The International Journal Of Engineering And Science (IJES)
||Volume||2 ||Issue|| 9 ||Pages|| 15-18||2013||
ISSN(e): 2319 – 1813 ISSN(p): 2319 – 1805

HUISA and DSICA: Privacy Preserving Utility Pattern Mining

C.Saravanabhavan[1], R.M.S.Parvathi[2]

[1] Research Scholar &Asst Professor, Department of CSE, Kongunadu College of Engineering and Technology, Tamil Nadu, India.
[2]Principal & Professor, Department of CSE, Sengunthar College of Engineering, Tamil Nadu, India.

ABSTRACT

Privacy Preserving Data Mining (PPDM) has become a popular research area. How to balance between privacy protection and knowledge discovery in the sharing process is an important issue. This paper focuses on Privacy Preserving Utility Pattern Mining (PPUPM) and presents two algorithms, ***HUISA*** *and* ***DSICA****, to achieve the goal of hiding sensitive itemsets so that the adversaries cannot mine them from the modified database. In addition, we reduce the impact on the sanitized database in the process of hiding sensitive itemsets. The experimental results show that HUISA achieves the lower miss costs than DSICA does on two synthetic datasets. In the other hand, DSICA generally has the smaller deviation between the actual databases and sanitized databases than HUISA.*

KEY TERMS*: Data Mining, Utility Pattern Mining, Sensitive Database, Privacy.*

Date of Submission: 30, August, 2013 Date of Acceptance: 20, September 2013

I. INTRODUCTION

The association rule mining is one of the most important techniques in data mining. It discovers all the itemsets which support values are greater than a given threshold. These methods are used for discovering interesting relations between variables in large databases. Based on the concept of strong rules Rakesh Agarwal et al. introduced association rule for discovering regularities between the products in large-scale transaction. There are lots of algorithms proposed for discovering the frequent itemsets in literatures. The Apriori algorithm [1, 2, 13] is considered as the most famous one. In order to measure how “useful” an itemset is in the database, utility Pattern mining is proposed [22]. It overcomes the limitations of association rule mining, which ignores the sale quantity and price (or profitability) among items in a transaction.

On the other hand, Privacy Preserving Data Mining (PPDM) [20] becomes a popular research area in data mining in the past few years. In 1996, Cliftonetal. [5] analyzed that data mining can bring about threat against databases and addressed possible solutions to achieve privacy protection of data mining. In 2002, Rizvietal. discussed the privacy preserving mining of association rules[17,18]. However, to the best of our knowledge, there is less research contributions done in Privacy Preserving Utility Pattern Mining (PPUPM). Therefore, this study focuses on PPUPM and presents two novel algorithms, HUISA and DSICA, to achieve the goal of hiding sensitive itemsets so that the adversaries cannot mine them from the modified database. In addition, we reduce the impact on the sanitized database in the process of hiding sensitive itemsets. The rest of this paper is organized as follows. Sections 2 are reviewed related works. Then, Section 3 proposes the HUISA and DSICA algorithms to improve the balance between privacy protection and knowledge discovery. Section 4 presents the experimental results and evaluates the performance of the proposed algorithms. Finally, section 5 presents the conclusions of our work.

II. RELATED WORKS

2.1. Utility Pattern Mining

Utility pattern mining discovers all itemsets whose utility values are equal or greater than a user specified threshold in a transaction database. The challenge of utility pattern mining is in avoid the size of the candidate set and simplifying the computation for calculating the utility. Recently, Li et al. developed some efficient approaches, including the FSM, SuFSM, and DCG methods for share mining [8,9]. Under appropriate adjustments on item count and external utility of items, share mining is equivalent to utility pattern mining. In mean while, Liuetal.[12] also presented the Multi-Phase(MP)algorithm for discovering all high utility itemsets.

