



A Heuristic Approach to Association Rule Mining

Saiyed Wafa Ahsan*

M.Tech student, Dept. of IT,
AIET, Lucknow, Uttar Pradesh,
India

R. K. Gupta

Asst. Professor of Dept. of CSE,
AIET, Lucknow, Uttar Pradesh,
India

Abstract— During data sharing, the protection of private data or sensitive information becomes critical. Privacy Preserving Data Mining (PPDM) is a research area that deals with the issue of violation of privacy during data mining. The main objective is to develop an algorithm to modify the original data such that the privacy is preserved even after the mining process. In this paper, we propose an algorithm that uses a heuristic approach to hide sensitive information or data to obtain a distorted database with minimum side effects such that private data remains private in a centralized database environment.

Keywords— association rule mining, sensitive itemsets, preserving private data, supporting transaction, side effects.

I. INTRODUCTION

Data mining and knowledge discovery in databases are two research areas that investigate the automatic extraction of previously unknown patterns in large volumes of data. However, these technologies can be a threat to data privacy. Privacy Preserving Data Mining (PPDM) is a novel research direction in Data mining and statistical databases, where data mining algorithms are analyzed for the side effects they incur in data privacy.

The main objective in PPDM is to develop algorithms for modifying the original data in some ways, so that there is no infringement of privacy constraint even after the mining process. However, the modifications will generate side effects.

Association rule data mining involves picking out the unknown inter-dependence of the data (or relationships) and finding out the rules between those items. The relationships can be represented in the form of frequent item sets or association rules. The rules can be characterized as sensitive or non sensitive depending upon the risk involved in their disclosure (based on some given threshold).

In this study, we suggest a heuristic approach to hide sensitive association rules that are specified by the users to incur minimum side effects (i.e., non sensitive rules falsely hidden and some spurious rules falsely generated). Firstly, the victim item is selected that needs to be hidden and then we find the appropriate supporting transactions in which the modifications are to be made such that we can obtain a sanitized database which when subjected to mining does not reveal the sensitive information.

Certain experiments are also conducted on synthetic dataset for both existing and proposed algorithm to check the algorithm for its effectiveness with reference to time and bulk of data.

II. PROPOSED ALGORITHM

Let the input be Database D and we have a set of user defined sensitive items SEN freq, we use apriori algorithm to generate frequent item set and store them in Dfreq with their support values. Non sensitive frequent item sets can be calculated and stored in NSEN freq.

- To identify the transaction required to update so as to incur minimum side effects, we need to calculate the weight of the transaction $W(T_i)$. The transaction with minimum weight is chosen first $W(T_i) = \text{no. of dependent transactions with victim item} - \text{no. of infrequent item sets associated with victim item}$.
- MinTrans is the number of transactions required to support an item set to be frequent.
- MinT is the set consisting of suitable number of transactions, which are to be modified to hide sensitive item set.

The split pattern technique is used if any item set in SENfreq has a length of more than two, so the pairs that are to be hidden for the purpose of sanitisation are identified using this technique. Thus, SENPAIRfreq is a vector consisting of all two pair sensitive items.

Now for every item set in SENPAIRfreq the following few steps are performed depending upon overlapping or non-overlapping patterns.

A. For Overlapping Patterns

Step 1 : Let (I,J)(J,K) be the sensitive item sets .

Selected item = J (overlapping item)

Step 2 : Find supporting transactions .

$$T(I,J,K) = T(I,J) \text{ AND } T(J,K)$$

Step 3: $C1 = \text{Support}(I,J) - \text{MinTrans} + 1$

$$C2 = \text{Support}(J,K) - \text{MinTrans} + 1$$

Step 4 : **if** $C1 > C2$

then Selected item values are replaced with 0 in C2 number of transactions from MinT

else

Selected item values are replaced with 0 in C1 number of transactions in MinT

Step 5 : **if** $|C1 - C2| \neq 0$

then **if** Dependency of I in NSEnfreq > Dependency of K in NSEnfreq

then K values are replaced with 0 in $|C1 - C2|$ number of transactions from MinT

else

J values are replaced with 0 in $|C1 - C2|$ number of transactions from MinT

step 6 : Remove (I,J),(J,K) from SENPAIRfreq .

B. For Non-Overlapping Patterns

Step 1 : Let (I,J) be the sensitive item set.

if Dependency of I in NSEnfreq > Dependency of J in NSEnfreq

then Selected item = J

else

Selected item = I

Step 2 : Find supporting transactions T(I,J).

