ISSN:0975-9646

Sensitive Itemset Hiding in Multi-level Association Rule Mining

#P.RajyaLakshmi, #C.Mohan Rao, **Madhavi Dabbiru and *Kopparthy Vinay Kumar
#Avanthi college of Engineering, Visakhapatnam, A.P
***Pydah College of Engineering & Technology, Visakhapatnam, A.P*
**Godavari Institute of Engineering & Technology, Rajahmundry, A.P.*

Abstract--- **Enormous numbers of intelligent data mining techniques are in usage to discover hidden patterns. Especially Association rule mining has a high impact on business improvement. However mining association rules at multiple-level may lead to discovery of more specific and concrete knowledge from data. Privacy is needed in order to withstand the business competence. Now-a-days privacy preserving data mining is an important research area. In this paper we propose to apply sensitive itemset hiding algorithm [3] on multi level association Rule mining results, such that frequent items information is retained while sensitive items are hidden.**

Keywords: **Association rule discovery system, Privacy preserving data mining, Multi-level association rules mining.**

I. INTRODUCTION

Association rule mining is a crucial task in the area of data mining. Many applications at mining associations require that mining be performed at multiple levels of abstraction. Applying different minimum thresholds, for mining frequent itemsets at different levels of abstraction, reduce the minimum thresholds at lower levels of abstraction. This leads to mining interesting frequent itemsets at multiple concept levels, which may not only discover rules at different levels, but may also have high potential to find nontrivial, informative association rules because of its flexibility for focusing the attention to different sets of data and applying different thresholds at different levels. For example, besides finding 60% of customers that purchase computer may also purchase printer, it is interesting to allow business persons to drill-down and show that 50% of people buy HP printer if they buy HP computer. The association relationship in the latter statement is expressed at a lower level of abstraction but carries more specific and concrete information than that in the former. Therefore, a data mining system should provide efficient methods for mining multiple-level association rules. Advances in data collection, processing, and analysis, along with privacy concerns regarding the misuse of the induced knowledge from this data, soon brought into existence the field of privacy preserving data mining [8]. Simple de-identification of the data prior to its mining is insufficient to guarantee a privacy-aware outcome since intelligent analysis of the data, through inference based attacks, may reveal sensitive patterns that were unknown to the database owner before mining the data. Thus, compliance to privacy regulations requires the incorporation of advanced and sophisticated solutions. This paper concentrates on a subfield of privacy preserving data mining, known as "knowledge hiding."

The remainder of this paper is organized as follows: Section 2 reviews related work in the fields of sensitive itemset hiding. Section 3 discusses the methodology with an example. Section 4 provides the algorithm. Section 5 concludes this paper followed by references.

II. RELATED WORK

In what follows, we review some of the fundamental related work in both research directions. Yinbo WAN,et al. Mining Multilevel Association Rules From Primitive Frequent Itemsets 2006 which improves the mining efficiency without making the additional deviations to the FP(*l*)-tree, while realizing the mining of cross level association rules. Further more, this approach can support dynamic hierarchies based on different views of organizing items, which allow different users to get their desired association rules from a customized point of view. L.K. Sharma1 et.al., proposed a novel approach of mining spatial positive and negative association rules. The approach applies multiple level spatial mining methods to extract interesting patterns in spatial and/or non-spatial predicates. A pruning strategy is used in their approach to efficiently reduce the search space. The first work to address the problem was presented in [8], where the authors proposed a greedy algorithm for selecting items to sanitize among the transactions supporting sensitive itemsets. Menon et al. [7] present an integer programming approach for the hiding of sensitive item sets. The algorithm treats the hiding process as a CSP that identifies the minimum number of transactions to be sanitized. The authors first reduce the size of the CSP by using constraints involving only the sensitive item sets and then solve it by using integer programming. A heuristic is then enforced to identify the actual transactions and sanitize them.

To the best of our knowledge, apart from ongoing research work regarding an additive model for sensitive item set hiding [5], this approach facilitates knowledge hiding in multi level databases. Extending the original database to accommodate knowledge hiding can be considered as a bridging between the itemset hiding and the synthetic database generation approaches.

III. HIDING ALGORITHM

Terminology:

- I – Set of Items, Transaction – A nonempty subset of I, uniquely identified by a TID, Transaction t contains an itemset X if X is a subset of t.
- Database (DB) – A nonempty sequence of transactions.
- Support of an itemset X – Fraction of transactions in DB that contain X.

- Given 0 < minsup <=1, all itemsets X that have support atleast minsup in DB are called FREQUENT ITEMSETS or LARGE ITEMSETS.

