International Journal of Current Trends in Engineering & Technology
ISSN: 2395-3152
Volume: 03, Issue: 05 (SEPTEMBER -OCTOBER, 2017)

Multiparty Privacy by Utilizing Modified Frequent Pattern Tree and Super modularity Approach

Archana Singh[1], Varun Singh[2]
Department of Computer Science & Engineering
RGPV, Bhopal
[1]erarchanasigh83@gmail.com

Abstract— Data mining systems can remove concealed however valuable data from substantial databases. Most effective methodologies for mining conveyed databases assume that the greater part of the information at each site can be shared. Be that as it may, source exchange databases normally incorporate extremely sensitive data this paper explores generation of association rule by using the modified frequent pattern tree. Here use of this tree reduces the database scan time. In this work super class substitution increases the privacy of the work. Here use of these techniques reduces the risk of data leakage to the third party. Experiment is done on real dataset for finding patterns and hiding information. Results show that proposed work has reduced the execution time while risk of sharing the data also gets reduced. At the same time utility of the proposed work is better as compare to previous work similar kind of work.

Keywords:-Frequent pattern mining, Data Perturbation, Privacy Preserving Mining, Substitution, Threats.

I. INTRODUCTION

The objective of previous advances in information mining systems is to proficiently find significant and non-clear learning from vast databases [9]. The mining of affiliation rules assumes a vital part in different information mining fields, for example, money related investigation, the retail business and business basic leadership [9]. Current associations have their own particular databases, situated in better places. Most mining procedures expect that the information is brought together or the appropriated measures of information can effectively move to a focal site to wind up noticeably a solitary model. In any case, associations might will to share just their mining models, not their information. These brought together strategies have a high danger of unforeseen data spills when information is discharged [5]. Companies earnestly expect assessment to diminish the danger of revealing data. Protection Preserving Data Mining (PPDM) can run an information mining calculation to acquire commonly gainful worldwide mining destinations without uncovering private information [7]. Accordingly, PPDM has turned into a critical issue in numerous information mining applications. A basic technique for PPDM in conveyed databases is to annoy the first information. The system of changing the first database into another one that conceals some sensitive affiliation rules is known as the sanitization procedure. Playing out a mining procedure on the purified database can lessen the danger of uncovering the delicate data [12]. Be that as it may, the mining result on the sterilized database is less exact than that of the first database. In some business situations, maybe the information mining ought to be handled among databases. By the by, information might be disseminated among a few destinations, however none of the locales is permitted to open its database to another site. Consider the accompanying situation: Some insurance agencies have their own databases that record their guarantee's data. For shared advantage, these organizations choose to collaborate for protection extortion identification by conveyed information mining. The information mining model must be high exact to distinguish extortion, in light of the fact that a mix-up brings about extraordinary loss of income or awesome measures of pay. Additionally, insurance agencies can't impart information about their clients to different organizations, attributable to the limitation laws (and having a high aggressive edge). They may share learning about false privacy records, yet not their information. Each organization endeavors to share their "square box" models to find all the more fascinating tenets overall shared data than that all alone database, and can ensure the selective records that different organizations may discover [5]. Secure Multiparty Computation (SMC) [6, 7] utilizes conveyed calculations in a safe way. SMC jam singular security, as well as means to safeguard spillage of any data other than the last outcome. Be that as it may, customary SMC strategies require a high correspondence overhead. They don't scale well with the database estimate. Along these lines, this investigation concentrates on the issue of security saving mining continuous item sets in various conveyed databases, in which every exchange completely has a place with just a single site, with a low correspondence necessity and without irritating the first information.

II. RELATED WORK

In [6] data anonymization is a promising procedure inside the train of security safeguarding information mining used to ensure the data contrary to character revelation. Data misfortune and since a long time ago settled assaults conceivable on the anonymized data are basic difficulties of anonymization. Not very far in the past, learning anonymization using data mining techniques has indicated huge change in data utility. In any case the predominant methodologies need in secured treatment of assaults. Thus J. Jesu Vedha Nayahi

et al. proposed an anonymization calculation built up on grouping and flexible to comparability assault and probabilistic derivation assault is proposed.

In [7] R. Rajeswari et al. Proposes a security continued on get to control instrument for information streams. For the privacy security component it makes utilization of the blend of both the Kanonymity technique and fracture framework. The kanonymity method makes utilization of the concealment and speculation. It keeps the privacy disclosure of the delicate data. The security barrier component evades the personality and qualities divulgence. The privacy is executed by methods for the high exactness and consistency of the individual skill, i.e., the accuracy of the individual information.