Step 3 : $C = \text{Support}(I,J) - \text{MinTrans} + 1$

Step 4 : Selected item values are replaced with 0 in C number of transactions from MinT

Step 5 : Remove (I,J) from SENPAIRfreq.

These steps for overlapping and non overlapping patterns are to be repeated until no more pairs in SENPAIRfreq are left to hide. Thus, we obtain a Sanitised database D' where sensitive frequent item sets are hidden.

III. PERFORMANCE ANALYSIS OF THE PROPOSED METHODOLOGY

- The main aim of the proposed methodology is to find a distorted database efficiently in the situations such as when overlapping patterns exist in the sensitive items sets, when non – overlapping patterns exist in the sensitive item sets.
- Generally a sensitive item set may consists of single item or it may consists more than one item.
- In heuristic approach, the efficiency of the algorithm to hide the sensitive item sets can be measured in terms of number of modifications which are required to hide the sensitive item sets with minimum side effects.
- When sensitive item consists of more than one item, the selection of most promising item (victim item) whose value can be considered for distortion is a crucial decision to hide the sensitive item sets. The suggested approach in the proposed algorithm helps to find the victim item efficiently instead of random selection.
- Another important aspect is heuristic approach in the selection of supporting transaction by victim item for modification purpose. The weight concept specified above helps to find the suitable supporting transactions which are to be considered for modification purpose in order to hide sensitive item sets in the process of minimizing side effects.
- In the process of minimizing the number of modification to hide the sensitive item sets, in case of overlapping patterns which are exist in the sensitive item set, the suggested procedure efficiently finds the common victim item. This procedure helps to hide the overlapping pattern hidden at one step with minimum of modification over the victim item.
- For each modification over the victim item, the related overlapping patterns support will be decreased simultaneously and with this, the overall execution time to hide overlapping patterns with minimum number of changes is minimized by reducing the number of modifications.

The experimentations are conducted on synthetic dataset which consists of boolean transactions with 6 attributes. For evaluation purpose, the algorithm specified is considered. Experiments are conducted for both existing and proposed algorithm on synthetic dataset. The distinction of execution time for both existing and proposed methodology for various size databases are revealed in the following graph.

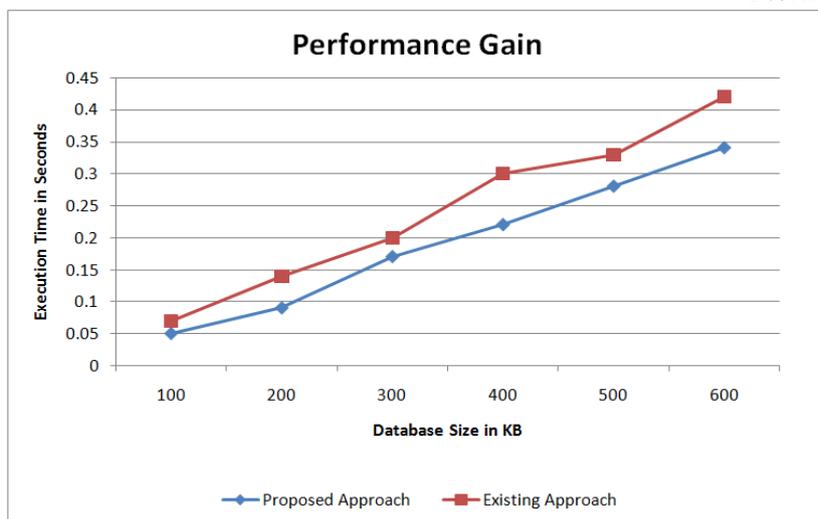


Fig 1: The performance of the proposed algorithm in case of non-overlapping patterns

