© 2019 IJSRSET | Volume 6 | Issue 3 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section : Engineering and Technology
DOI : https://doi.org/10.32628/IJSRSET196368

Survey on Association Rule Hiding Techniques

G. Bhavani, Dr. S. Sivakumari
Department of Computer Science and Engineering, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, Tamil Nadu, India

ABSTRACT

Data mining process extracts useful information from a large amount of data. The most interesting part of data mining is discovering the unseen patterns without unpacking sensitive knowledge. Privacy Preserving Data Mining abbreviated as PPDM deals with the issue of sustaining the privacy of information. This methodology covers the sensitive information from disclosure. PPDM techniques are established for hiding the sensitive information even after performing the data mining. One of the practices to hide the sensitive association rules is termed as association rule hiding. The main objective of association rule hiding algorithm is to slightly adjust the original database so that no sensitive association rule is derived from it. The following article presents a detailed survey of various association rule hiding techniques for preserving privacy in data mining. At first, different techniques developed by previous researchers are studied in detail. Then, a comparative analysis is carried out to know the limitations of each technique and then providing a suggestion for future improvement in association rule hiding for privacy preservation.

Keywords : Association Rule Hiding, Data Mining, Privacy Preserving Data Mining, Sensitive data.

I. INTRODUCTION

One of the most popular activities in data mining is association rule mining [1] which is used to find frequent patterns, casual structures, correlations or associations among a set of objects or items in information repositories such as transaction databases, relational databases etc. The objective of data mining is to obtain information that is unknown, including secured information like, personal identification numbers, credit card numbers, telephone numbers, and so on. While extracting information from databases using data mining techniques there may be a chance to disclose the sensitive information. So, it is more important to protect the data during mining. This technique of protection is called Privacy Preserving Data Mining (PPDM) [2] which alleviates the problem regarding privacy revealing during the data mining process.

The PPDM techniques secure the details that are confidential during the data extraction from the database. Association rule hiding [3] is an approach adopted in PPDM so as to hide sensitive association rules. The rule hiding techniques perform sanitization process in the original database which hides the sensitive rules having the sensitive information. The association rule hiding techniques provide a sanitized database with certain conditions as the sanitized database contains non-sensitive information, the sanitized database will not expose sensitive association rules and there is no modification in the database. The sanitized database will not affect the quality of the data as it has no association rule generated newly.

Association rule hiding techniques induce some modifications in the databases, which causes certain side effects such as ghost rule, false rule, and lost rule, to the database. Association rules mined in the database using hiding algorithm is known as the ghost rule whereas the false rules are association rules which a hiding algorithm cannot mine. If the hiding algorithm cannot mine non-sensitive rules, which are in the original database from the sanitized database, it is the lost rule. The main intention of this article is studying detailed information on different techniques to hide the association rules. In addition, their limitations are addressed to further improve the association rule hiding process effectively.

The rest of the article is organized as follows: Section II provides the previous researches related to association rule hiding techniques for privacy preservation. Section III compares the performance efficiency of those techniques and Section IV concludes the survey that reviews an entire discussion.

II. REVIEW OF LITERATURE

A novel approach [4] was proposed for association rule hiding. In this approach, data distortion technique is used by which the points of the sensitive items were altered but their support value rests untouched. This technique uses the indication of illustrative rules to trim the rules and at that point hide the sensitive rules. Initially, in this approach, all association rule containing sensitive items were selected and then the rules were represented in Representative Rules (RR) format with a sensitive item. Then chose a rule from RR's which had the sensitive item on the rule and chose a transaction that completely supported RR. From the selected transaction replace the sensitive item with an alternate partially supported RR item.

Another algorithm based on intersection lattice has been developed for ARH which is known as AARHIL [5] was proposed to hide the formed set of sensitive association rules for preserving the sensitive information in the database. The theory of intersection lattice of frequent itemsets was analyzed and applied it into the issue of association rule hiding by formulating two heuristics. The first heuristic determines the victim item and concentrates on sustaining itemsets to reduce lost rules. The second heuristic allots weight to each transaction based on sensitive rules, transactions degree of safety and the number of non-sensitive association rules in the transaction.