*Maximal frequent itemset***:** A frequent itemset is called maximal if it is not a subset of any other frequent itemset.
Minimal infrequent itemsets:
MIF= {I ЄIF/ for all XЄ IF there does not exist XↄI}
Where IF is the set of all infrequent itemsets and MIF is the set of all minimal infrequent itemsets.
Database Extension Size (Q)
Q= | [sup(Im, Do) / mfreq-N]+1 |
Where sup (Im, Do) is the maximum frequency of sensitive item in original database Do, N is the number of records in Do and mfreq is the user given minimum support threshold.
Input: Sensitive items
Output: Hiding sensitive items
Process:

1. Find Maximal frequent itemsets using any one of the popular Max algorithms.[A]
2. Find Minimal infrequent itemsets. [AIN].
3. From one and two identity itemsets containing sensitive itemsets and remove them. The new sets are now [AS^-] and [$AINS^-$].
4. Now the original transaction database size is to be extended in order to hide the sensitive items in the database using EXTEND algorithm.
5. Now calculate the new minimum support for the extended database.
6. Repeat step1 to ensure whether sensitive items are hidden or not.

EXTEND Algorithm:
Step 1: 1. Find the support counts of all the sensitive items.
Step 2: Find the size of the extension for the original database in order to make the sensitive itemsets infrequent

a. Find the Maximum support count among all sensitive items
b. The min support threshold should be chosen such that it dominates the Maximum frequency among sensitive itemsets.
c. Basing on the newly obtained minimum support and original database min support calculate the size of the additional transactions to be added.

Step 3: Now in order to fill the newly added transactions with items. Select 75% of items from [AS^-] and 25% of items from [$AINS^-$].
Step 4: END.

IV. APPLYING SENSITIVE ITEMS HIDING ON MULTI LEVEL ARM.

Procedure:
Step 1: Find maximal frequent itemsets at each level with a different minimum support threshold at each level and store them in a set MaxMultiLevel.
Step 2: Remove hierarchical redundancy[ ] from MaxMultiLevel set and store in NewMax.
Step 3: Now apply the sensitive itemset hiding algorithm developed by the authors on NewMax.
Step 4: The database is updated with the new extended database which hides frequent sensitive itremsets.

Example: consider a simple multi-level data set as shown in figure "1".the dataset has three levels with each item belonging to the lowest level. Thus the item 221 can be decoded as follows: The first digit '2' represents 'bread' at level 1,the second '2' for 'wheat' at level2 and the third, '1' for the brand 'freshchoice' at level-3. Repeated items at any level will be treated as one item in one transaction.

TID	ITEM LIST
1	{111,121,211,221}
2	{111,211,222,323}
3	{112,122,221,411}
4	{111,121}
5	{111,122,211,221,413}
6	{113,323,524}
7	{131,231}
8	{323,411,524,713}

Figure-1 Transactional database

From this transactional database, we apply the MLT2L1 algorithm with the cross level add on [4] and a minimum support threshold of 4 for level 1 and 3 for level 2 etc.,.figure '2' shows all frequent itemsets derived from all three levels.

1 –ITEMSETS	2-ITEMSETS	3-ITEMSETS
{1**}	{1**,2**}	{1**,21*,22*}
{2**}	{1**,21*}	{2**,11*,12*}
{11*}	{1**,22*}	{11*,12*,22*}
{12*}	{2**,11*}	{11*,21*,22*}
{21*}	{2**,12*}	{1**,21*,22*}
{22*}	{11*,12*}	{11*,211,22*}
{111}	{11*,21*}	{11*,221,12*}
{211}	{11*,22*}	{21*,111,22*}
{221}	{12*,22*}	{22*,111,211}
	{21*,22*}	
	{1**,211}	
	{1**,221}	
	{2**,111}	
	{11*,211}	
	{11*,221}	
	{12*,111}	
	{12*,221}	
	{21*,111}	
	{22*,111}	
	{22*,211}	
	{111,211}	

Figure:-2 All frequent itemsets

Now, we apply a maximal frequent items algorithm and derive maximal frequent itemsets as shown in figure '3'

[1**,2**], [1**,21*, 22*]
[1**, 211], [2**, 11*, 12*]
[1**, 221], [11*, 12*, 22*]
[2**, 111], [11*, 21*, 22*]
[12*, 111], [11*, 211, 22*]
[11*, 221, 12*]
[21*, 111, 22*]
[22*, 111, 211]

Figure 3 Maximal Frequent itemsets

However some hierarchical redundancy exits in the maximal frequent Itemsets. For example the item 111 is a child of the more general item 11*. After eliminating hierarchical redundancy the frequent items are tabulated as shown figure 4.

[1**, 2**], [12*, 111],
[1**, 21*, 22*], [2**, 11*, 12*]
[11*, 12*, 22*], [11*, 21*, 22*]

Figure: 4 After removal of hierarchical redundancy

Set of sensitive items = {2**} Now apply the sensitive hiding algorithm stated in previous section. Eliminate all itemsets containing items starting with '2'
Revised Maximal frequent itemsets = {[12*, 111]}
Minimal infrequent itemsets = {323, 411, 413, 524, 131, 713}
Now minimum support =6. [Since maximum support count of sensitive item is 5]
Data base Extension size = |[5/4-8]+1|=6.
New Data base size =8+6=14
After extension, the extended data base hides all the sensitive itemsets and is shown in the figure '5'.
Thus we perform data compression and reduce the size of the output of multilevel transactional data and at the same time we hide the sensitive itemsets.

