Hybrid Conflict Ratio for Hiding Sensitive Patterns with Minimum Information Loss

R.R.Rajalaxmi and A.M.Natarajan

Abstract— **Information sharing among the organizations promotes business growth. Recent trends in data mining techniques impose an intimidation to data sharing. A key problem here is the need to balance the privacy of the data with the genuine need for the users. To address this problem, data sanitization process modifies the original data to conceal sensitive knowledge before release. Several researchers addressed the privacy preservation of sensitive knowledge in the form of association rules by suppressing the frequent itemsets. This paper proposes an effective data sanitization algorithm which minimizes the side effects caused in the original database. A hybrid conflict ratio approach is proposed to pick the victim transactions and items to minimize the legitimate lost during sanitization. To study the effectiveness of the algorithm, experimental analysis is carried over the real and synthetic datasets. The results illustrate that the algorithm show good results compared with the other approaches.**

Index Terms— **Data Sanitization, Association Rule mining, Privacy Preserving Data Mining, Conflict Ratio.**

I. INTRODUCTION

Many organizations widely adopted different ways to collect large volume of data. Such data is processed to yield useful information [4]. However, the data analysis techniques do not reveal hidden knowledge. Subsequently, a novel research area emerged which focus primarily on extracting meaningful and previously unknown patterns from these large volumes of data. Data mining is often viewed as a step in the knowledge discovery process to explore hidden knowledge.

Due to the increase in use of data mining, it may pose a threat to data privacy and security [13]. However, it is important to note that most of the data mining applications do not even touch personal data. Prominent examples include applications involving natural resources, meteorology, astronomy, geology and other scientific and engineering data. The focus of data mining technology is on the discovery of general patterns, not on specific information regarding individuals. In this sense, the real privacy concerns are with unconstrained access of individual records, like credit card and banking applications. For those applications that do not involve personal data, in many cases, simple methods such as removing sensitive IDs for data may protect the privacy of individuals.

Recent trends in business demands the need to adopt collaborative business model which ensures a new way to achieve mutual benefit among the organizations. Data sharing enables to achieve this. Even though there are potential benefits of data sharing, it could reveal confidential information [17] which could be detrimental to the data owner. The latent risks with data sharing can be significant, as highlighted by Wal-Mart’s decision to stop selling its point-of-sale data to market-research companies [14]. Collaboration requires trust and a leap of faith that once customers get a good look inside the business, they will like what they see. Hence, the organizations may not want to share some data which can reveal sensitive patterns [9][15][22] with other parties. It is beneficial to share data without revealing sensitive patterns, which makes a trust among the organizations for collaboration.

In order to preserve the privacy of confidential data, they would like to transform their data in such a way that these sensitive patterns cannot be discovered but others can be. Data sanitization process was introduced [20] which legitimately modify the original data to seek a balance between sensitive knowledge protection and non-sensitive knowledge discovery. Privacy preservation in data mining is the emerging research area which addresses the different ways to protect the sensitive knowledge discovery [2] [21]. It is mandatory to gain the benefit of data mining and as well as maintaining privacy. This paper focuses on the privacy preservation in frequent pattern mining. The rest of the paper is organized as follows: Section 2 discusses the background and related work in privacy preserving frequent pattern mining. In section 3 we describe the problem to be solved and the proposed approach for privacy preservation in frequent pattern mining. Section 4 discusses the various experimental results carried and the detailed discussion on the analysis of results. Finally we conclude the paper in Section 5.

II. Background and Related Work

A. *Association Rule Mining*

Let I = {i1, i2, i3, …, in} be a set of items. Let D, the task-relevant data, be a set of database transactions where each transaction T is a set of items such that $T \subseteq I$. Each transaction is associated with an identifier, called TID. Let A be a set of items. A transaction T is said to contain A if and only if $A \subseteq T$. A set of items is referred to as an itemset. An itemset that contains k items is a k-itemset. The occurrence frequency or support of an itemset is the number of

Manuscript received May 20, 2009. R.R.Rajalaxmi is with Computer Science and Engineering Department, Kongu Engineering College, Erode, TamilNadu, India(phone no : 04294-226580)

A.M.Natarajan is CEO of Bannari Amman Institute of Technology, Sathtyamangalam, Erode, TamilNadu, India.