In [19] typical information distributing ways will get rid of the delicate qualities and produce the extensive records to achieve the objective of security wellbeing. . In the enormous information condition, the necessity of utilizing data (e.g., information mining) come to be increasingly a considerable amount of, which is past the extent of the typical strategy. Tong Li et al. Presents a cryptographic information distributing framework that jam the data respectability (i.e., the since a long time ago settled learning structure is protected) and accomplishes namelessness without erasure of any property or use of repetition. The wellbeing investigation recommends that their procedure is secured underneath proposed security demonstrate.
In [10]Surbhi Sharma et al. Show how the elite department of same gathering join their information without hurting the privacy of the customer for making secured choices in effective and redress way. Thus the methodologies vertically data mix, cryptography and choice mining is built up. To mine the decisions from the data a C4.5 determination tree is utilized. The usage of the proposed privacy safeguarding information mining and basic leadership strategy is done utilizing JAVA innovation. Also the proficiency of the strategy is registered in expressions of exactness, blunder rate, memory utilization and time utilization. At last to legitimize the impacts of the proposed information mining framework the ordinary J4.5 tree using WEKA instrument is utilized with same information for similar execution learn. The test comes about demonstrate the powerful execution and insurance inside the given privacy safeguarding technique.

III. PROPOSED METHODOLOGY

In this section whole work is explained where two level of security is maintained. First is replacing the super modularity content with its subclass, while second is to hide sensitive information from the dataset.

Pre-Processing: As dataset are available in various format so conversion of data as per required environment is done in this step. Here some of information which uniquely identifies any individual or organization is directly removed from the set. In this progression entire multi qualities are supplant by its chain of command a value in the super modularity tree, while supplanting it is required to adjust the dataset utility and risk by rolling out required improvements. This was done in [4]. This substitution is designed to the point that utility of the information get increment while chance stay beneath under some limit esteem.

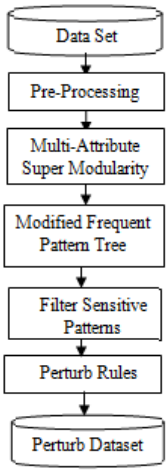
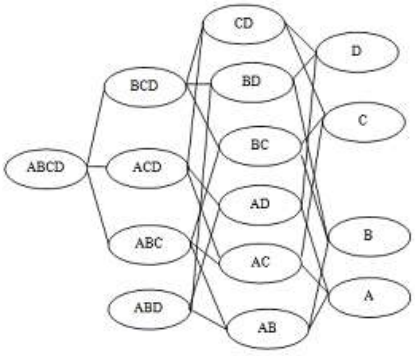


Fig. 1 Proposed work block diagram.

Multi-quality Super-modularity
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 passing in the tree where items present in this transaction are count by various parent in the tree. This can be understood 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 children.



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.

Now pass first transaction in the set whereas per item presence in the transaction child counter value get raise by one. This can be understood as first transaction set is {A, B, D}, than node whose count raise by one are {A, B, D, AB, AD, BD, ABD}. In similar fashion other set of transactions are pass in the tree and each set of pattern count get raise by one.

Separate Sensitive pattern

Presently from the set of rules one can get cluster of rules that are sensitive for some threshold value. This can be understood as let support of any pattern is 20% which can be obtained by dividing the total number of rule count to the dataset size. Here if pattern ABC is present in 200 transaction of the dataset and total transactions are 1000, than (200/1000)*100 is the support of that pattern. So pattern crossing minimum threshold support value is considered as the sensitive patterns.

Perturb Transaction: Now next step is to calculate number of transaction that need to be modified in order to hide the pattern. For this follow steps *(1)* Now find support that is lacking from the maximum support Noise = (Max_support – Rule_support) *(2)* Find fake transaction number for each rule by Transaction = (Noise x Dataset_size)/100. For this A, B, C is a pattern then change A by its complementary value A' in case of binary items or remove A if complementary value is not present. So number of transaction where this A value is present can be convert into A'. In this way transaction can be modified by sensitive patterns.

IV. EXPERIMENT AND RESULTS

This segment exhibits the trial assessment of the proposed work of perturbation and encryption procedure for protection of multiparty dataset. All calculations and utility measures were executed by utilizing the MATLAB tool. The tests were performed on a 2.27 GHz Intel Dual Core machine, furnished with 2 GB of RAM, and running under Windows 7 operating system.

Dataset:

To analyze proposed calculation, it needs the dataset. One basic adult dataset is utilized that has total fourteen attributes. Here individual data are present like gender, education, marital status, salary, etc. Whole dataset consist of 32561 sessions. In this work, an arrangement of calculations and systems were proposed to take care of security safeguarding information mining issues. The analyses demonstrated that the proposed calculations perform well on huge databases. From the given table it is obtained that proposed work has highly maintained the originality of the dataset after applying the perturbation algorithm.

