IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-2, Issue-9 E-ISSN: 2347-2693

Trade-off between Utility and Security using Group Privacy Threshold Sanitization

Cynthia Selvi P[1], Mohamed Shanavas A.R[2]

[1*] Dept. of Computer Science, KNGA College(W), Thanjavur, Bharathidasan University, Tiruchirapalli, TamilNadu
[2] Dept. of Computer Science, Jamal Mohamed College, Tiruchirapalli , Bharathidasan University, Tiruchirapalli, TamilNadu

www.ijcaonline.org

Received: Aug/25/2014 Revised: Sep/10/2014 Accepted: Sep/22/2014 Published: Sep/30/2014

***Abstract*—** Data mining is a well-known technique for automatically and intelligently extracting useful information or knowledge from a large amount of data, but it can also disclose sensitive information of an individual or a company. This promotes the need for privacy preserving data mining which is becoming an increasingly important field of research and many researchers have proposed techniques for handling this concept. However, most of the privacy preserving data mining approaches concentrate on fixed disclosure threshold strategy for all sensitive information. This article proposes an approach for group-based threshold strategy which may help facilitate to use varying sensitivity level for the information to be hidden.

***Keywords*—** Restricted patterns, Sanitization, Sensitive transactions, Group-based Threshold

I. INTRODUCTION

Data mining technology is emerging as an effective tool for identifying patterns and trends from huge volume of datasets [1]. The growth of data mining applications in both the public and private sectors promotes multifold benefits along with new challenges and more essential issues of which privacy preservation is becoming an increasingly important issue. In a collaborative business environment, multiple organizations may want to reap the extra benefits from their information systems by applying data mining algorithms; but at the same time they may not want to disclose any extra information about their most important sensitive data to other parties for various legal reasons or competition.

Moreover, in recent years with the rapid development in Internet, data storage and data processing technologies, privacy preserving data mining has been drawn increasing attention. The privacy preserving data mining problem was extensively researched on privacy constraints. A number of effective methods for privacy preserving data mining have been proposed [2-8]. But most of these methods might result in information loss and side-effects to some extent like reduced data utility, degraded data mining efficiency. When some sensitive data is completely hidden by some approaches, ultimately this may result in information loss. In addition, sensitivity level may not be common for all sensitive data but may vary for different group of associated items or patterns. Particularly, as the users or experts often have an insight as to which groups are more important than others, it is sometimes more desirable to set up user-specific or group-based privacy thresholds. Hence, preserving sensitive information by introducing a group-based threshold value would establish a balance between the privacy gain and data utility.

This article is an extension work of [9], with the introduction of group-based privacy threshold value. Section-2 briefly states the definitions and proposes the improved algorithm. Section-3 shows the experimental results on measures of effectiveness and efficiency

II. SANITIZATION WITH GROUP-BASED PRIVACY THRESHOLD

A. Definitions

Transactional Database: A transactional database consists of a file where each record represents a transaction that typically includes a unique identity number (*trans_id*) and a list of items that make up the transaction.

Association Rule: It is an expression of the form $X \Rightarrow Y$, where X and Y contain one or more paterns(categorical values) without common elements ($X \cap Y = \phi$).

Frequent Pattern: A pattern(itemset) that forms an association rule is said to be frequent if it satisfies a prespecified minimum support threshold(*min_sup*).

Restrictive Patterns: A set of all patterns rp_i denoted by R_P is said to be *restrictive*, if $R_P \subset P$ and if and only if R_P would derive the set R_H. $\sim R_P$ is the set of *non-restrictive patterns* such that $\sim R_P \cup R_P = P$.

Group Privacy Threshold: A privacy measure(numeric) which determines the sensitivity level of different group of associated items especially in transactional databases.

Sensitive Transactions: A set of transactions are said to be *sensitive*, denoted by S_T, if every $t \in S_T$ contain atleast one restrictive pattern rp_i . ie $S_T = \{ t \in T \mid \exists rp_i \in R_P, rp_i \subseteq t \}$.

Null Transactions: A set of transactions is said to be *null transactions*($\sim S_T$) if they do not contain any of the patterns being examined.