.

transactions that contain the itemset [11]. If the relative support of an itemset I satisfies a prespecified minimum support threshold, then I is a frequent itemset. Several algorithms have been proposed to efficiently discover the frequent patterns.Level-wise algorithms generate the frequent patterns which generate candidate itemsets [6][12], where as pattern-growth algorithms does not generate candidate itemsets[16][18].

B. Privacy preserving frequent pattern mining

The problem of privacy preserving frequent pattern mining can be defined as follows:

For the given database D, frequent patterns P and sensitive patterns Ps, the main goal of sanitization process is to transform D into a modified database D’ such that the sensitive patterns are hidden in D’ with minimal side effects in D.

The potential threats of data mining are analyzed [5] and some possible approaches to prevent the discovery of sensitive knowledge in a data mining context are suggested. The proposed solutions include, limit the access to the data by releasing only samples, fuzzyfying and augmenting the source database. In [7] the last approach is adopted. The author has suggested a solution to determine the sample size in such a way that data mining tools cannot obtain reliable results. The solution proposed is independent from any specific data mining technique.

Data sanitization is the process of modifying the original database to preserve the privacy of sensitive knowledge [10]. Preserving the privacy of sensitive rules [3] which restrict the disclosure of sensitive rules was performed by decreasing the support of their corresponding frequent itemsets. The problem of sanitization is NP-hard and heuristic approaches can be used to provide a solution. Based on this, Atallah et al. proposed a sanitization heuristic to hide the sensitive frequent itemsets. On the other hand, privacy preservation can be performed by perturbing the data and reconstructing the orginal distributions at an aggregate level in order to perform mining. This method enables to retain privacy while accessing the information implicit in the original attributes [1]. Data perturbation protects individual confidential data values while data sanitization protects sensitive knowledge hidden in the database in the form of sensitive association rules.

Rizvi et al. [19] addressed the privacy preservation of sensitive rules by distorting the user data before it is subject to the mining process. The authors present [21] three strategies and five algorithms for hiding a group of association rules, which is characterized as sensitive. Deterministic algorithms [23] were proposed to hide sensitive association rules. A new framework for rule hiding is proposed in [11]. Stanely et al. [18] presented algorithms for hiding sensitive patterns in the form of frequent itemsets. Hiding the sensitive rules with database modification may produce side effects which must be controlled so that the legitimate patterns are not lost in mining. To address this [23] proposed an approach to hide sensitive association rules with limited side effects. Item conflict degree helps to minimize the non-sensitive patterns lost during sanitization [24].The conflict degree of an item with the other sensitive itemsets is considered to choose a victim item for deletion. But it does not consider the number of legitimate itemsets affected when a victim item is selected for deletion. Here we introduce the concept of conflict ratio of transaction and an item which pays attention towards the legitimate itemsets affected during the sanitization. The proposed approach shows improved results in terms of misses cost with minimal changes in the database.

III. Methodology

Given a set of sensitive patterns and original database, the main goal of the sanitization process is to modify the original database such that minimal numbers of non-sensitive patterns are affected. In general, the process of sanitization involves the following steps.

1. Identify the list of transactions which support the sensitive itemsets.
2. Select the partial list of transactions for sanitization
3. Choose the victim item in the transactions and delete
4. Rewrite the modified transactions in the original database to get the sanitized database.

Steps 1 do not produce any significant change in sanitization. Steps 2 and 3 are the crucial part which plays a vital role in selecting the victim transaction and item for deletion which introduces side effects in the database.

Definition 1 : The transaction conflict ratio (TCR(t)) of a transaction t $\in$ D is defined as follows.

$$TCR(t) = \# Ps \div \# \sim Ps$$

where
$\# P_s$ - the number of sensitive itemsets supported by t
$\# \sim P_s$ - the number of non-sensitive itemsets supported by t

Definition 2: Let P_s be the set sensitive itemset in P .The set of non-senstive itemsets is denoted as $\sim P_s$ (=P- P_s).