Table 1. An example of transaction database

(a) Transaction Tables

TID	A	B	C	D
T1	3	2	15	4
T2	0	6	3	5
T3	7	0	4	2
T4	1	7	0	3
T5	6	9	5	2
T6	7	2	8	1

(b) External Utility Table

ITEM	PROFIT($) per unit
A	3
B	2
C	5
D	4

Let $I=\{i_1,i_2,...,i_k\}$ be a set of items, where m is the total number of items. Let $DB=\{T_1,T_2,...,T_k\}$, the task-relevant database, be a set of transactions where each transaction T_q is a set of items, that is, $T_q \subseteq I$. A set of items is also referred as an *itemset*. An itemset that contains *k*-items is called a *k-itemset*.

- The *itemcount* of item $i_k \subseteq I$in transaction T_q, $c(i_k,T_q)$, is the number of item i_k purchased in transactionT_q .For example, $c(A,T_1)$=3, $c(B,T_1)$=20,and$c(C,T_1)$=15,inTable1(a).
- Each item i_k has an associated set of transactions $T_j=\{T_q \in DB | i_k \in T_q\}$.
- A *k*-itemset $X=\{x_1,x_2,...,x_k\}$ is a subset of *I*, where $1 \le k \le m$.
- Each *k*-itemset *X has* an associated set of transactions $T_X=\{T_q \varepsilon DB | X \subseteq T_q\}$.
- The utility of item $i_k \subseteq I$ in transaction tq, u(i_k, T_q) , is the quantitative measure of utility for item i_k in transaction T_q.

Utility Pattern mining is to find all the itemsets whose utility values are beyond a user specified threshold. An itemset X is a high utility itemset, if u(X) ≥ ε, ε is the ε is the minimum utility threshold.

2.2.Privacy Preserving Mining on Association Rules

The sanitizing algorithms for the privacy preserving mining on association rules can be divided into two categories :**(1)Data-Sharing approach** and **(2) Pattern-Sharing approach**

(1)Data-Sharing approach: The sanitization process acts on the data to remove or hide the group of restrictive association rules that contain sensitive knowledge. Among the algorithms of the data-sharing approach. They are classified the following sub-categories[18]. (a)Item Restriction-Based [21],(b) Item Addition-Based[21], and (c)Item Obfuscation-Based[19,20]. The *external utility* of item $ik \subseteq I$,$eu(i_k)$, is the value associated with itemi_k in the external utility table. This value reflects the importance of an item, which is independent of transactions. For example, in Table1(b),the external utility of item *A*,*eu*(*A*), is 3.

(2) Pattern-Sharing approach: the sanitizing algorithm acts on the rules mined from a database instead of the data itself. Regarding pattern-sharing techniques, the only known approach that falls into this category was introduced in [21].

Rule avoid-Based: This approach blocks some inference channels to ensure that an adversary cannot reconstruct restrictive rules from the non-restrictive ones. In doing so, we can reduce the inference channels and minimize the side effect.

III. PROPOSED ALGORITHMS

In this section, we present two algorithms for the privacy preserving utility mining:(1)Hiding Utility Itemset Algorithm (**HUISA**)and(2) Divergence Sensitive Itemsets Conflict Algorithm (**DSICA**).

3.1. Hiding Utility Itemset Algorithm (HUISA)

For each sensitive itemset, the sanitization process decreases the utility value of the sensitive itemset by modifying the quantity value of an item with the highest utility value in some transaction containing the sensitive itemset. The process repeats until the utility values of all sensitive itemsets are below the minimum utility threshold.

AlgorithmHUISA

Input: the original database*D*; the minimum utility threshold; the sensitive itemsets $U=\{S_1,S_2,\ldots,S_i\}$.