Corresponding Author: *Cynthia Selvi P*

Transaction Size: The number of items which make up a transaction is the size of the transaction.

Transaction Degree: Let D be a source database and S_T be a set of all sensitive transactions in D. The *degree of a sensitive transaction t*, denoted as *deg(t),* such that $t \in S_T$ is defined as the number of restrictive patterns that *t* contains.

Cover: The *Cover*[9] of an item A_k can be defined as, $C_{Ak} = \{ rp_i \mid A_k \in rp_i \subset R_P, 1 \leq i \leq |R_P| \}$
i.e., set of all restrictive patterns which contain A_k.
The item that is included in a maximum number of rp_i's is the one with *maximal cover or maxCover;*
i.e., maxCover = $max(|C_{A1}|, |C_{A2}| , \ldots |C_{An}|)$ such that $A_k \in rp_i \subset R_P$.

B. Sanitization Algorithm with Group Privacy Threshold

Input : (i) D – Source Database

(ii) Γ - Group Privacy Threshold(%)

(iii) n- No. of Groups

(iv) R_P – Set of all Restrictive Patterns

Output : D' – Sanitized Database

***Algorithm*:**

Step 1 : calculate *supCount*(rp_i) $\forall\ rp_i \in R_P$ and sort in decreasing order ;

Step 2 : find *Sensitive Transactions*(S_T) w.r.t. R_P ;

a) calculate *deg(t), size(t)* $\forall\ t \in S_T$;

b) sort $t \in S_T$ in decreasing order of *deg & size ;*
// t- sensitive transaction//

Step 3 : find $\sim S_T \leftarrow D - S_T$;
// $\sim S_T$ - non sensitive transactions //

Step 4 : // Find S_T' //
get n;
do while (n ≥1)
{
get Γ and R_P for every group;
for each $rp_i \in R_P$ do
{
extract S_{Trpi} ; //initially all *t* are *nonvictim* //
find nTs =min[($|S_{Trpi}| \times (1- \Gamma)$,(|*nonVictimTransactions*|)]
// *nTs- no.of transactionToSanitize*//
repeat
for each $t \in nTs$
{
find *cover* for every item A_k such that $A_k \in rp_i \subset t$;
delete A_k with *maxCover*
(*round robin* in case of tie); // A_k – *victimItem* //
decrease *supCount* of all rp_i's which contain *victimItem;* // $A_k \in rp_i \subset t$
mark *t* as *victimTransactions* w.r.t each rp_i ;
}
until (*supCount* = 0) ;
}
Step 5 : D' ← $\sim S_T \cup S_T$'

Before initiating the sanitization process, a disclosure threshold value is set by the owner or the user of the database. This disclosure threshold enables the restrictive patterns not completely to be hidden but to a certain percentage value which in turn reduces the rate of accidental hiding of some legitimate patterns during sanitization process. In other words, this threshold represents a tradeoff between privacy and utility. The choice of setting suitable privacy threshold value is left to database owner or user based on the sensitivity level of each and every group of associated sensitive data.

The sensitive patterns (that are to be hidden) are identified and grouped according to their privacy threshold. For every pattern in a group find the non-victim transactions which avoid redundant visit and decrease of the support count for the transactions which are already considered. This look-ahead procedure speed up the process and eliminate redundancy. Sometimes the number of non-victim transactions would be less than the actual number of sensitive transactions, as many of the transactions would be visited in the previous iterations; hence the minimum of these two is considered for sanitization which would definitely reduce the sanitization rate.

III. EXPERIMENTAL RESULTS

The algorithm was tested for real dataset T10I4D100K[10] by considering the number of transactions ranging between 1000 to 10000 and the details of Restricted Patterns used are given in the table-I. The test run was made on Intel core i5 processor with 2.3 GHz speed and 4GB RAM operating on 32 bit OS; The implementation of the proposed algorithm was done with windows 7-Netbeans 6.9.1-SQL 2005. The frequent patterns were obtained using Matrix Apriori[11] approach.

