Privacy-Preserving Data Sharing Using Data Reconstruction Based Approach

Kshitij Pathak
MIT, Ujjain

Narendra S. Chaudhari
IIT, Indore

Aruna Tiwari
SGSITS, Indore

ABSTRACT
Data mining services require accurate input data for their results to be meaningful, but privacy concerns may influence users to provide spurious information. To preserve client privacy in the data mining process, a variety of techniques based on random perturbation of data records have been proposed recently. One known fact which is very important in data mining is discovering the association rules from database of transactions where each transaction consists of set of items. There are many approaches to hide certain association rules which take the support and confidence as a base for algorithms ([1, 2, 6] and many more). This research work discusses privacy and security issues that are likely to affect data mining projects. This research work focuses on further investigating reconstruction-based techniques for association rule hiding, the problem of sharing sensitive knowledge by sanitization and hope that proposed solution will fetch up the new reconstruction-based research track and work well according to the evaluation metrics including hiding effects, data utility, and time performance.

Keywords
Frequent Item sets, Data Mining, Cursors, Association Rules.

1. INTRODUCTION
Data mining is the process of extracting hidden patterns from data. As more data is gathered, with the amount of data doubling every three years, data mining is becoming an increasingly important tool to transform this data into knowledge. It is commonly used in a wide range of applications, such as marketing, fraud detection and scientific discovery. Data mining can be applied to data sets of any size, and while it can be used to uncover hidden patterns, it cannot uncover patterns which are not already present in the data set. Data mining extracts novel and useful knowledge from data and has become an effective analysis and decision means in corporation. Data sharing can bring a lot of advantages for research and business collaboration. However, large repositories of data contain private data and sensitive rules that must be preserved before published. Motivated by the multiple conflicting requirements of data sharing, privacy preserving and knowledge discovery, privacy preserving data mining (PPDM) has become a research hotspot in data mining and database security fields.

Two problems are addressed in PPDM: one is the protection of private data; another is the protection of sensitive rules (knowledge) contained in the data. The former settles how to get normal mining results when private data cannot be accessed accurately; the latter settles how to protect sensitive rules contained in the data from being discovered, while non-sensitive rules can still be mined normally. The latter problem is called knowledge hiding in database in (KHD) which is opposite to knowledge discovery in database (KDD). And association rule hiding problem we focus is one of problems in KHD. The basic idea of data reconstruction is to perform knowledge sanitization rather than data sanitization.

2. ASSOCIATION RULES
In data mining, association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases. Authors describe analyzing and presenting strong rules discovered in databases using different measures of interestingness. Based on the concept of strong rules, Agrawal et al introduced association rules for discovering regularities between products in large scale transaction data recorded by point-of-sale (POS) systems in supermarkets For example, the rule

$$\{onions, potatoes\} \Rightarrow \{beef\}$$

Found in the sales data of a supermarket would indicate that if a customer buys onions and potatoes together, he or she is likely to also buy beef. Such information can be used as the basis for decisions about marketing activities such as, e.g., promotional pricing or product placements. In addition to the above example from market basket analysis association rules are employed today in many application areas including Web usage mining, intrusion detection and bioinformatics.

3. BACKGROUND AND RELATED WORK
The effect of security impact of Data Mining is analyzed in [12] and some possible approaches to the problem of inference and discovering sensitive association rule in a DM context are investigated. The proposed strategies include fuzzyfying and augmenting the source database and also limiting the access to the source database by releasing only samples of the original data. Clifton [13] adopts the last approach as author studies the correlation between the amount of released data and the significance of the sensitive rules which are discovered. He shows how to determine the sample size in a way that data mining tools cannot obtain sensitive results.

Clifton and Marks in [12] also recognize the necessity of analyzing the various data mining algorithms in order to increase the efficiency of any adopted strategy that deals with disclosure limitation of sensitive data and knowledge. The solution proposed by Clifton in [13] is independent from any specific data mining technique; other researchers [14], [15] propose solutions that prevent disclosure of confidential information for specific data mining algorithms such as association rule mining and classification rule mining.

Classification mining algorithms may use sensitive data to rank objects; each group of objects has a description given by a combination of non sensitive attributes. The sets of descriptions, obtained for a certain value of the sensitive attribute, are referred to as description space. For Decision-Region-based algorithms, the description space generated by each value of the sensitive attribute can be determined a priori. The authors in [8] first identify two major criteria which can be used to assess the output of a classification inference system and then they use these criteria, in the context of Decision-Region based algorithms, to inspect and

to modify, if necessary, the description of a sensitive object so that they can be sure that it is not sensitive.

There is a large amount of work related to association rule hiding. Maximum researchers have worked on the basis of reducing the support and confidence of sensitive association rules ([1-4,6,7,9-11]). ISL and DSR are the common approaches used to hide the sensitive rules. Actually any given specific rules to be hidden, many approaches for hiding association, classification and clustering rules have been proposed. Some of the researchers have used data perturbation techniques ([5]) to modify the confidential data values in such a way that the approximate data mining results could be obtained from the modified version of the database. Some researchers also recognize the necessity of analyzing the various data mining algorithms in order to increase the efficiency of any adopted strategy that deals with disclosure limitation of sensitive data and knowledge. Also disclosure limitation of sensitive knowledge by data mining algorithms, based on the retrieval of association rules, has been recently investigated.