Defintion 3. The item conflict ratio (ICRi) of an item i in a sensitive transaction t is defined as follows.

$$ICRi(t) = \# Psi \div \# \sim Psi$$

where
$\# P_s$ - the number of sensitive itemsets supported by i in t
$\# \sim P_s$ - the number of non-sensitive itemsets supported by i in t

Algorithm HCRS

Input

Source database D
Set of sensitive itemsets P_s
Set of non-sensitive itemsets $\sim P_s$
Disclosure threshold δ

Output

Modified database D’

Steps

```
n  Sort $P_s$ in ascending order of support
n  for each $s_i$ in $P_s$
n  extract the transactions $Ts_i$ from D supporting $s_i$
n  Compute Tcr($Ts_i$) and sort in descending order
n  Ntrans = $|Ts_i| * (1-\delta)$
n  for k = 1 to Ntrans
n  for each item j,m $\in s_i$
n    compute ICRj($T_p$)
n         if ICRj($T_p$) > ICRm($T_p$)
n              victim item = i
```

```
n       else
n           victim item = m
n     delete victim item from k
n   for each transaction t in D
n       if t is modified as t' then
n           write t' to D'
n        else
n            write t to D'
```

The set of sensitive itemsets are sorted based on the support (1). This leads to the selection of a sensitive itemset with smaller support to be selected first which cause minimal changes in the database. For each sensitive itemset the sensitive transactions are extracted and the conflict ratio is computed. Then the transactions are sorted in descending order of conflict ratio (2-4). Based on the disclosure threshold δ, the number of transactions to be modified is calculated (5). The victim item in the transaction is identified as follows: For each item of the sensitive itemset, the conflict ratio of the item is calculated and the item with maximum conflict ratio is selected as victim. If the items have the same item conflict ratio, then the item with minimum support is selected as victim, since it cause minimal changes in the database (6-13). The modified transactions are rewritten to the disk which yields the sanitized database D'(11-15).

The algorithm is memory based, in which the transactions are loaded into the memory and processing is done. As the size of the database grows, it leads to memory bottleneck. To alleviate this problem, it employs a partitioning approach which performs sanitization by loading the partitions [24].

IV. Experimental Analysis

A. Evaluation Metric

In this section, we present the data-sharing measure related to performance of the sanitization approach [20]. This measure quantifies the side-effects regarding legitimate patterns that were missed during sanitization of the original database. The performance measure is specified as follows:

Misses Cost: It denotes the percentage of legitimate patterns that are not discovered from D'

$$MC = \frac{\#\sim Ps(D) - \#\sim Ps(D')}{\#\sim Ps(D)} \qquad (1)$$

where #~Ps(X) denotes the number of non-sensitive patterns in the database X.

B. Results and discussion

To verify the effectiveness of the proposed approach, experiments were conducted on the synthetic and real datasets to compare the misses cost and difference. The results are compared with that of MaxFIA, MinFIA and MICF.All experiments were performed on a PC with Prentium IV processor having 1 GB of main memory.

For the given database and minimum support threshold, the frequent itemsets are generated using the implementation in [8]. A set of sensitive itemsets (non-singleton itemsets) are selected randomly from the frequent itemsets, in which none of the itemsets was a subset of the other. For each sensitive itemset, the algorithms sanitize the transactions based on the disclosure threshold (δ).For our experimental analysis the disclosure threshold is set to zero (δ=0%).

Table 1 lists the summary of the datasets used in the experiment. These datasets are used to measure the effectiveness of the algorithms. IBM synthetic data generator [25] was used to generate the simulated datasets. The parameters are similar to those in [24] to produce transaction-like datasets.

Table 1. Datasets used

Dataset	No.of Transactions	Distinct items	Size
T10I6D100KN500	1,00,000	495	5.2MB
BMS-WebView-1	59,602	497	0.78MB

Fig. 1-4 show the performance of misses cost associated with the algorithms over a range of sensitive itemsets. For varying number of sensitive itemsets, the proposed approach (HCRS) has less misses cost compared with MICF (Fig.1). As noted from fig.2 that the minimum support threshold is varied with a fixed number of sensitive itemsets (200).