Table-I. Sensitive Patterns with Group Threshold

	Threshold	Patterns
Group-1	20%	354,58
		438,75
		217,346
Group-2	30%	354,752
		217,283,515

A. Effectiveness Measures:

Privacy Loss : It is measured as the ratio of the total support count of the restrictive patterns in sanitized dataset(D') to source dataset(D).

$$PL = \frac{\sum_{i=1}^{k} supCount(rpi) in\ D'}{\sum_{i=1}^{k} supCount(rpi)\ in\ D}$$

Privacy Gain : It is measured as the ratio of the difference in the total support count of the restrictive patterns in sanitized dataset(D') and source dataset(D) to the total support count of the restrictive patterns in source dataset(D).

$$PG = \frac{\sum_{i=1}^{k} supCount(rpi)\ in\ D - \sum_{i=1}^{k} supCount(rpi) in\ D'}{\sum_{i=1}^{k} supCount(rpi) in\ D}$$

Information Loss : It is measured as the ratio of the difference in the total number of non-restrictive patterns in the sanitized dataset(D') and source dataset(D) to the total number of non-restrictive patterns in the source dataset(D).

$$IL : \frac{|\sim RP(D)| - |\sim RP(D')|}{|\sim RP(D)|}$$

As there are functional dependencies between restricted and non-restricted patterns, some rules would accidentally be removed, which may happen when some of the non-restrictive patterns lose support in the dataset during sanitization process.

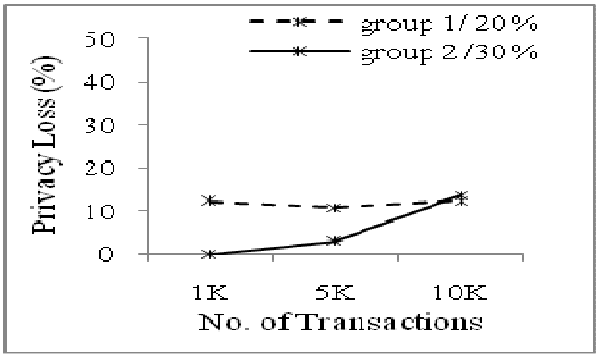

Fig.1. Privacy Loss

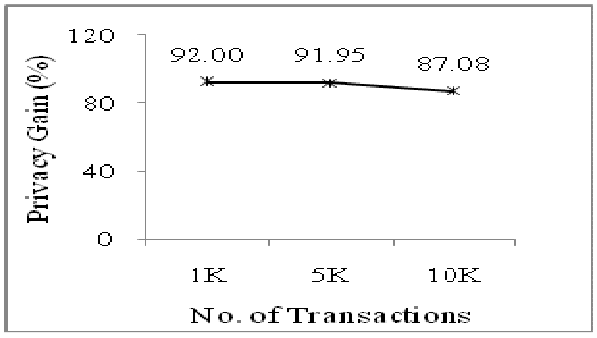

Fig.2. Privacy Gain

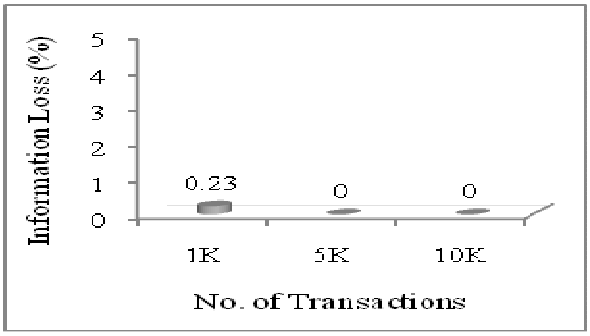

Fig.3. Information Loss

B. Efficiency Measures:

Dissimilarity(dif): The dissimilarity between the original(D) and sanitized(D') datasets is measured by comparing their contents instead of their sizes and it is calculated by,

$$dif(D, D') = \frac{1}{\sum_{i=1}^{n} fD(i)} x \sum_{i=1}^{n} [fD(i) - fD'(i)]$$

where *fx(i)* represents the i^{th} item in the dataset *X*.