Table 5.1 Comparison of proposed and previous work on the basis of Originality percentage.

Dataset Percentage	Originality Percentage	
	Previous work	Proposed Work
20	99.4463	85.4681
40	99.1750	85.7907
60	99.0738	86.1519

Here by change in the sigma value originality of the previous work is move around 69% while proposed work has originality around 93%. Here pattern preservation has less affect on dataset while previous approach was having higher affect.

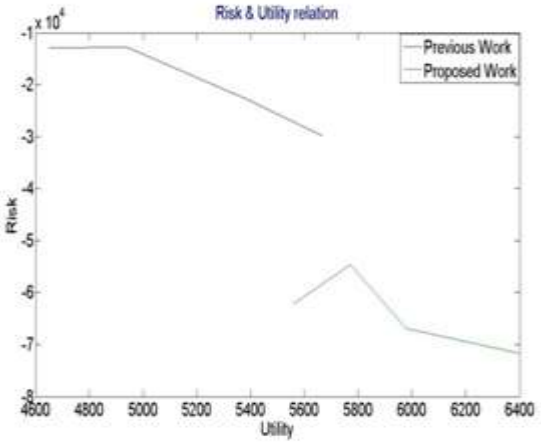
Table 5.2 Comparison of proposed and previous work on the basis of Risk values.

Dataset Percentage	Risk Value	
	Previous work	Proposed Work
20	7.1508e+04	5.5334e+04
40	5.1876e+04	5.3239e+04
60	6.2694e+04	5.0824e+04

From table 5.2 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. From given table it is obtained that proposed work has increase the utility value of the dataset after applying the proposed work. As the previous work was having lower utility value. From given graph it is obtained that proposed work has increased the utility value of the perturbed dataset as compared to the previous work. While one more evaluation is obtained that risk of the proposed outcome is quite low as compare to the previous work. In other words, previous work was having lower utility value.

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

Dataset Percentage	Utility Value	
	Previous work	Proposed Work
20	2.6834e+03	7.5672e+03
40	773.4074	7.3371e+03
60	1.8730e+03	7.0827e+03



From above table 6 it is obtained that proposed work have less space cost as compare to the previous work. As high sensitive rules are perturb below some threshold confidence so no need to increase the number of fake transaction for increasing the confusion in dataset.

V. CONCLUSIONS

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 differential privacy and then directly hides 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]. 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]. B.Vani, D.Jayanthi, (2013), "Efficient Approach for Privacy Preserving Microdata Publishing Using Slicing" IJRCTT.

[5]. J. Liu, J. Luo and J. Z. Huang, "Rating: Privacy Preservation for Multiple Attributes with Different Sensitivity requirements", in proceedings of 11th IEEE International Conference on DataMining Workshops, IEEE 2011.

[6]. J. Jesu Vedha Nayahi and V. Kavitha," Privacy and utility preserving data clustering for data anonymization and distribution on Hadoop", Future Generation Computer Systems, 0167-739X/© 2016 Elsevier.

[7]. R. Rajeswari and Mrs R. Kavitha ,"Privacy Preserving Mechanism for anonymizing data streams in data mining", International conference on current research in Engineering Science and Technology(ICCREST-2016).

[8]. Elahe Ghasemi Komishani, Mahdi Abadi, and Fatemeh Deldar," Preserving personalized privacy in trajectory data publishing by sensitive attribute generalization and trajectory local suppression", Future Generation Computer Science(2016).

[9]. Tong Li, Zheli Liu, Zin Li, Chunfu Jia and Kuan-Ching Li, "A Cryptographic Data Publishing System", J. Computer System Science(2016).

[10]. Surbhi Sharma and Deepak Shukla, "Efficient multi-party privacy preserving data mining for vertically partitioned data", Inventive Computation Technologies (ICICT), 10.1109/INVENTIVE.2016.7824852, © 2017 IEEE.

[11]. Mohamed R. Fouad, Khaled Elbassioni, And Elisa Bertino. "A Super modularity-Based Differential Privacy Preserving Algorithm For Data Anonymization". IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 7, July 2014 1591

[12]. Jerry Chun-Wei Lin, (Member, IEEE) 1, Qiankun Liu1, Philippe Fournier-Viger2, And Tzung-Pei Hong. "Pta: An Efficient System For Transaction Database Anonymization". August 25, 2016, Date Of Current Version October 31, 2016. Digital Object Identifier 10.1109/Access.2016.2596542.