JCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-1, Jan 2019 E-ISSN: 2347-2693

An optimized algorithm for association rule hiding technique using Hybrid Approach

Apoorva Joshi[1*], Pratima Gautam[2]
[1]Career College Bhopal, India
[2]AISECT University Bhopal, India

Corresponding author: apoorvajoshi16@gmail.com

Available online at: www.ijcseonline.org

Accepted: 06/Dec/2018, Published: 31/Jan/2019

Abstract - Data mining is basically a process of identifying different patterns from huge amount of data. Association rule hiding technique target to find out dependency relationships transversely attributes. It may reveal sensitive information. With widespread applications of data mining techniques to diverse domains, privacy preservation becomes compulsory. Association rule hiding is one of the important techniques of privacy preserving data mining to keep the sensitive association rules generated by association rule mining. In this paper we talk about heuristic approach for hiding sensitive rules. The proposed technique provides us solution of privacy disclosure and hides the sensitive rules.

Keywords- Association Rule Hiding, Data Mining, Privacy Preserving Data Mining**,** Distortion, Representative Rule

I. INTRODUCTION

Data mining is broadly used in many different applications. But distant from the reward it provides, it is also viewed as a danger to privacy. Confidentiality of data is a very vital feature that each & every information system has to satisfy. When a person or group is able to veil information about them and expose themselves the manner they want is termed as privacy. The whole data may be positioned at a distinct location or scattered at various sites referred to as centralized or distributed respectively. In a scattered environment, several logically unified data are dispersed among various sites [1]. The sharing of data may be horizontal, vertical or hybrid. In horizontal partitioning all sites contains a subset of minutes of the original ecstasy R. In vertical partitioning, every site contains only a division of the attributes of a table R. As the data is fragmented in each site client can access its segment of the data without intervention of other site location users. In This paper we talk about association rule hiding in the scattered environment (horizontal and vertical), at the similar time sharing exposed worldwide non-sensitive system to many legitimate users. Agrawal at al. introduced association rule mining.

Association rules can be established lacking disclosing a few sensitive information [2]. The most common method of transformation, the conversion that used in PPDM usually using one technique or mixture from several techniques. The combination technique frequently focused on one method, not optimizing each technique. In research by Li [13] conversion is completed by using entropy partition-based and frequency distortion-based and in earlier research by Geng and Zhang [14] conversion is done using attribute partition-based and combined distortionto optimize the transformation thus. This transformation called hybrid transformation. The hybrid transformation will be completed by partitioning based on entropy and then they apply combined distortion.

II. ASSOCIATION RULE HIDING TECHNIQUE

Association rule hiding is basically sub area of privacy preserving data mining to studies the demerits of data mining methods that occurs from the unveil the sensitive information belong to persons or organizations. They accessible of many extensive set of application scenarios in which composed data or information patterns extracted from the data have to be public with others entities to serve up holder or organization for a particular purposes. The allocation of data or information might complete at a cost to privacy, mainly due to two main reasons:

(i) If the data pass on to individuals, then its exposé can defy the privacy of the persons who are recorded in the data. If their uniqueness is exposed to not trust third parties or if receptive knowledge concerning them can be extracted from the data.

(ii) If the data regard business or organizations information, then it exposé of this data or any information mined from the data may potentially expose responsive job secrets, whose information can suggest a significant profit to business competitors and thus can origin the information holder to misplace business information over the peers. The difficulty of hiding association rule can be measured as a type of database deduction control, but its chief objective is to guard the sensitive rules not the sensitive data [15] (the infringement of privacy is coming from sensitive association rules rather than the data itself). In association rule hiding contents a set of sensitive association rules, which are particular by the protection of administrator or data owner, the main task of the association rule hiding algorithms is to disinfect the data so that the association rule mining algorithms functional to the data. (i) It will be unable to dig out the responsive rules (ii) it mine all the non-responsive rules. Presently several methodologies have been used to hide sensitive association rules by applying certain changes in the main data set. Due to these certain changes, a few non-sensitive patterns may be misplaced or missing is called lost rules, and new patterns are furthermore generated is called as ghost rules [3].

