Volume 4, Issue 8, August 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Literature Survey of Privacy Preserving in Data Mining Techniques

Abhinav Kumar Sharma
Svits Indore, M.P.,
India

***Abstract*—In this paper, we provide here an overview of the new and rapidly emerging research area of privacy preserving data mining. Privacy preserving in data mining is a very popular research topic. A large number of researchers are working on improving security in data mining. Also a detailed review of the work accomplished in this area is also given along with the coordinates of each work to the classification hierarchy. The critical review of some modern data hiding approaches is also performed.**

***Keywords*— Data mining, Confidence, Support, Association Rule, Transaction.**

I. INTRODUCTION

Data mining is the non-trivial process of identifying valid and potentially useful patterns in data. Many governmental organization, businesses etc. are finding a way to collect, analyse and report data about individuals ,households or businesses, in order to support (short and long term) planning activities. Information system contains private or confidential information like their social security number, income of employees, purchasing of customer etc, that should be properly secured.

Privacy preserving data mining [9,18] is a new investigation in data mining and statistical databases [1]. In PPDM data mining algorithms are analysed for side effects obtain in data privacy. There is a twofold consideration in privacy preserving data mining. The first is sensitive raw data that are kept secure from unauthorized access like identifiers, names ,addresses should be modified from original database. The second one is sensitive knowledge is excluded that can be mined from a database by using data mining algorithms as such type of knowledge compromises data privacy.

The purchasing of one product when another product is purchased represents an association rule. Association rules are frequently used by retail stores to support in marketing, advertisement and inventory control. Let us suppose a grocery store chain keeping record of weekly transaction where each transaction represents the item bought during one cash register transaction. The executives of this chain recipient summarized report of transaction that indicated want type of item? Sold at what quantity? They periodically take information about which items are commonly purchased together. Then they find that 100% of time that peanut butter is purchased bread is also purchased. 33.3% of time peanut butter is purchased ,jelly is also purchased .Peanut butter is exist in only 50% of overall transaction [21].

II. LITERATURE SURVEY

This approach involves efficient fast and scalable algorithms that selectively sanitize a set of transactions from the original database to hide the sensitive association rules [1]. Various heuristic algorithms are based on mainly two techniques: (1) Data distortion technique (2) Blocking technique.

Data distortion is done by the alteration of an attribute value by a new value. It changes 1's to 0's or vice versa in selected transactions. There are two basic approaches for rule hiding in data distortion based technique: Reduce the confidence of rules and reduce the support of rules.

Consider sample database given in Table I. Selecting minimum support = 20% and minimum confidence = 80% and applying association rule mining algorithm, two association rules AB -> C (confidence = 100%) and BC-> A (confidence= 100%) are mined. Now suppose rule AB-> C is sensitive and needs to be hidden. Decreasing the confidence of a rule AB-> C can be done by either increasing the support of AB in transactions not supporting C (as shown in Table II) or by decreasing the support of C in transactions supporting both AB and C (as shown in Table III). Decreasing support of rule AB-> C can be done by decreasing the support of the corresponding large item set ABC (as shown in Table IV). The problem for finding an optimal sanitization to a database against association rule analysis has been proven to be NP-Hard [2]. In [3], authors presented three algorithms 1.a, 1.b and 2.a for hiding sensitive association rules. Algorithm 1.a hides association rules by increasing the support of the rule's antecedent until the rule confidence decreases below the minimum confidence threshold. Algorithm 1.b hides sensitive rules by decreasing the frequency of the consequent until either the confidence or the support of the rule is below the threshold. Algorithm 2.a decreases the support of the sensitive rules until either their confidence is below the minimum confidence threshold or their support is below the minimum support threshold. In 1.a algorithm large number of new frequent item sets is introduced and, therefore, an increasing number of new rules are generated. Algorithm 1.b and 2.a affects number of non-sensitive rules in database due to removal of items from transaction [3].

