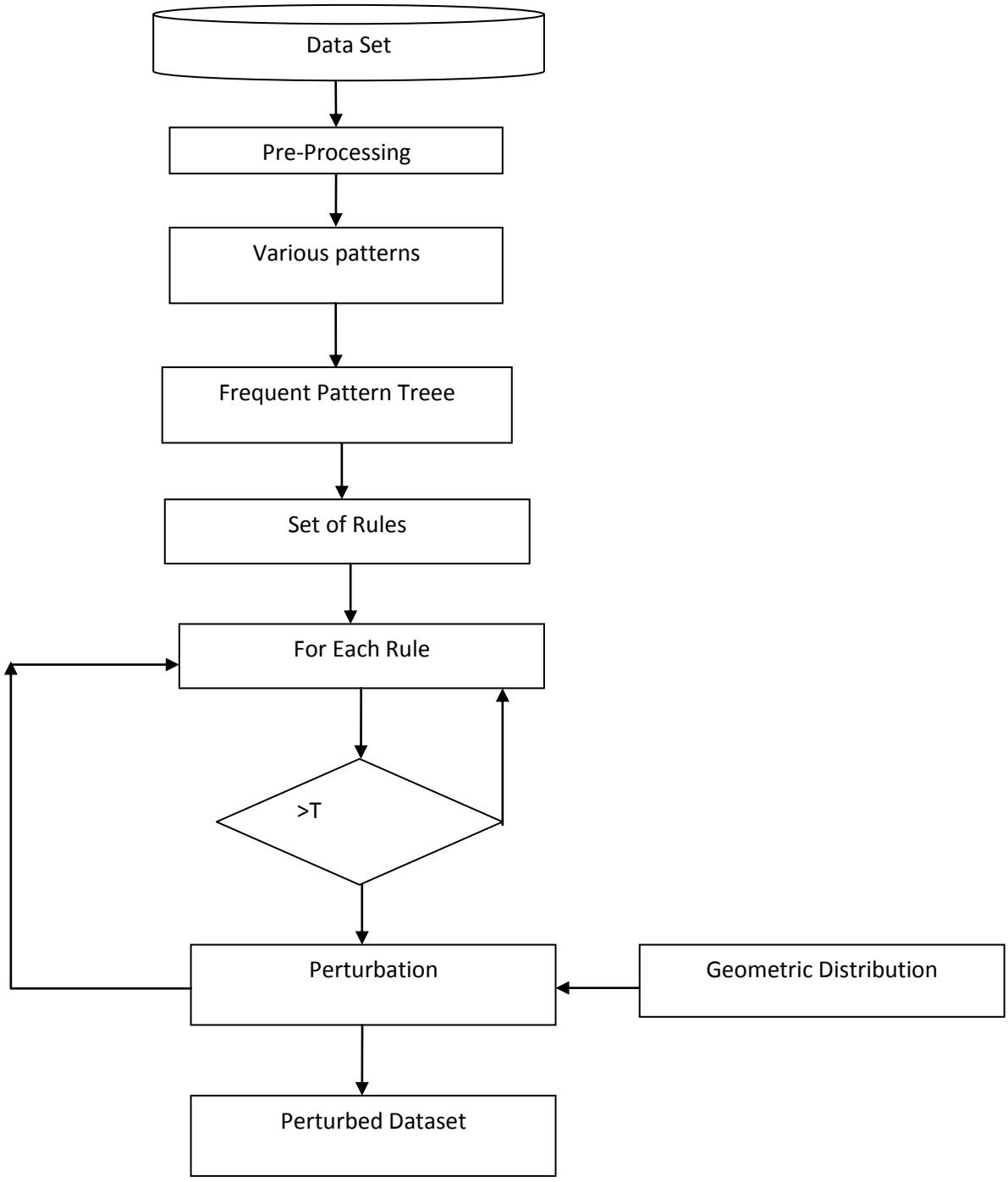

Fig. 1 Block diagram of proposed work rule generation.

