Review of Association Rule Hiding in Privacy Preserving Data Mining

Nidhi Gondalia[1] Sagar Kothadiya[2]
[1,2]Department of Computer Engineering
[1,2]Noble Group of Institutions, Junagadh, Gujarat, India- 362001

Abstract— Data mining is a popular analysis tool to extract information from collection of huge amount of data. The objective of this paper is to review an association rule hiding algorithm for privacy preserving data mining which would be used for providing confidentiality and improve the performance when the database stores and retrieves large amount of data. Association rule hiding which is one of the techniques of PPDM to protect the association rules generated by association rule mining. Association rule hiding refers to the process of changing the original database in such a way that certain sensitive association rules hide without seriously affecting the data and the non-sensitive rules. Association rule mining is an significant data-mining technique that finds riveting association among huge amount of data items.

Key words: Privacy Preserving Data Mining, Confidence, Support, association rules, Item sets

I. INTRODUCTION

Data mining is the knowledge discovery process which is extracting useful information from huge amounts of data. data mining is a knowledge discovery process to find patterns. discovered knowledge is specified in terms of decision tree, clusters or association rules. data mining has number of applications in marketing, engineering design, bioinformatics, scientific exploration, business, medical analysis, etc. for example, consider indian superstores like big bazaar and reliance store. suppose shopkeeper of reliance store mines the association rules related to big bazaar, where shopkeeper found that most of the customers who buy bread also buy butter. seeing this, shopkeeper of reliance store exploits this information and puts some discount on the cost of bread. this is how customers of big bazaar will now move to reliance store. this scenario leads to the research of sensitive knowledge rule hiding in database. therefore, before sharing the dataset to the other party, each supermarket is willing to hide sensitive association rules of its own sensitive products. so, the sensitive will be protected.

Privacy preserving data mining (PPDM) is measured to maintain the privacy of data and information extracted from data mining. Data mining allows the extraction of relevant knowledge and information from large amount of data, while protecting sensitive information. To preserve data privacy of information, one can change the original database in such a way that the sensitive information is excluded from the mining analysis and non-sensitive knowledge will be extracted. Association rule hiding to protect the sensitive association rules. The main aim of association rule hiding algorithms is to reduce the changes on original database to hide sensitive information, deriving non sensitive information and do not producing some other information.

II. ASSOCIATION RULE MINING

Association rule hiding refers to the process of changing the original database in such a way that certain sensitive association rules hide without affecting the data and the non-sensitive rules.

Let I = { i1,i2,i3,i4,…. in} which is a set of items. Let D which is a set of transactions or database. Each transaction t Є D is an item set such that t which is a proper subset of I. A transaction t supports X, a set of items in I, if X is a proper subset of t.

An association rule is an implication of the form X➔Y, where X is called antecedent and Y is called consequent, where X and Y which are subsets of I and X∩Y= Ø. The support of rule X➔Y can be calculated by this equation:

$$\text{Support}(X \rightarrow Y) = |X \cup Y| * 100 / |D|,$$

Where |X➔Y| represents the number of occurrences in the database that contains X U Y.
|D| represents the number of the occurrences in the database D.
The confidence of rule is calculated by this equation:

$$\text{Confidence}(X \rightarrow Y) = |X \cup Y| * 100 / |X|,$$

Where |X| is number of occurrences in database D that contains itemset X.

A rule X➔Y is strong if support(X➔Y) ≥ min_support and confidence(X➔Y) ≥ min_confidence, where min_support and min_confidence are two given minimum thresholds.

III. APPROACHES OF ASSOCIATION RULE HIDING

Many approaches are there to preserve privacy for sensitive information or sensitive association rules in database. These approaches can be classified into following approaches: heuristic based approaches, border based approaches, exact approaches, reconstruction based approaches, and cryptography based approaches.

A. Heuristic Based Approaches

These approaches can be divided into two types based on data distortion techniques and data blocking techniques.

1) Data Distortion Techniques:

There are two basic approaches for rule hiding in data distortion based technique. First one is reducing the confidence of rules and second is reducing the support of rules. Association rules can be hiding by decreasing or increasing support (or confidence). They replace 0-value to 1-value or 1-value to 0-value in selected transactions.

Verykios et al. [1] proposed five algorithms 1.a, 1.b, 2.a, 2.b, 2.c which is used to hide sensitive information of database by decreasing support or confidence of sensitive rule.Algorithms 1.a, 1.b, and 2.a which is used to hide rules and algorithms 2.b, 2.c which is used to hide large item sets.

Shyue-Liang Wang [2] proposed namely two algorithms which is Increase Support of LHS (ISL) and Decrease Support of RHS (DSR). The first algorithm

proposed for increase Support at Left hand Side to decrease Confidence of rule. The second algorithm proposed for decrease the support of the right hand side to decrease confidence of the rule.

Shyue-Liang Wang [3] proposed two algorithms which is namely DCIS (Decrease Confidence by Increase Support) which is used to increase support left hand side to decrease confidence for rule hiding and DCDS (Decrease Confidence by Decrease Support) which is used to decrease support to right hand side.

Data blocking techniques replace the existing value 0's and 1's by unknowns symbol ("?") in selected transaction instead of inserting or deleting items.

Y.Saygin et. al[4] proposed two algorithms which is used for blocking for rule hiding. first one used to hiding the rules by decreasing the minimum support of the item sets that generated these rules . The second used to decreasing the minimum confidence of the rules.