In [4] two algorithms are proposed ISL Increase Support of LHS and DSR Decrease Support of RHS. Predicting items are given as input for both algorithms to automatically hide sensitive association rules without pre-mining and selection of hidden rules. In [5] two algorithms DCIS Decrease Confidence by **I**ncrease **S**upport and DCDS Decrease Confidence by Decrease **S**upport are proposed to automatically hide collaborative recommendation association rules without pre-mining and selection of hidden rules. The ISL and DCIS algorithms try to increase the support of left hand side of the rule and algorithms DSR and DCDS try to decrease the support of the right hand side of the rule. It is observed that ISL requires more running time than DSR. Also both algorithm exhibit contrasting side effects. DSR algorithm shows no hiding failure (0%), few new rules (5%) and some lost rules (11%). ISL algorithm shows some hiding failure (12.9%), many new rules (33%) and no lost rule (0%). Algorithm DCIS requires more running time than DCDS. Similar to ISL and DSR algorithms DCIS and DCDS also exhibit contrasting side effects. DCDS algorithm shows no hiding failure (0%) few new rules (1%) and some lost rules (4%). DCIS algorithm shows no hiding failure (0%), many new rules (75%) and no lost rule (0%). In [6] an algorithm DSC (Decrease Support and Confidence) is proposed in which pattern-inversion tree is used to store related information so that only one scan of database is required. The proposed algorithm can automatically sanitize informative rule sets without pre-mining and selection of a class of rules under one database scan. There are about 4% of new rules generated and about 9% of rules are lost on the average for DSC algorithm and it also shows hiding failure for two predicting items.

Table 1. sample database

TID	Items
1	A, B, C
2	A, B, C
3	A, C
4	A, E
5	C, D

Rule	Confidence
AB→C	100%
BC→A	100%

Table 2. Hiding ab-> c by increasing support of ab

TID	Items
1	A, B, C
2	A, B, C
3	A, C
4	A, B, E
5	C, D

Rule	Confidence
AB→C	66%
BC→A	100%

Table 3. Hiding ab-> c by decreasing support of c

TID	Items
1	A, B
2	A, B, C
3	A, C
4	A, E
5	C, D

Rule	Confidence
AB→C	50%
BC→A	100%

Table 4. Hiding ab-> c by decreasing support of abc

TID	Items
1	A, C
2	A, B
3	A, C
4	A, E
5	C, D

Rule	Confidence
AB→C	0%
BC→A	0%

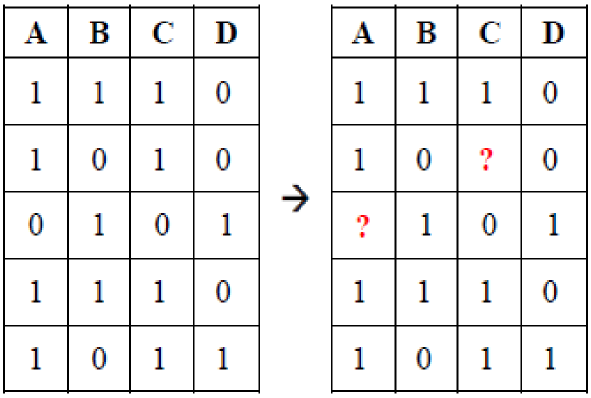
In [7] authors proposed an efficient algorithm, FHSAR (Fast Hiding Sensitive Association Rules) for fast hiding sensitive association rules. The algorithm can completely hide any given sensitive association rule by scanning database only once which significantly reduces the execution time. In this algorithm correlations between the sensitive association rules and each transaction in the original database are analyzed, which can effectively select the proper item to modify. In [9] four heuristic algorithms are proposed: Algorithm Naïve MinFIA (Minimum Frequency Item Algorithm) MaxFIA (Maximum Frequency Item Algorithm) and IGA (Item Grouping algorithm). Each algorithm selects the sensitive transactions to sanitize based on degree of conflict. Naive Algorithm removes all items of selected transaction except for the item with the highest frequency in the database. The MinFIA algorithm selects an item with the lowest support in the item set as a suspicious item and it removes the suspicious item from the sensitive transactions. Unlike the MinFIA & algorithm MaxFIA selects the item with the maximum support in the restrictive pattern as a victim item. Algorithm IGA groups restricted patterns in groups of patterns sharing the same item sets so that all sensitive patterns in the group will be hidden in one step. In [8] a heuristic algorithm named DSRRC (Decrease Support of R.H.S. item of Rule Clusters) is given. which provides privacy for sensitive rules at certain level while ensuring data quality. Proposed DSRRC algorithm clusters the sensitive association rules based on R.H.S. of rules and hides as many as possible rules at a time by modifying fewer transactions. Because of less modification in database it helps maintaining data quality. Algorithm DSRRC cannot hide rules having multiple RHS items. ***Blocking*** is the replacement of an existing value with a "?". This technique inserts unknown values in the data to fuzzify the rules. In some applications where publishing wrong data is not acceptable, then unknown values may be inserted to blur the rules. When unknown values are inserted support and confidence values would fall into a range instead of a fixed value. Consider the database shown in Table V. For rule A-> C, Support (A-> C) = 80% and Confidence (A-> C) = 100%. After fuzzifying the values, support and confidence becomes marginal. So in new database: 60% ≤ Confidence (A □-> C) ≤ 100% and 60% ≤ Support (A-> C) ≤ 80%.