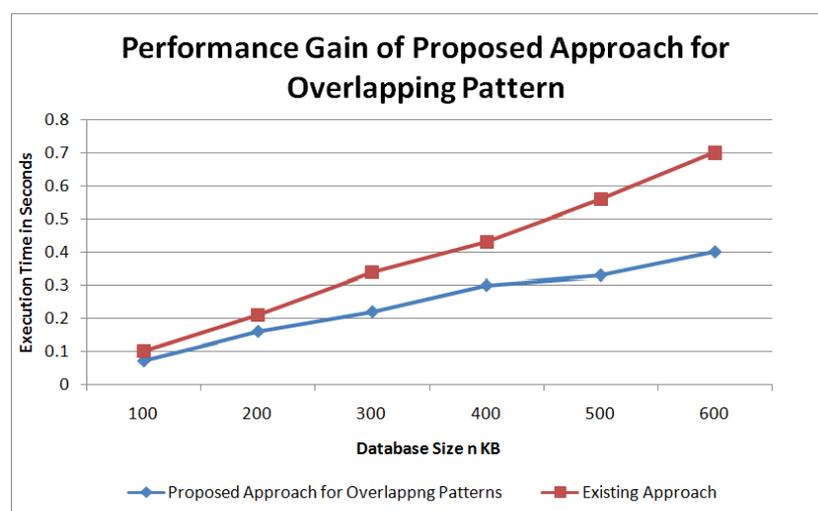


Fig 2: The performance of proposed algorithm in case of overlapping patterns.

IV. CONCLUSIONS

We have proposed an algorithm that hides the sensitive association rules with limited side effects. Thus obtaining a database which when subjected to data mining process preserves the privacy of data. It is also noted that the proposed algorithms effectiveness increases with the increase in the volumes of data in comparison with the existing approaches.

In addition, we have used split pattern technique to help accelerate the hiding process and to avoid the difficulty of forward inference attack by splitting all the sensitive item sets with length greater than two into pairs of sub pattern. From these pairs only significant pair sub patterns are hidden. To avoid forward inference attack problem, at least one such sub pattern with the length of two of the patterns should be hidden.

REFERENCES

- [1] C. Clifton, M. Kantarcioglu, and J. Vaidya. "Defining Privacy For Data Mining". In Proc. of the National Science Foundation Workshop on Next Generation Data Mining, pp 126-133, Baltimore, MD, USA, November 2002.
- [2] Oliveira and Zaiane Oliveira, S. R. M., Zaiane, O.R., "Towards Standardization in Privacy-Preserving Data", Mining, Edmonton, 2004.
- [3] Vassilios S. Verykios¹, Elisa Bertino², Igor Nai Fovino², "State-of-the-art in Privacy Preserving Data Mining", SIGMOD Record, Vol. 33, No. 1, March 2004.
- [4] Jaideep Vaidya, Chris Clifton, "Privacy-Preserving Data Mining: why, how and when", IEEE Security & Privacy, , 2004
- [5] Verykios, V.S., Bertino, E., Nai Fovino, I., Parasiliti, L., Saygin, Y., and Theodoridis, Y. "State-of-the-art in privacy preserving data mining", SIGMOD Record, SIGMOD Record, Vol. 33, No. 1, 50-57, March 2004.
- [6] Martin Meints and Jan Möller, "Privacy Preserving Data Mining: A Process Centric View from a European Perspective", 2008.

- [7] Elisa Bertino , Igor Nai Fovino Loredana Parasiliti Provenza ,”A Framework for Evaluating Privacy Preserving Data Mining Algorithms”, Data Mining and Knowledge Discovery, Vol.: 11, Issue: 2, pp 121–154, 2005,
- [8] Jian Wang ,Yongcheng Luo, Yan Zhao ,Jiajin Le, “A Survey on Privacy Preserving Data Mining”, 2009 First International Workshop on Database Technology and Applications, pp 11-114.
- [9] Vassilios S. Verykios and Aris Gkoulalas-Divanis, “A Survey of Association Rule Hiding Methods for Privacy Preserving Data Mining”, The Kluwer International Series on Advances in Database Systems, 2008, Vol 34 , 267- 289,2008.
- [10] Chirag Modi, U. P. Rao and Dhiren R. Patel, “A Survey on Preserving Privacy for Sensitive Association Rules in Databases”, Communications in Computer and Information Science, Vol 70, Information Processing and Management, pp 538-544, 2010.
- [11] Yongcheng Luo, Yan Zhao and Jiajin Le, “A Survey on the Privacy Preserving Algorithm of Association Rule Mining”, IEEE,2009.
- [12] R Agarwal, T Imielinski and A Swamy, “Mining Association Rules between Sets of Items in Large Databases”, Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, page 207-210, 1993.
- [13] Michael Goebel , Le Gruenwald, “A Survey Of Data Mining And Knowledge Discovery Software Tools”, SIGKDD Explorations, ACM SIGKDD, Vol:1, Issue 1, pp 20- 33, June 1999.