III. APPROACHES OF ASSOCIATION RULE HIDING ALGORITHMS

A. Heuristic Approach

Data Distortion technique Data Distortion technique is used for changing data using a random value. This technique apparently distorts responsive data values by adding random value, data reverse matrix, or adding mysterious values etc. This method can grip over dissimilar data types: character, Boolean, arrangement and numbers. Discrete data require unique data set to be processed. These type of processing fragmented into the attribute coding as well as obtaining sets coded data set [3]. The demerits of distortion techniques it is extremely difficult to conserve the original data.

Data blocking techniques Blocking mechanism is by reduction of the level of support and confidence of responsive association rules and changing some attribute values of data sets with unknown values or swap ‘1’ by ’0’ or ‘0’ by ‘1’. This technique of privacy preservation is completed in two steps. First is to identify dealings of responsive rule and second is to swap the known values to the unknown, so that to sustain of definite items goes downward to a assured level and rule mining algorithm is to not able to mine the responsive rules [16]. Main problem with block-based privacy preserving association rule mining is it is too tough to compute the support and confidence of a responsive association rule while the some of the novel data is replaced with unidentified value [17]. This problem can be solved by using tentative secret code which can be restored with real support and confidence.

B. Border Revision Approach

Border revision approach it modifies borders in the web of the frequent and infrequent item sets to veil responsive association rules. This approach basically tracks the border of the non responsive frequent item sets and covetously applies on data alteration that may have negligible impact on the value to lodge the hiding sensitive rules. Researchers planned numerous border revision algorithms such as (BBA) Border Based Approach like Max– Min1 and Max-Min2 to conceal responsive association rules. This algorithm uses odd techniques such as deleting explicit sensitive items and tries to minimize the amount of non sensitive item sets that may be lost even as sanitization is performed more than the novel database in arrange to guard sensitive rules [4].

C. Exact approaches

This approach is to follow the hiding procedure as a constraints fulfillment trouble which is later sort out by binary integer programming (BIP). These approaches offer improved result. However they undergo from sky-scraping time complexity. Gkoulalas and Verykios [18] projected an approach for verdict most favorable solution for hiding the rule dilemma which tries to reduce the space between the unique data set and its sanitized data set. The authors in [19] planned a border-based approach to facilitate a most favorable solution to veil the responsive frequent item sets by extending the novel data set by an unnaturally generated data set. Extending the novel data set for responsive item set hiding is proved to offer most favorable solutions to an comprehensive set of hiding troubles compared to earlier approaches and to offer solutions of advanced quality[5]

D. Reconstruction Based Approach

This method is implemented by disturbing the data first and reconstructing the distributions at an cumulative level in sort to execute the association rules mining. Mielikainen [6] was the primary analyzed the computational complexity of inverse frequent set mining and showed in numerous problems these problems are computationally hard. In this method firstly put the novel data to the side and begins from knowledge base. To disinfect, it conceals the responsive rules by sanitizing itemset lattice more willingly than sanitizing original dataset. Moreover FP tree approach which is based on opposite frequent set mining algorithm. The planned model has three dissimilar phases, first phase generates frequent item sets from the novel database, second phase performs refinement algorithm above the frequent item sets by selecting hiding approach and identifying responsive frequent items sets according to sensitive association rules. The last phase generates sterile database by using inverse frequent item set mining algorithm and afterward releases this record. In reconstruction based approaches, first frequent sets are generated, from these non responsive frequent set, fresh dataset is generated which basically conserve the solitude of sensitive data.

E. Cryptographic Technique

Sensitive information can be able to encrypt by cryptography technique. In [7], authors introduced cryptographic technique which is extremely popular since it provides protection and security of sensitive values. There are further cryptography algorithms obtainable. But these methods contain many demerits like to protect the amount produced by computation. The algorithm proposed in [7] does not offer wealthy results in case of multiparties and also it is extremely complicated to pertain this algorithm to vast databases.