The work done in [10] contains two algorithms are built based on blocking for rule hiding. The first one focuses on hiding the rules by reducing the minimum support of the item sets that generates these rules. The second algorithm focuses on reducing the minimum confidence of the sensitive rules. In [11] and [12] algorithms based on blocking technique are proposed and analyzed. In blocking technique the maximum confidence of a sensitive rule cannot be reduced. If the blocking algorithm does not add much uncertainty in the database, adversary can infer the hidden values if he applies a smart inference technique. In database both 0's and 1's must be hidden during blocking because if only 1's were hidden the adversary would simply replace all the ?'s with 1's and would restore easily the initial database and many ?'s must be inserted if we don't want an adversary to infer hidden data.

This approach hides sensitive association rule by modifying the borders in the lattice of the frequent and the infrequent item sets of the original database. The item sets which are at the position of the borderline separating the frequent and infrequent item sets forms the borders. The algorithms in this approach differ in the methodology they follow to enforce the new revised borders in the modified database. Border based approach uses the theory of borders presented in [13]. The first frequent item set hiding methodology that is based on the notion of the border is proposed in [14, 15]. It maintains the quality of database by greedily selecting the modifications with minimal side effect. Then in [16, 17] more efficient algorithms based on border theory are presented.

This approach contains non heuristic algorithms which formulates the hiding process as a constraints satisfaction problem or an optimization problem which is solved by integer programming. These algorithms can provide optimal hiding solution with ideally no side effects. In [18] an exact algorithm for association rule hiding is proposed which tries to minimize the distance between the original database and its sanitized version. In [19] proposed an exact border based approach to achieve optimal solution as compared to previous approaches.

Table 5. hiding a-> c by blocking



A	B	C	D
1	1	1	0
1	0	1	0
0	1	0	1
1	1	1	0
1	0	1	1

→

A	B	C	D
1	1	1	0
1	0	?	0
?	1	0	1
1	1	1	0
1	0	1	1

III. CONCLUSION

In this paper, we have presented a survey various privacy preserving data mining algorithms. The work presented in here, indicates the ever increasing interest of researchers in the area of securing sensitive data and knowledge from unauthorized users. The conclusions that we have reached from reviewing this area, manifest that privacy issues can be effectively considered only within the limits of specific data mining algorithms. The critical review, of privacy preserving in data mining techniques, done in this paper will help the researchers to overcome the drawbacks of existing algorithms and make an efficient algorithm.

ACKNOWLEDGEMENT

This research paper is made possible through the help and support from everyone, including: parents, teachers, family, friends, and in essence, all sentient beings. Especially, please allow me to dedicate my acknowledgment of gratitude toward the following significant advisors and contributors :

First and foremost, I would like to thank Mr Anand Rajavat and Mr Narendra Rathod (Department of CSE in SVITS Indore) for his most support and encouragement. He kindly read my paper and offered invaluable detailed advices on grammar, organization, and the theme of the paper.

Finally, I sincerely thank to my parents, family, and friends, who provide the advice and financial support. The product of this research paper would not be possible without all of them.