A heuristic-based algorithm which was a modified version of rule clusters with decrease support of the right hand side item MDSRRC [6] was proposed to hide the sensitive association rules. A Matrix Apriori algorithm was proposed based on the analysis of two association algorithm called Frequent Pattern-Growth (FP-growth) and Apriori algorithm. The matrix Apriori algorithm generated association rules and the hides the sensitive information of databases using MDSRRC. In MDSRRC, transactions were ordered in decreasing order of their sensitivity. In the ordered transaction, the deleted item was assigned as 0 and formed a binary matrix. After that, the support and confidence of sensitive rules which contain deleted items were updated. In the remaining sensitive rules, the sensitive rules which were below the minimum support threshold and minimum confidence threshold were deleted from sensitive rules. This process was continued until all the sensitive rules were hiding.

Border Rule-based Distortion Algorithm [7] was proposed for covering sensitive association rules by removing certain items from a database. The above technique reduces the threshold levels of support and confidence for the sensitive rules. The positive border rule and negative border rule concepts were described to identify the rules which can be easily affected by the database modifications to generate side effects. The supporting transactions were evaluated based on their relation with negative border rules and positive border rules. The weakly relevant ones were chosen preferentially for modification.

A rule hiding approach based on the optimization of an evolutional multi-objective EMO [8] was proposed for hiding association rule. During the sanitization process, a tradeoff relation was analyzed and collaborated with the association rule hiding process using the EMO algorithm. In this process the transactions are found out by the decoded chromosomes to determine which items to be removed. Next to sanitization, the side effects were calculated to the original support and the updated supports. The sensitive association rules were hidden by removing items and it was impossible to increase the support of any rule.

A fuzzy logic approach [9] was proposed for hiding association rule hiding in big data. This approach hides association rules in big data using anonymization techniques. It removes the unwanted side effects of eliminating frequent item-sets on the entry of data. The sensitive grade of each association rule was found using suitable membership functions and it was performed based on these functions. In this approach, the association rules were hiding based on two concepts. The first one is that sensitive rules have values near threshold and the one with low confidence value are non- sensitive. Another concept is that sensitive rules have higher threshold and confidence value.

Whale optimization and Least Lion Optimization Algorithms (LLOA) [10] were introduced for privacy-preserving association rule hiding. This whale optimization is used for association rule mining of certain database and creates the rules with a recently trapped fitness function. LLOA was an advanced version of existing optimization algorithm with the addition of Least Mean Square (LMS) which is secret key to privacy. Using the secret key, LLOA transforms the old original database into the new sanitized database. The sensitive information in the database was hidden by privacy and utility factor of the objective function.

A MAXARH algorithm [11] was proposed to find the sensitive rules and offer the privacy to the sensitive rules. The novel algorithm split the association rule hiding process as conversion, mining, identification, and hiding of the best rules. It hides the association rules based on threshold of minimum support and threshold of minimum confidence. In the conversion phase, the input transaction database was converted into binary values. In the mining phase, Apriori-based association rule mining was applied on the dataset to mine the association rules. In the identification phase, the best rules were identified based on maximum support and maximum confidence. In the hiding phase, the best rules are hidden by replacing the binary values.

A genetic algorithm based hiding technique HGA and a technique for creating dummy items DIC [12] were proposed for association rule hiding. Initially, the cost of individual transactions was calculated and then selected the sensitive items one by one for modifications. All transactions were arranged in decreasing order of transaction cost then each transaction was modified from 1 to 0 and then the new cost values was calculated to form a new modified database. DIC technique was used to hide the sensitive rules and also produced mock items for the altered sensitive items.

Cuckoo optimization algorithm [13] was proposed to hide sensitive association rules. The process of hiding rules was performed in cuckoo optimization algorithm using the distortion technique. In order to avoid the increasing sanitization time, a pre-process with two phases were introduced. In the first phase, only the critical transactions were selected by processing the original database. In the second phase, preprocess operation were addressed only those sensitive items with a critical role in sanitization. Then, three fitness functions were defined which were used to achieve the best solution to hide the sensitive association rules. An immigration function was introduced which had the ability to escape from any local optimum.