IV. PROBLEM STATEMENT

Numerous PPDM methods have been planned to defend sensitive data in each PPDM coating. But, a few troubles still stay to be addressed in the upcoming time. Firstly, personalized PPDM should be deliberate. Because confidentiality is a individual concept regarded as a personal matter, typically privacy preservation required to be customized. Even though there are slight efforts have been conducted for solving this problem, e.g., condensation-based method [9] that were build up for personalized data preservation, we haven't up till now notified personalized PPDM techniques that are formerly applied in real live applications. Moreover, sometimes it is not realistic to require all entity to discover its individual privacy preservation phase. This might be an enormous usability complexity Therefore; additional research needs to be carried out for solving modified PPDM towards reasonable applications.

Secondly, the faith of data miner must be evaluated in organize to optimize PPDM. Distant from personalized privacy preservation for data holder, huge trust of data-miners is also a release issue that affects the level of alteration of raw data for privacy preserving reason. Li et al. proposed a technique to facilitate multi-level faith of data-miners by given that multiple unique disturbed data copies [20]. But, this technique did not deal with how to describe faith levels of data-miners and how to choose the exact number of troubled copies must be used to corresponding trust levels. Thirdly, harmonizing privacy defeat and information failure is also a challenge in PPDM. There is a normal tradeoff among privacy preservation and information loss in PPDM. Usually, the superior level of privacy preservation, the more novel data are required to be troubled, which causes additional information loss. How to maintain a balance among privacy loss and information failure is really not a simple task. Some methods have been proposed to tackle this issue by quantify the transaction among these two indexes.

Table 1

Approach		**Advantage**	**Limitation**
Heuristic approach	**Data Distorti**	More capable, scalable	Difficult to revert the changes made in database
	Data Blocking	It maintains authenticity of database, since substitute of insert false value it just blocks unique value.	Suffer from various side effects like ghost rule, lost rule etc.
Border approach		Preserve database quality by selecting the transaction that produces minimal side effect.	Theory of border complicated to distinguish Based on heuristic approach
Exact approach		Provides an optimal solution lacking any side effects	High complexity due to linear integer programming
Reconstruction approach		slighter side effect than heuristic based approaches	Number of operation is limited in new released database
Cryptography approach		Provide security in multi party computation	Does not offer security for the output of the calculation and it is very difficult to pertain on massive databases

However the accessible methods were planned below definite assumptions, which are hardly ever supported in apply. For example, each and every responsive attributes are uniformly significant and have the similar consideration burden in considered algorithms. Finally, calculation of cost and effectiveness are crucial in sensible applications of PPDM. Computation competence is still an unlock matter in PPDM, particularly for combined data mining among multiple parties.At the identical time when these schemes offer a high protection level, they also begin luxurious computation and communication expenditure. As a result, even though many planned methods are theoretically protected, they may not be possible to relate in actual life applications with huge data sets and sky-scraping dimensionality. We should memo that there is no ideal privacy preserving methods for data mining. This is because data sets usually have their own individuality and different data mining responsibilities have different privacy necessities. The objective of privacy preservation throughout data mining process is to find a equilibrium between information loss and privacy loss, in order to guard sensitive knowledge from exposé while at the same time to stay the correctness of data mining results. In our view, future research should focus on optimizing the competence of PPDM solutions and studying their generalization for sensible usages and acceptance.

Table 2

Approach	Algorithm	Conclusion
Data Distortion	DCIS	Huge number of fresh rule generation and less number of rules are lost.
	ISL	
	DCIS	
Data Blocking	DSR	Huge number of rules are missing and fewer number of new rule generation.
	DCDS	
	DSRRC	

V. PROPOSED APPROACH

After analysis no. of technique in association rule we find that there is no proper approach here we optimized hybrid methodology which provide better privacy in database.

HEURISTIC BASED APPROACH

This approach involves well-organized, quick and scalable algorithms that selectively sanitize a set of contact from the original database to conceal the sensitive association rules [7].It is divided additional into two types of method that are Distortion method and blocking method.