Output: the sanitized database *D*' so that S_i cannot be mined.

```
For each sensitive itemset $S_{i \mathcal{E} U}$
  $d=u(S_i)-\varepsilon$
  While ($d>0$) {
      $(ik, T_j)=\mathrm{argmax}(i\in si,\ sj\subseteq T)\ (u(i,T))$
    d={d-u(ik, Tj), if u(ik, Tj)<d
              0          ,if u(ik, Tj)>d
    return sanitized database D’
    End
```

3.2. Divergence Sensitive Itemsets Conflict Algorithm (DSICA)

To reduce the number of the modified items from the original database, DSICA selects an appropriate item which has the conflict among items in the sensitive itemsets. First calculate the I_{count}(U) and sort the i_k in decreasing order. Second it finds the transaction T_j such can be mined.

4. Experimental Results

The experiment was conducted with Pentium IV 3.2GHzPC with 2GB memory on the Linux platform. All algorithms were implemented in C/C++.We have conducted the experiments to study the effect of the execution time of our proposed model and given in table 2.

Table 2. Difference between *D* and *D’* for Supplier Transaction Data set

Min utility	3000	4000	5000
HIUSA	0.45%	0.92%	1.34%
DSICA	0.42%	0.86%	1.3%

IV. CONCLUSION

In our proposal, we present HIUSA and DSICA algorithms to reduce the issues on the original database for the Privacy Preserving Utility Pattern Mining. These algorithms are based on modifying the database transactions containing the sensitive itemsets so that the utility value can be reduced below the given threshold. There is no possible way to reconstruct the original database from the sanitized one. The experimental results show that HIUSA has the lower miss costs than DSICA does in two synthetic datasets. On the other hand, DSICA has the lower difference between the actual and sanitized databases than HIUSA.

REFERENCES

[1] R. Agrawal, T. Imielinski, and A. N. Swami, “Mining association rules between sets of items in large databases,” in *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*,pp.207-216,1993.

[2] R.Agrawaland R. Srikant, “Fast algorithms for mining association rules,”in *Proceedingsof 20th International Conference on Very Large Data Bases*,pp.487-499,1994.

[3] M.Atallah,E.Bertino,A.Elmagarmid,M.Ibrahim, andV.Verykios,“Disclosure limitation of sensitive rules,” in *Proceedings of IEEE Knowledge and Data Engineering Workshop*,pp.45-52,1999.

[4] R.Chan,Q.Yang,andY.D.Shen,“Mining high utility itemsets,”in *Proceedings of the 2003 IEEE International Conference on Data Mining (ICDM2003)*,pp.19-26,2003.

[5] C.Clifton and D.Marks,“Security and privacy implications of data mining,”in *Proceedings of the 1996 ACMSIGMOD Workshop on DataMining and Knowledge Discovery*,pp.15-19,1996.

[6] E.Dasseni,V.S.Verykios,A.K.Elmagarmid,and E.Bertino,“Hiding association rules by using confidence and support,” in *Proceedings of the 4th Information Hiding Workshop*,pp.369-383,2001.

[7] Y.C.Li,J.S.Yeh, and C. C. Chang, "Efficient algorithms for mining share-frequent itemsets," in *Proceedings of Fuzzy Logic, Soft Computing and Computational Intelligence -11thWorldCongressof International Fuzzy Systems Association (IFSA 2005)*, pp.534-539,2005.

[8] Y.C.Li,J.S.Yeh,andC.C.Chang,“A fast algorithm for mining share-frequent itemsets,”*Lecture Notes in Computer Science 3399*,pp.417-428,2005.

[9] Y.C.Li,J.S.Yeh,andC.C.Chang,“Direct candidates generation: a novel algorithm for discovering complete share-frequent itemsets,”*Lecture Notes in Artificial Intelligence 3614*, pp. 551-560,2005.

[10] Y.C.Li,J.S.Yeh,andC.C.Chang,"Isolated items discarding strategy for discovering high utility itemsets," *Data & Knowledge Engineering*,vol.64,no.1,pp.198-217,2008.