REFRENCES

[1] Aris Gkoulalas–Divanis;Vassilios S. Verykios “Association Rule Hiding For Data Mining” Springer, DOI 10.1007/978-1-4419-6569-1, Springer Science + Business Media, LLC 2010

[2] M. Atallah, E. Bertino, A. Elmagarmid, M. Ibrahim, and V. S. Verykios “Disclosure limitation of sensitive rules,”.*In Proc. of the 1999 IEEE Knowledge and Data Engineering Exchange Workshop (KDEX’99)*, pp. 45–52, 1999.

[3] Vassilios S. Verykios, A.K. Elmagarmid, E. Bertino, Y. Saygin, and E. Dasseni, “Association Rule Hiding,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 4, pp. 434-447, 2004.

[4] Shyue-Liang Wang; Bhavesh Parikh,; Ayat Jafari, “Hiding informative association rule sets”, ELSEVIER, Expert Systems with Applications 33 (2007) 316–323,2006

[5] Shyue-LiangWang *;Dipen Patel ;Ayat Jafari ;*Tzung-Pei Hong, “Hiding collaborative recommendation association rules”, Published online: 30 January 2007, Springer Science+Business Media, LLC 2007

[6] Shyue-Liang Wang; Rajeev Maskey; Ayat Jafari; Tzung-Pei Hong “ Efficient sanitization of informative association rules” ACM , Expert Systems with Applications: An International Journal, Volume 35, Issue 1-2, July, 2008

[7] Chih-Chia Weng; Shan-Tai Chen; Hung-Che Lo, “A Novel Algorithm for Completely Hiding Sensitive Association Rules”, IEEE Intelligent Systems Design and Applications, 2008.,vol 3, pp.202-208, 2008

[8] Modi, C.N.; Rao, U.P.; Patel, D.R., “Maintaining privacy and data quality in privacy preserving association rule mining”, IEEE 2008 Seventh International Conference on Machine Learning and Applications, pp 1-6, 2010

[9] Stanley R. M. Oliveira; Osmar R. Za¨_ane, “Privacy Preserving Frequent Itemset Mining”, IEEE International Conference on Data Mining Workshop on Privacy, Security, and Data Mining, Maebashi City, Japan. Conferences in Research and Practice in Information
Technology, Vol. 14.2002

[10] Y.Saygin, V. S. Verykios, and C. Clifton, “Using Unknowns to Prevent Discovery of Association Rules,” *ACM SIGMOD*, vol.30(4), pp. 45–54, Dec. 2001.

[11] Y. Saygin, V. S. Verykios, and A. K. Elmagarmid, “Privacy preserving association rule mining,” *In Proc. International Workshop on Research Issues in Data Engineering (RIDE 2002)*, 2002,pp. 151–163.

[12] E. Pontikakis, Y. Theodoridis, A. Tsitsonis, L. Chang, and V. S. Verykios. “A quantitative and qualitative analysis of blocking in association rule hiding”. In Proceedings of the 2004 ACM Workshop on Privacy in the Electronic Society (WPES), pages 29–30, 2004.

[13] H. Mannila and H. Toivonen, “Levelwise search and borders of theories in knowledge discovery”, *Data Mining and Knowledge Discovery*, vol.1(3), pp. 241–258, Sep. 1997.

[14] X. Sun and P. S. Yu. “A border–based approach for hiding sensitive frequent itemsets”, In Proceedings of the 5th IEEE International Conference on Data Mining (ICDM), pages 426– 433, 2005.

[15] X. Sun and P. S. Yu. Hiding sensitive frequent itemsets by a border–based approach. Computing science and engineering, 1(1):74–94, 2007.

[16] G. V. Moustakides and V. S. Verykios. A max–min approach for hiding frequent itemsets. In Workshops Proceedings of the 6th IEEE International Conference on Data Mining (ICDM), pages 502–506, 2006.

[17] G. V. Moustakides and V. S. Verykios. A maxmin approach for hiding frequent itemsets. Data and Knowledge Engineering, 65(1):75–89, 2008.

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

[20] Ila Chandrakar, Manasa, Usha Rani, and Renuka. *Hybrid Algorithm for Association Rule mining*. Journal of Computer Science 6(12), pages 1494-1498, 2010

[21] Belwal, Varsheney, Khan, Sharma, Bhattacharya. *Hiding sensitive association rules efficiently by introducing new variable hiding counter*. Pages 130-134, 978-2008, IEEE.