According to the author [7] obtainable an algorithm designed for hiding sensitive association rules based on heuristic approach. Algorithm called DSSR (Decrease support of R.H.S items in rule)

The algorithm based on DSR. It hides association rule by manufacture the bunch of sensitive association rule based on right hand items. After that calculate the compassion of each bunch, sensitivity of bunsh or cluster is the addition of sensitivity of each item present in the cluster. Then index responsive transactions for each cluster and sort each and every clusters decreasing order of their sensitivities. The thrashing process hides rule by deleting ordinary R.H.S. item of rules in bunch, from the sensitive connections. This approach does not make main changes in database but it hides the responsive association rules which hold a solitary item in the right hand side of the rule only.

The algorithm is based on DSR method.

Anonymization based PPDM

Anonymization refers to an approach where identity or/and sensitive data about record owners are to be hidden. It even assumes that sensitive data should be retained for analysis. It's obvious that explicit identifiers should be removed but still there is a danger of privacy intrusion when quasi identifiers are linked to publicly available data. Such attacks are called as linking attacks. For example attributes such as DOB, Sex, Race, and Zip are available in public records such as voter list. Such records are available in medical records also, when linked, can be used to infer the identity of the corresponding individual with high probability [12].

Steps to implement hybrid approach

Step 1: Select database and Table T

Step 2: Select Key attribute, Quazi-identifier attribute and Sensitive Attribute.

Step 3: Select the set of most sensitive values M

Step 4: For each tuple whose sensitive value belongs to set M they move all these tuples to Table T1 and rest to table T2.

Step 5: Find the statistics of quazi attributes of table T1 i.e. distinct values for that attribute and total no of rows having that value.

Step 6: Apply generalization on quazi identifiers of table T1 to make it anonymized

Step 7: Attach rows of table T1 and table T2. T*=T1+T2 which is table ready to release.

Step 8: Algorithm first generates possible number of association rules by using FP-Growth from database D.

Step 9: Generated association rules are selected as a sensitive rule set by database holder, Rules contain only single R.H.S. item are one as sensitive.

Step 10: Find N clusters based on common R.H.S. item in a sensitive rule set RH.

Step 11: Check sensitivity of each cluster N.

Step 12: Index responsive transactions for each cluster and sort each and every clusters in decreasing order of their sensitivities.

Step 13: Hiding Process hides all sensitive rules by deleting common R.H.S. item of rules in cluster from the sensitive transactions.

VI. CONCLUSION

There are various technologies of hiding the association rules in the database and optimizing support and confidence. Each methodology has its advantages and disadvantages. The outcome of the proposed research is to hide the sensitive association rules of data mining with following criterions –

i. No false rules generation
ii. No information loss
iii. Modification Degree
iv. Robustness against intentional or unintentional attacks [10]

Generation of frequent item sets from a dataset plays an important role in association rule mining. At the same time privacy preserving in data mining is important to prevent the sensitive association rules from getting revealed from the data to the unwanted user. There are various algorithms for generating frequent item sets and for hiding sensitive rules from the outcome of association rule mining. Apriori algorithm is mostly used to generate frequent item sets. But with apriori, as the database increases the number of database scans required also increases thus increasing the execution time.[11] The proposed algorithm uses FP-Growth algorithm for generating frequent item sets as it needs slighter database scans while generating frequent item sets, along with heuristic approach, to hide sensitive rules from unwanted user. Proposed algorithm was tested for its performance on distributed medical database taking care that not much modifications were done here we work with annomnization technique which block the sensitive data and heuristic approach which hide the sensitive rule

REFERENCES

[1] Masooda Modaka,Rizwana Shaikhb." Privacy Preserving Distributed Association Rule Hiding Using Concept Hierarchy", 7th International Conference on Communication, Computing and Virtualization 2016. Procedia Computer Science 79 (2016)993–1000.

[2] Putri, A walia W. Laksmiwati Hira." Hybrid Transformation in Privacy-Preserving Data Mining" 978-1-5090-5671-2/16/$31.00 ©2016 IEEE.