International Journal of Advance Research in Science and Engineering
Volume No.07, Special Issue No. (03), January 2018
www.ijarse.com
IJARSE
ISSN: 2319-8354

Geometric Distribution

$$f(x \mid p) = \sum_{i=0}^{x} pq^{x}$$

Where q = 1 – p, x is a matrix of numbers while p is a probable value for the generation of series.

In this approach original dataset is change in random portion where amount of change is depend on the minimum threshold. The original values but not in the same order as was in the original dataset. In [10] noise is generate by a Gaussian function that produce a sequence of number then add those sequence in the original position, so a kind of variation is develop over here for the privacy of the original one, but that was limited to the numeric only.

Sensitive Pattern Hiding:

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] ← MFPT(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 substitution has been included.

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:

International Journal of Advance Research in Science and Engineering
Volume No.07, Special Issue No. (03), January 2018
www.ijarse.com
IJARSE
ISSN: 2319-8354

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.RESULTS

Dataset Size	Originality percentage	
	Previous work [5]	Proposed Work
400	400	452
1200	1200	1356
5000	5000	5650

Table 5.1 Comparison of proposed and previous work on the basis of dataset size.

From the above figure and table it is obtained that proposed work has maintain the same dataset size after applying the perturbation algorithm. Here by change in the dataset value dataset size of the previous work is increase than proposed work.

Dataset Percentage	Risk Value	
	Robfrugal [5]	Proposed Work
400	6800	7203
1200	20400	21469
5000	85000	88879

Table 2 Comparison of proposed and previous work on the basis of Risk values.

International Journal of Advance Research in Science and Engineering
Volume No.07, Special Issue No. (03), January 2018
www.ijarse.com
IJARSE
ISSN: 2319-8354

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 session addition have reduce risk as compare to previous but not that much as done by substitution algorithm proposed in this work. Here proposed work replace less informative data so risk of the outsourced dataset was quit less.

Dataset Percentage	**Utility Value**	
	Robfrugal [5]	**Proposed Work**
400	205.5197	117.9165
1200	**631.9918**	**347.7699**
5000	**2.7110e+03**	**1.4598e+03**

Table 3 Comparison of proposed and previous work [5] on Utility Value basis.

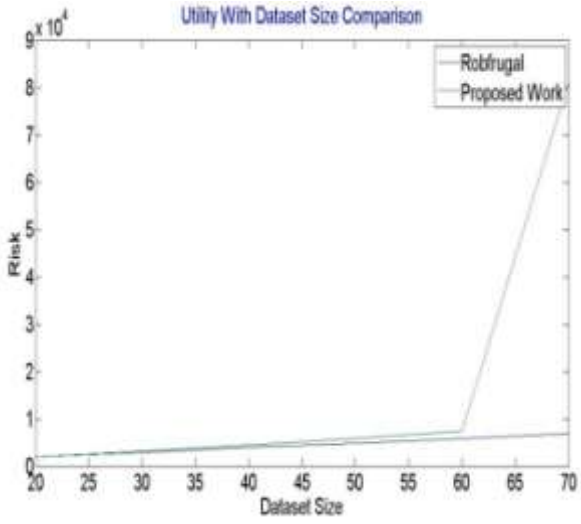

Figure 3 Comparison of dataset variation with utility value obtained from various approaches.

From above figure and table it is obtained that proposed work has increase the utility value of the dataset after applying the proposed work. In other words previous work has increased the utility value but to less extent. It was obtained that utility of session addition have some time increase while some time decrease as well as compare to previous but not that much as done by substitution algorithm proposed in this work. Here proposed work replace less informative data so risk of the outsourced dataset was quit less. But perturbation done by fake transaction or removing item has reduce the utility to large values.

VI.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 frequent pattern tree. By the utilization of substitution security of the information at server side get upgrade too. Results demonstrates that proposed work risk value get decrease. While utility was high as compare to previous work. By the utilization of programmed same session space cost is additionally maintained. As research is never end handle so in future one can embrace other example era method for enhancing the server execution.

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, "PrivacyPreserving-Outsourced Association Rule Mining on Vertically Partitioned Databases," IEEE Trans. Info. Foren. Secur., vol. 11, no. 8, pp. 1847–1861, Aug. 2016.