III. RESULTS AND DISCUSSION

This section presents a detail about the merits and demerits of different association rule hiding techniques whose functional information is discussed in the previous section. Through the review on different association rule hiding techniques, the following challenges are addressed. A novel approach is applicable only for small databases. In AARHIL based association rule hiding technique, there is no improvement in terms of accuracy. In MDSRRC, the side effects due to reducing database need to be reduced. The Border Rule-based Distortion Algorithm has a high CPU time problem. The density of dataset affects the performance of Evolutionary Multi-Objective Optimization. If there are any changes in membership function of the fuzzy logic approach, then it causes some change in height of appropriate generalization. The convergence speed of LLOA depends on their stopping criterion. Still, MAXARH based association rule hiding technique has side effects due to hiding association rules. The major drawback of HGA and DIC is a high artifactual error rate. The hiding failure of the cuckoo optimization algorithm is high when the number of iterations ranges from 0 to 5 for chess dataset. From the following Table 1, the most challenging issues in association rule hiding techniques for privacy preservation are observed and an ideal solution is identified to overcome those issues for association rule hiding.

TABLE I

COMPARISON BASED ON METHODS

Methods Used	Merits	Demerits	Results
Novel Approach [4]	Requires less number of database scans	Applicable only for small database	Dataset 1: Number of pruned rules = 6 Dataset 2: Number of pruned rules = 6 Dataset 3: Number of pruned rules = 3 Dataset 4: Number of pruned rules = 6 Dataset 5: Number of pruned rules = 7
AARHIL [5]	Attain lower lost rules	There is no improvement in terms of accuracy in AARHIL when compared with HCSRIL	Average lost rule = 4.20% Average Accuracy = 99.74% Average CPU-Time = 55 secs
MDSRRC [6]	Improves the speed of the mining process	Side effects due to reducing database needs to be reduced	Time Consumption (@Support =0.3) = 4 secs
Border Rule-based Distortion Algorithm [7]	Fewer Side effects	High CPU time	Mushroom Dataset: CPU Time (@5 set of sensitive rules) = 78.17 secs Bms-1 Dataset: CPU Time (@5 set of sensitive rules) = 12.19 secs Bms-2 Dataset: CPU Time (@5 set of sensitive rules) = 12.36 secs Chess Dataset: CPU Time (@5 set of sensitive rules) = 105.05 secs

Evolutionary Multi-Objective Optimization [8]	Less side effects	Density of dataset affects the performance of EMO	Mushroom Dataset (Sensitive Rule set=10):
			Missing = 2.574%
			Ghost = 0.155%
			Data loss = 23.331%
			Bms-1 Dataset (Sensitive Rule set=10):
			Missing = 1.312%
			Ghost = 0.161%
			Data loss = 1.548%
			Bms-2 Dataset (Sensitive Rule set=10):
			Missing = 3.447%
			Ghost = 0%
			Data loss = 3.201%
			Chess Dataset (Sensitive Rule set=10):
			Missing = 5.105%
			Ghost = 0%
			Data loss = 22.829%
			Retail Dataset (Sensitive Rule set=10):
			Missing = 0.453%
			Ghost = 0%
			Data loss = 2.395%

Fuzzy logic approach [9]	Decrease unnecessary side effect of sensitive rule hiding on non-sensitive rules	If any changes in membership function, then it causes some change in height of appropriate generalization	Brijis Dataset: Percentage of lost rule = 45% Clue Web Dataset: Percentage of lost rule = 62%

Whale optimization and Least Lion Optimization Algorithm [10]	Attains maximum privacy	Convergence speed of LLOA depends on the stopping criterion	Chess Dataset:
			Privacy = 84.36%
			Utility = 81.37%
			T10I4D100K Dataset:
			Privacy = 83.74%
			Utility = 83.96%
			Retail Dataset:
			Privacy = 82.76%
			Utility = 83.96%

MAXARH [11]	Maximize the identification of lost rules	Still MAXARH has side effects due to hiding association rules	Transactional Data: Misses Cost = 22% Dissimilarity = 5.3% Side Effect Factors = 24% Lost Rule Recovery = 89% Ghost Rule Generation = 5%
HGA and DIC [12]	Maintains the same cost of the transaction for both original and new databases	Artifactual error rate is high	Hiding Failure (@σ20 C40, 1k Datasets) = 0.7% Misses Cost (@σ20 C40, 1k Datasets) = 6% Artifactual Error (@σ 20 C40, 1k Datasets) = 6% Time (@σ20 C40, 1k Datasets) = 5754 secs
Cuckoo Optimization Algorithm [13]	Converged with high speed	Hiding failure is high when the number of iterations in Cuckoo Optimization Algorithm from 0 to 5 for chess dataset	Mushroom Dataset: Hiding Factor = 0% Ghost Rules= 0% Lost Rules = 0.5% Chess Dataset: Hiding Factor = 0% Ghost Rules= 0% Lost Rules = 0.17% Synthetic Dataset: Hiding Factor = 0% Ghost Rules= 0.005% Lost Rules = 0.16%