[11] Y.Liu,W.K.Liao,andA.Choudhary,“A two- phase algorithm for fast discovery of high utility itemsets, ”*Lecture Notes in Computer Science 3518 (PAKDD 2005)*,pp.689-695,2005.

[12] H.Mannila,H.Toivonen,andA.I. Verkamo, “Efficient algorithms for discovering association rules,” in Proceedings *of AAAI Workshop on Knowledge Discovery in Databases(KDD’94)*,pp181-192,1994.

[13] S.R.M.Oliveira and O.R.Zaïne,“A framework for enforcing privacy in mining frequent patterns,” Technical Report, TR02-13, Computer Science Department, University of Alberta, Canada, June 2000.

[14] S.R.M.Oliveira and O.R Zaïane,”Privacy preserving frequent itemset mining, “in *Proceedings of the IEEE ICDM Workshop on Privacy, Security, and Data Mining*,pp.43–54,2002.

[15] S.R.M.Oliveira,O.R.Zaïane,andY.Saygin, “Secure association rule sharing,”in *Proceedings of 8th Pacific-Asia Conference on Knowledge Discovery and Data Mining(PAKDD’04)*,pp.74-85,2004.

[16] S.Rizvi,andJ.Haritsa,“Maintaining data privacy in association rulemining,”in *Proceedingsof28thIntl. Conf. on Very Large Databases (VLDB)*,2002.

[17] S.J.RizviandJ.R.Haritsa,“Privacy-preserving association rule mining”,in *Proceedings of the 28th Int’l Conference on Very Large Databases*,2002.

[18] Y. Saygin,V. S. Verykios, and C.Clifton,“Using unknowns to prevent discovery of association rules,” *SIGMOD Record*,vol.30,no.4,pp.45-54,2001.

[19] V.Verykios,E. Bertino, I. G. Fovino, L. P. Provenza,Y.Saygin,and Y. Theodoridis, “State-of- the-art in privacy preserving data mining”,*SIGMOD Record*,vol.33,no.1,pp.50-57,2004.

[20] V.Verykios,A.Elmagarmid,E.Bertino, Y. Saygin, and E.Dasseni, “Association rules hiding,” *IEEE Transactions on Knowledge and Data Engineering*,vol.16,no.4,pp.434-447,2004.

[21] H.Yao,H.J.Hamilton,andC.J.Butz,“A foundational approach to mining itemset utilities from Databases,” in *Proceedings of the 4th SIAM InternationalConferenceonDataMining*,pp.482-486,2004.

[22] Z.Wang,W.Wang,B.Shi, and S. H. Boey, "Preserving private knowledge in frequent pattern mining,"in *Proceedings of 6th IEEE International Conference on Data Mining -Workshops(ICDMW'06)*, pp.530-534,2006

[23] Z. Wang, W. Wang, and B. Shi, "Blocking inference channels in frequent pattern sharing," in 53–87, 2004.

Mr.C.Saravanabhavan received his M.C.A degree from K.S.R. College of Technology, Namakkal in 2003 and M.Tech. degree in Information Technology from Sathyabama University, Chennai in 2007. He is currently pursuing Ph.D., in Computer Science and Engineering from Anna University of Technology, Coimbatore. Presently he is working as an Assistant Professor in Kongunadu College of Engineering and Technology, Trichy,Tamilnadu, India.His research areas of interest are Data Mining and Networking.

Dr.R.M.S.Parvathi has completed her Ph.D., degree in Computer Science and Engineering in 2005 in Bharathiyar University, Tamilnadu, India.Currently she is a Principal and Professor, Department of Computer Science and Engineering in Sengunthar College of Engineering, Tamilnadu, India, She has completed 22 years of teaching service. She has published more than 28 articles in International / National Journals. She has authorized 3 books with reputed publishers. She is guiding 20 Research scholars. Her research areas of interest are Software Engineering, Data Mining, Knowledge Engineering, and Object Oriented System Design.