E. Pontikakis et.al[5] Proposed algorithm Reduce the minimum confidence of sensitive rules below and Do not reduce the minimum confidence of non-sensitive rules.

Border based approaches

X. Sun, and P. Yu [6] proposed algorithm used to changing the borders in the lattice of the frequent and the infrequent item sets of the actual database to hide sensitive association rule.The item sets at the position of the borderline lie between the frequent and infrequent item sets forms the borders. It uses the border of non-sensitive frequent item and compute the positive and negative borders in the item set. Then minimal affected modification has been selected. If modification has been done by greedy selection then it leads to minimum side effects.

G. V. Moustakides [7] The Max-Min approach proposed to hiding of the sensitive item sets while at the same time reduce the impact of the hiding process to the non-sensitive item sets. Effects achieved by the hiding process to the item sets on the positive border, after the border revised by taking into consideration the sensitive large item sets.

B. Exact Based Approaches

Exact approaches express hiding problem to constraint satisfaction problem (CSP) and solve it by using binary integer programming (BIP). They provide an optimal solution that satisfies all the constraints. However if exact solution does not exist in database, some of the constraints which are relaxed. Exact Approaches provide better solution than other approaches.

A. Gkoulalas-Divanis et. Al.[8] Proposed an algorithm for association rule hiding which tries to reduce the distance between the original database and its sanitized version.

A. Gkoulalas-Divanis et. Al.[9] proposed an exact border based approach to get optimal solution.

C. Reconstruction Based Approaches

First in this approach perturbing data and reconstructing the distributions at an aggregate level in order to perform the association rules mining. These approaches generate lesser side effects in database than heuristic approaches.

Mielikainen [10] has been analyzed the reckoning complexity of inverse frequent set mining and showed in many cases the problems are computationally difficult.

Y. Guo [11] proposed a FP tree based algorithm which is used to reconstruct the original database by using non characteristic of database and efficiently generates number of secure databases.

D. Cryptography Based Approaches

Cryptography based approaches which are used for multiparty computation, when database is distributed among several sites. Multiple parties wish to share their private data, without leaking any sensitive information at their end. This approach is classified as vertically partitioned distributed data and horizontally partitioned distributed data. In these approaches inplace of distorting the database, it encrypts original database itself for sharing.

Vaidya and Clifton[12] proposed a secure approach for sharing association rules when data are vertically partitioned. In terms of communication cost this approach is very effective and very expensive for large amount of datasets.

M. Kantarcioglu, and C. Clifton [13] proposed for the secure mining of association rules over horizontal partitioned data. This approach mines association rules securely with reasonable communication cost and computation cost.

IV. CONCLUSIONS

Association rule hiding is good concept in the area of privacy preserving data mining. It secures the privacy of sensitive knowledge in databases against the association rule mining approaches. In this paper, we have surveyed methods of hiding association rules by identifying some open challenges that will be useful to research community in privacy preserving data mining. It is found that finding an optimal solution for sanitizing database is NP-Hard. Existing approaches provide only the approximate solution to hide sensitive knowledge. There is need of finding feasible solution to the privacy problem in database disclosure. we strongly believe that the association rule hiding area will come into play in the evolution of other related fields in data mining and will cause new waves of research.

REFERENCES

[1] V.S. Verykios, A. Elmagarmid, E. Bertino, Y. Saygin, and E. Dasseni,"Association rule hiding," IEEE Transactions on Knowledge and Data Engineering, Vol. 16, No.4, 434–447,2004.

[2] Shyue-Liang Wang; Bhavesh Parikh,; Ayat Jafari, "Hiding informative association rule sets", ELSEVIER, Expert Systems with Applications 33 316–323,2006

[3] Shyue-LiangWang;Dipen Patel ;Ayat Jafari ;Tzung-Pei Hong,"Hiding collaborative recommendation association rules", Springer Science+Business Media, LLC 2007.

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

[5] 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.

[6] X. Sun, and P. Yu, "A Border-Based Approach for Hiding Sensitive Frequent Itemsets," In: Proc. Fifth IEEE Int'l. Conf. Data Mining (ICDM 2005), pp. 426–433,2005.

[7] G. V. Moustakides, and V. S. Verykios, "A Max-Min Approach for Hiding Frequent Itemssets," In: Proc. Sixth IEEE Int'l. Conf. Data Mining (ICDM 2006), pp. 502–506.

[8] Gkoulalas-Divanis and V.S. Verykios,"An Integer Programming Approach for Frequent Itemset Hiding," In Proc. ACM Conf. Information and Knowledge Management,2006.

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

[10] T. Mielikainen,"On inverse frequent set mining", In Proc. of 3rd IEEE ICDM Workshop on Privacy Preserving Data Mining. IEEE Computer Society, pp.18-23,2003.

[11] Y. Guo,"Reconstruction-Based Association Rule Hiding" In Proc. of SIGMOD2007 Ph.D. Workshop on Innovative Database Research 2007, pp.51-56,2007.

[12] J. Vaidya, and C. Clifton,"Privacy preserving association rule mining in vertically partitioned data," In proc. Int'l Conf. Knowledge Discovery and Data Mining, pp. 639–644,2002.

[13] M. Kantarcioglu, and C. Clifton,"Privacy-preserving distributed mining of association rules on horizontally partitioned data," IEEE Transactions on Knowledge and Data Engineering, Vol. 16, No. 9, pp. 1026-1037,2004.