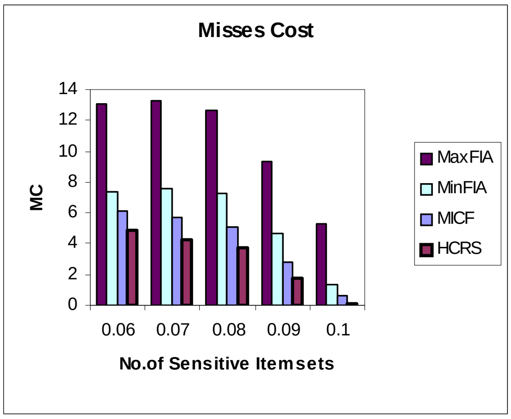

Fig 2. Misses cost for T10I6D100kN500, |Ps|=200 and varying MST

It is evident to observe that from fig. 1-2, HCRS shows better results for the simulated dataset. As the support threshold of the sensitive itemsets becomes closer to the minimum support threshold, there are only minimum numbers of non-sensitive itemsets lost.

The performance of the algorithms are also analysed for the real datasets BMS-WebView1. This dataset contain several months click stream data from e-commerce web site. Fig.3 and 3 shows that for the real dataset BMS-WebView1, the approach (HCRS) has shown better results. The results indicate that it has improved results for the varying sensitive itemsets and support threshold. The proposed approach outperforms MaxFIA, MinFIA and MICF. Likewise, for the varying support threshold HCRS outperforms the other methods.(Fig.2 and 4).

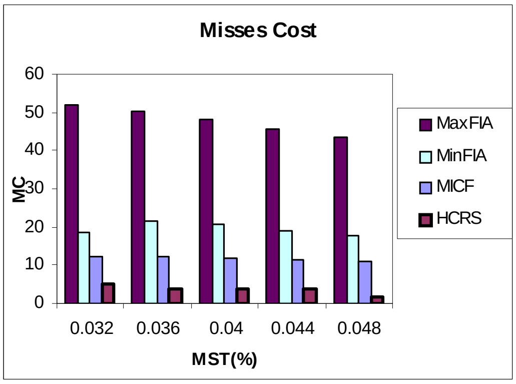

Fig 3. Misses cost for BMS-WebView1, MST=0.064%

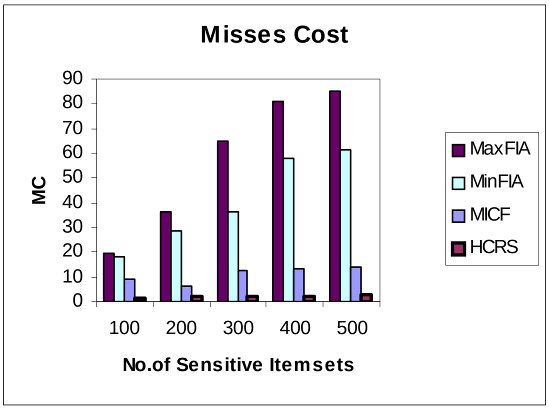

Fig 4. Misses cost for BMS-WebView1, |Ps|=200 and varying MST

V. Conclusion

Privacy becomes a critical issue when the data owner decides to share the database since it reveals sensitive patterns. Data sanitization helps to resolve this problem by making appropriate modifications in the original database. Since data mining helps to explore the hidden knowledge, modified database may not yield accurate mining results. To make a compromise between data sharing, accuracy and privacy, we proposed an approach to minimize the impact on source database in preserving frequent patterns. The proposed approach employs hybrid conflict ratio of the transaction and item to hide the sensitive itemsets such that minimal number of legitimate itemsets is affected. Experimental results clearly indicate that the approach outperforms the earlier algorithms in terms of misses cost. As a future work, we have planned to develop new strategies to minimize the impact on the source database.

Acknowledgment

The authors would like to thank Blue Martini Software, Inc, for kindly providing the BMS datasets.

References

[1] Agrawal D and Aggarwal C.C, "On the Design and Quantification of Privacy Preserving Data Mining Algorithms", In Proc. of ACM SIGMOD/PODS, (2001), pp: 247–255, Santa Barbara, CA.

[2] Agrawal R and Srikant R, "Privacy Preserving Data Mining,", Proc. ACM SIGMOD Conf. Management of Data, (2000), pp: 439-450.

[3] Atallah M., Bertino E., Elmagarmid A., Ibrahim M., and Verykios V.. Disclosure Limitation of Sensitive Rules. Proc. of IEEE Knowledge and Data Engineering Workshop, pages 45–52, Chicago, Illinois, November 1999.