IV. CONCLUSION

In this article, a detailed comparative study on different association rule hiding techniques for privacy preservation is presented. From this comparative analysis, it is clearly noticed that the cuckoo optimization algorithm hides the sensitive association rules with satisfied performance. Among those techniques, a cuckoo optimization algorithm based association rule hiding has better performance. Even though, few limitations are addressed in cuckoo optimization algorithm based association rule hiding where in some point the hiding failure is high. Therefore, the future extension of this study could be focused on using different methods to improve the hiding failure of cuckoo optimization algorithm that further increases the efficiency of a cuckoo optimization algorithm based association rule hiding.

V. REFERENCES

[1]. Sathiyapriya, K., Sudhasadasivam, G., & Suganya, C. J. P. (2014). Hiding Sensitive Fuzzy Association Rules Using Weighted Item Grouping and Rank Based Correlated Rule Hiding Algorithm. WSEAS Transactions on Computers, 13, 78-89.

[2]. Shah, A., & Gulati, R. (2016). Privacy preserving data mining: Techniques classification and implications—A survey. Int. J. Comput. Appl., 137(12), 40-46.

[3]. Verykios, V. S. (2013). Association rule hiding methods. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3(1), 28-36.

[4]. Gulwani, P. (2012). A Novel Approach for Association Rule Hiding. International Journal of Advance Innovations, Thoughts & ideas, 1(3), 1-9.

[5]. Quoc Le, H., Arch-int, S., & Arch-int, N. (2013). Association rule hiding based on intersection lattice. Mathematical Problems in Engineering, 2013.

[6]. Ponde, P. R., & Jagade, S. M. (2014). Privacy Preserving by Hiding Association Rule Mining from Transaction Database. IOSR Journal of Computer Engineering (IOSR-JCE), 16(5), 25-31.

[7]. Cheng, P., Lee, I., Pan, J. S., Lin, C. W., & Roddick, J. F. (2015). Hide association rules with fewer side effects. IEICE TRANSACTIONS on Information and Systems, 98(10), 1788-1798.

[8]. Cheng, P., Lee, I., Lin, C. W., & Pan, J. S. (2016). Association rule hiding based on

evolutionary multi-objective optimization. Intelligent Data Analysis, 20(3), 495-514.

[9]. Afzali, G. A., & Mohammadi, S. (2017). Privacy preserving big data mining: association rule hiding using fuzzy logic approach. IET Information Security, 12(1), 15-24.

[10]. Menaga, D., & Revathi, S. (2018). Least lion optimisation algorithm (LLOA) based secret key generation for privacy preserving association rule hiding. IET Information Security, 12(4), 332-340.

[11]. Murthy, T. S., Gopalan, N. P., & Venkateswarlu, Y. (2018). An Efficient Method for Hiding Association Rules with Additional Parameter Metrics. International Journal of Pure and Applied Mathematics, 118(7), 285-290.

[12]. Mohan, S. V., & Angamuthu, T. (2018). Association Rule Hiding in Privacy Preserving Data Mining. International Journal of Information Security and Privacy (IJISP), 12(3), 141-163.

[13]. Afshari, M. H., Dehkordi, M. N., & Akbari, M. (2016). Association rule hiding using cuckoo optimization algorithm. Expert Systems with Applications, 64, 340-351.

[14]. Ponde, P. R., & Jagade, S. M. (2014). Privacy Preserving by Hiding Association Rule Mining from Transaction Database. IOSR Journal of Computer Engineering (IOSR-JCE), 16(5), 25-31.

Cite this article as :

G. Bhavani, Dr. S. Sivakumari, "Survey on Association Rule Hiding Techniques", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 6 Issue 3, pp. 300-305, May-June 2019. Available at doi : https://doi.org/10.32628/IJSRSET196368
Journal URL : http://ijsrset.com/IJSRSET196368