[3] Mohamed Refaat Abdellah H. Aboelseoud '
Khalid Shafee M. Badr Senousy "Privacy Preserving Association Rule Hiding Techniques: Current Research Challenges" *International Journal of Computer Applications (0975 – 8887) Volume 136 – No.6, February 2016.*

[4] Kasthuri S1 and Meyyappan T2 "Hiding Sensitive Association Rule Using Heuristic Approach", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.1, January 2013DOI.

[5] Umesh Kumar Sahu Anju Singh "Approaches for Privacy Preserving Data Mining by Various Associations Rule Hiding Algorithms – A Survey", *International Journal of Computer Applications (0975 – 8887) Volume 134 – No.11, January 2016.*

[6] Khyati B. Jadav, Jignesh Vania Dhiren R. Patel, Ph.D "A Survey on Association Rule Hiding Methods", *International Journal of Computer Applications (0975 – 8887) Volume 82 – No 13, November 2013.*

[7] K. Naga Prasanthi "A Review on Privacy Preserving Data Mining Techniques" International Journal of Advanced Research in Computer Science and Software Engineering Volume 6, Issue 3, March 2016.

[8] Vinita Shah Divya C. Kalariya Jay Vala "Association Rule Hiding based on Heuristic Approach by Deleting Item at R.H.S. Side of Sensitive Rule", *International Journal of Computer Applications (0975 – 8887) Volume 122 –No.8, July 2015.*

[9]Ruchi.P.Kanekar1, Prof. Rachel Dhanaraj2 "Adding Dummy Items To Hide Sensitive Association Rules" *IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727 PP 06-10"*

[9] Xueyun Li Zheng Yan Peng Zhang "A Review on Privacy-Preserving Data Mining", December 2014.

[10]Mrs. Geeta S. Navale Dr. Suresh N. Mali "A Survey on Sensitive Association Rules Hiding Methods", 978-1-5386-4008-1/17/$31.00 ©2017 IEEE.

[11]Melissa Fernandes Joanne Gomes *"Heuristic Approach for Association Rule Hiding using ECLAT",* 978-1-5090-4381-1/17/$31.00 © 2017 IEEE

[12] Apoorva Joshi,, Pratima Gautam "A survey on Sanitizing Methods in Association Rule Hiding Technique" International Journal Of Scientific Research in Computer science, Engineering and Information Technology" ,volume2 Issue 6 2017.

[13] Li Jingquan. "Privacy-Enhancing Data Mining: Issues, Techniques and Measures". University of Illinois, United State of America, 2004.

[14] Xingyu Geng BP., Zhang J. "Combined Data Distortion Strategies for Privacy-Preserving Data Mining". Southwest Petroleum University, China and University of Kentucky, United States of America. 2010.

[15] Jajodia., C.F.a.S., The inference problem: A survey, in SIGKDD Exploration Newsletter. 2002. p. 6-11.

[16] A. Parmar, U.P.R., D. R. Patel. "Blocking based approach for classification Rule hiding to Preserve the Privacy in Database". In proceedings of International Symposium on Computer Science and Society, IEEE 2011.

[17] Animesh Tripathy, M.P. "A novel framework for preserving privacy of data using correlation analysis". In Proceedings of the International Conference on Advances in Computing, Communications and Informatics (ICACCI '12). ACM. 2012. NY, USA.

[18] Gkoulalas-Divanis and V.S.Verykios, "An Integer Programming Approach for Frequent Itemset Hiding", In Proc. ACM Conf. Information and Knowledge Management (CIKM'06), Nov. 2006

[19] Y. Li, M. Chen, Q. Li, and W. Zhang, "Enabling multilevel trust in privacy preserving data mining," IEEE Transactions on Knowledge and Data Engineering, vol.24, pp.1598-1612, 2012.

[20] Gkoulalas-Divanis and V.S. Verykios, "Exact Knowledge Hiding through Database Extension," IEEE Transactions on Knowledge and Data Engineering, vol. 21(5), May 2009, pp. 699-713.