Execution Time: The execution time and the scalability of the proposed algorithm is obtained by varying the size of the dataset. It is observed that the execution time is linear.

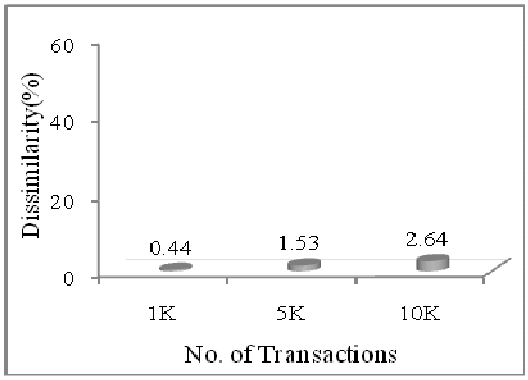

Fig.4. Dissimilarity

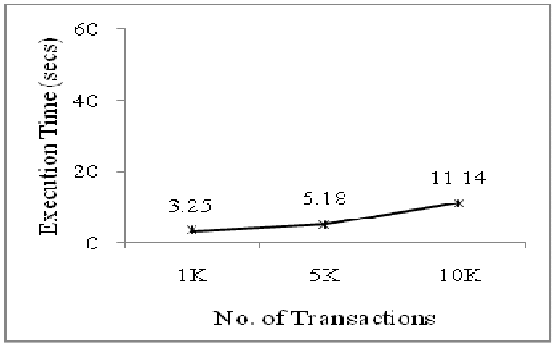

Fig.5. Execution Time

VIII. CONCLUSION

This article proposes a strategy to introduce a variegated sensitivity level for the sensitive patterns to be protected against disclosure that facilitate an effective improvement in maintaining privacy and utility with reduced information loss and it is also proved by the experimental results.

REFERENCES

[1] Richard Yevich, "Data Mining," In: Joyce Bischoff, Ted Alexander and Sid Adelman (editors). *DataWarehouse: Practical Advice from the Experts.* Upper Saddle River, N.J.: Prentice Hall, (1997).

[2] A. Evfimievski, R. Srikant, R. Agrawal, J. Gehrk, "Privacy Preserving Mining of Association Rules", In Proceedings the 8th ACM SIGKDD International Conference on Knowledge Discovery in Databases and Data Mining, pp.217-228, 2002.

[3] S. Rizvi, J. Haritsa, "Maintaining Data Privacy in Association Rule Mining", In Proceedings the 28th International Conference on Very Large Data Bases, pp.682-693, 2002.

[4] L. Sweeney, "k-Anonymity: A Model for Protecting Privacy", International Journal of Uncertainty, Fuzziness and Knowledge-based Systems, vol.10, no.5, pp.557-570,2002.

[5] X.K. Xiao, Y.F. Tao, "Personalized Privacy Preservation", In Proceedings of the ACM Conference on Management of Data (SIGMOD), pp.229-240, 2006.

[6] M. Kantarcioglu, 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.

[7] Lindell, Yehuda, Pinkas, "Privacy preserving data mining", In Proceedings of the Advances in Cryptology–CRYPTO, pp.36–54,2000.

[8] J. Vaidya, C. Clifton, "Privacy Preserving Association Rule Mining in Vertically Partitioned Data", In Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.639-644, 2002.

[9] Cynthia Selvi P., Mohamed Shanavas A.R., "Towards Information Privacy Using Transaction-Based Maxcover Algorithm", World Applied Sciences Journal 29 (Data Mining and Soft Computing Techniques): 06-11, 2014, ISSN 1818-4952, © IDOSI Publications, 2014, DOI:10.5829/ idosi.wasj.2014.29. dmsct.2, Pages. 06-11

[10] The Dataset used in this work for experimental analysis was generated using the generator from IBM Almaden Quest research group and is publicly available from http://fimi.ua.ac.be/data/.

[11] Pavon J, Viana S, Gomez S, "Matrix Apriori: speeding up the search for frequent patterns," Proc. 24th IASTED International Conference on Databases and Applications, 2006, pp. 75-82.