[4] Bhavani Thuraisingham, 2002, Data Mining, National Security, Privacy and Civil Liberties, ACM SIGKDD Explorations Newsletter,. Volume 4, Issue 2 ,pp:1-5

[5] Brankovic L and Estivill-Castro V, "Privacy Issues in Knowledge Discovery and Data Mining", Proc. of Australian Institute of Computer Ethics Conference (AICEC99), Melbourne, Victoria, Australia, July 1999.

[6] Brin S, Motwani R, Ullman J.D, Tsur S, Dynamic itemset counting and implication rules for market basket data,Proceedings of ACM SIGMOD International Conference on Management of Data, Tucson, AZ, 1997, pp. 255–264. 74–85.

[7] Charu C. Aggarwal, Jan Pei, Bo Zhang," On Privacy Preservation against Adversarial Data Mining",Proceedings of Knowledge Discovery in Databases (2006).

[8] Christian Borgelt," Efficient Implementations of Apriori and Eclat",Frequent Itemset Mining Implementations Repository,2004,(http://fimi.cs.helsinki.fi/src/)

[9] Elisa Bertino, Ravi Sandhu,"Database Security—Concepts, Approaches, and Challenges", IEEE transactions on dependable and secure computing, (2005),vol.2,no. 1

[10] Guanling Lee, Chien-Yu Chang and Arbee L.P Chen,"Hiding sensitive patterns in association rule mining" Proceedings of the 28th Annual International Computer Software and Applications Conference (COMPSAC'04),2004

[11] Han J and Kamber M, "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers, San Francisco, CA, 2006.

[12] Han J, Pei J, Yin Y, Mao R, Mining frequent pattern without candidate generation: a frequent pattern tree approach, Data Mining and Knowledge Discovery, Vol :8,No:1 (2004) 53–87.

[13] Herman T. Tavani, 1999, Informational privacy, data mining, and the Internet, Ethics and Information Technology 1: pp:137–145.

[14] Heun, C. 2001. When to share: Wal-Mart and other companies reassess their data-sharing strategies, InformationWeek.com. http://www.informationweek.com/838/data.htm.

[15] Josep Domingo-Ferrer,2005, "Privacy in Data mining", Data mining and Knowledge Discovery,Vol 11, pp: 117–119.

[16] Li Y.C,Chang C.C, A new FP-tree algorithm for mining frequent itemsets,Lecture Notes in Computer Science, vol. 3309, Springer-Verlag, New York, 2004, pp. 266–277.

[17] McDougall, P. 2001. Collaborative business: Companies that dare to share information are cashing in on new opportunities. InformationWeek.com. http://informationweek.com/836/collaborate.htm.

[18] Park J.S,Chen M.S, Yu P.S, An effective hash-based algorithm for mining association rules,Proceedings of ACM-SIGMOD International Conference on Management of Data, San Jose, CA, 1995, pp. 175–186.

[19] Rizvi S. J. and Haritsa J. R., "Maintaining data Privacy in Association Rule Mining", Proc. of the 28th International Conference on Very Large Data Bases, (2002),Hong Kong, China.

[20] Stanley R.M. Oliveira and Osmar R. Zaïane,"Unified framework for sensitive ar hiding", Int. J. Business Intelligence and Data Mining, Vol. 1, No. 3, 2006

[21] Vassilios S. Verykios, Ahmed K. Elmagarmid, Elisa Bertino, Yucel Saygin and Elena Dasseni, "Association rule hiding", IEEE Transactions On Knowledge And Data Engineering, Vol. 16, No. 4, April 2004.

[22] Verykios V.S., Bertino E., Fovino I.N., Provenza L.P., Saygin Y., and Theodoridis Y., "State-of-the-Art in Privacy Preserving Data Mining," ACM SIGMOD Record, (2004),vol. 3, no. 1, pp. 50-57.

[23] Yi-Hung Wu, Chia-Ming Chiang, and Arbee L.P. Chen,"Hiding sensitive association rules with limited side effects", IEEE Transactions On Knowledge And Data Engineering, Vol. 19, No. 1, January 2007.

[24] Yu-Chiang Lia, Jieh-Shan Yeh, Chin-Chen Chang, "MICF: An Effective sanitization algorithm for hiding sensitive patterns on data mining", Advanced Engineering Informatics 21 (2007) 269–280.

[25] IBM Almaden Research Center, Synthetic data generation code for associations and sequential patterns. http://www.almaden.ibm.com/software/quest/Resources/index.shtml, 2003.