TID	ITEM LIST
1	111 121 211 221
2	111 323 211 222
3	112 122 411 221
4	111 121
5	111 122 413 211 221
6	113 323 524
7	131 231
8	323 411 524 713
9	111 12*
10	131 12*
11	131 411
12	111 12*
13	713 524
14	111 12*

Figure-5 Extended database

V. RESULTS

All of our experiments are performed on a 2.00GHz, 1 GB memory Intel PC, running Windows XP. We implemented the algorithm in dot Net. We collected data from a local store and successfully implemented the algorithm. The performance of the algorithm is shown in figure 6 for various sizes of databases.

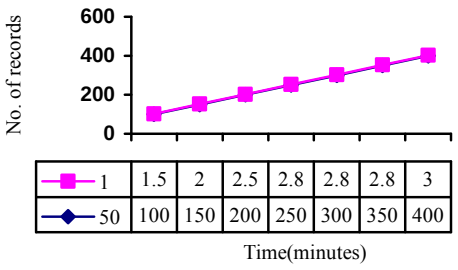

Figure: 6 Performance of the hiding algorithm

VI. CONCLUSIONS

In this paper, we have presented a novel approach to sensitive knowledge hiding, through the introduction of a minimal extension to the original database. The proposed methodology is capable of identifying an ideal solution whenever one exists, or approximate the exact solution, otherwise. We also propose to apply sensitive itemset hiding algorithm [3] on multi level association Rule mining results, such that frequent items information is retained while sensitive items are hidden.

REFERENCES

[1] A Novel Approach of Multilevel Positive and Negative Association Rule Mining for Spatial Databases L.K. Sharma1, O.P. Vyas1, U.S. Tiwary2, and R. Vyas1pp. 620 – 629, 2005. Springer-Verlag Berlin Heidelberg 2005
[2] M.L. Antonie and O.R. Za¨ıane, "Mining Positive and Negative Association
Rules: an Approach for Confined Rules", Proc. Intl. Conf. on Principles and Practice of Knowledge Discovery in Databases, 2004, pp 27–38.
[3] RajyaLakshmi.P , Dr.C.Mohan Rao , D. "A Novel Algorithm for Hiding Sensitive itemsets" in "National Seminar on Corporate e-Governance". Organized by Computer Society of India- Visakhapatnam Chapter (eCOG-2011).
[4] Jiawei Han and Yongjian Fu, "Mining Multiple Level Association Rules in Large Databases", Proc. Intl. Conf. on Principles and Practice of Knowledge Discovery in Databases, 2000.
[5] A G Divanis and VS Verykios, "Exact Knowledge Hiding through Database Extension", IEEE TKDE Vol 21, No 5, May 2009.
[6] M.J. Zaki and C.J. Hsiao, "CHARM: an Efficient Algorithm for Closed Itemset Mining", Proc. Second SIAM Intl. Conf. on Data Mining, Arlington, VA, 2002, pp 12–28.
[7] Menon et al. "An integer programming approach for the hiding of sensitive item sets".
[8]C.C Aggarwal and P.S. YU, "Privacy preserving Data Mining:Models and algorithms", Springer-Verlag, 2008

P Rajyalaxmi is studying M.Tech, in CSE, Avanthi College of Engg & Tech, Tamaram, Visakhapatnam,A.P., India. she has received her B.Tech (CSE) from Andhra University, Visakhapatnam A.P., INDIA.

Dr. C. Mohan Rao, received M.Tech in Computer Science and Technology from Andhra University College of Engineering and awarded Ph.D by Andhra University during 2000. He has 17 years teaching and research experience and guided number of M.Tech and MCA Students for their projects. He has published 17 papers in National and International Journals. He received " Best Teacher award" from JNTU, Kakinada during 2009.

D.Madhavi received her M.Sc Degree in Computer Science from P.B. Siddhartha college, Vijayawada and M.E. Degree in Computer Engineering with distinction from Andhra University. She is presently working as an Associate Professor in the department of Computer Science and Engg, Pydah College of Engineering and Technology, Visakhapatnam, Andhra Pradesh, India. She submitted her Ph.D thesis recently. Her areas of interest include Data Mining, Information Retrieval, and Data base systems. She received " Best Teacher award" from JNTU, Kakinada.

K.Vinay Kumar is studying M.Tech, in CSE, Godavari Institute of Engineering & Technology, Rajahmundry, A.P., India. He has received her B.Tech (IT) from JNTU,Hyderabad,A.P.