4. PROBLEM STATEMENT

Let $I = \{i_1, i_2, \ldots, i_n\}$ be a set of *n* binary attributes called *items*.

Let $D = \{t_1, t_2, \ldots, t_m\}$ be a set of transactions called the *database*.

Each transaction in *D* has a unique transaction ID and contains a subset of the items in *I*. A *rule* is defined as an implication of the form

$$X \Rightarrow Y$$

where

$$X, Y \subseteq I \text{ and } X \cap Y = \emptyset$$

and .

The sets of items (for short *item sets*) *X* and *Y* are called *antecedent* (left-hand-side or LHS) and *consequent* (right-hand-side or RHS) of the rule.The *support* supp(*X*) of an item set *X* is defined as the proportion of transactions in the data set which contain the item set.Confidence can be interpreted as an estimate of the probability $P(Y \mid X)$, the probability of finding the RHS of the rule in transactions under the condition that these transactions also contain the LHS

An association rule is an implication of the form $X \Rightarrow Y$, where X ,Y $\subset$ Itemsets, and X intersect Y= Φ . We say the rule X $\rightarrow$ Y holds in the database D with confidence c if $|X \cup Y| / |X| \geq c$. It can also be said that the rule X $\rightarrow$ Y has support s if $|X \cup Y| / |D| \geq s$.. Note while the support is a measure of the frequency of a rule, the confidence is a measure of the strength of the relation between sets of items. The well-known association rule mining problem aims to find all significant association rules. A rule is significant if its support and confidence is no less than the user specified minimum support threshold (*MST*) and minimum confidence threshold (*MCT*). To find the significant rules, an association rule mining algorithm first finds all the frequent itemsets and then derives the association rules from them. On the contrary, the association rule hiding problem aims to prevent some of these rules, which we refer to as "sensitive rules", from being mined.

5. PROPOSED APPROACH

Three phases of the framework are Generate Itemset Lattice, Knowledge Sanitization and Data Reconstruction. Work also proposes a new sanitization algorithm. The approach presented in this research work has been applied to many sample databases for testing its accuracy and it generates satisfactory Results.

5.1 Generate Itemset Lattice

Let us reiterate a well established fact by Chen [16]: the power set of the items in a database together with the subset relation form a lattice of the item space P(Ĩ) that contains all possible subsets of items.

The supports of all subsets of items according to data-base D form a frequency set S(P(Ĩ), D) associated with the itemsets lattice and it can be obtained by applying an Apriori-like algorithm[16]. Mining frequent itemsets is the primary task for generating association rules. Apriori-like algorithms employ a bottom-up, breadth-first search in the itemsets lattice space level by level. At the kth level, the supports of itemsets of cardinality k are counted. In this process, normally there are two ways to scan the dataset. If the number of itemsets in the lattice is much bigger than the number of transactions, a more efficient approach is to scan the dataset once, looking at one transaction at a time and finding all itemsets that occur in the lattice, incre-menting the count by 1. Otherwise, the dataset can be scanned once for each itemset in the lattice. Frequent Itemset Mining is to find all itemsets whose support is higher than the pre-defined threshold from itemsets lattice space. Once all frequenct itemsets are found, association rules can be easily derived.

There are two important principles in frequent itemset mining:

Monotonicity Principle: Let J (subset of I) & I be two itemsets, the support of I will be at most as high as the support of J.

5.1.1 Procedure for generating 1-itemset in PL/SQL

// items is a table having

collection of all transactions

//item_lattice is a table storing frequent itemsets with their support counts

Begin

copy all items from items to item_lattice table with their support counts.

End;

5.1.2 Procedure for generating n-itemset in PL/SQL

PROCEDURE n-itemset IS
Buffer items from item table in buffer 1 ;
// Cursors in PL/SQL are used as buffers
// items is a table having collection of all transactions
//table market contains basket data

Buffer items from item_lattice table in buffer2 ;
fetch from buffer1 into x until no items found ;
{
fetch from buffer2 into k until no items found ;
if (x not like '%' || k || '%' and k > x and length(x) = n- 1) then
// if items from buffer1 and buffer2 are not same
//then it's concatenation is a candidate

```
//for next frequent itemset
count record into Q from market where
items like '%' || x || '%' and
items like '%' || k || '%' ;
// check concatenation of x and k are present in how many transaction out of total number of transactions
insert into item_lattice values( (x || k) , Q ) ;
}
commit ;
END;
```

Considering an example for explaining the approach and checking its validity. Tables 1 show a sample transaction datasets on the some set of items I={W, X, Y, Z} and D1={T1,…,T15}.
In the above example, I represents the set of items and W,X,Y,Z are the various items. D1 consists of 15 transactions. The details of various transactions in D1 are as follows:-

Table 1. Sample Database D1

Transactions	Items
T1	WXZ
T2	XZ
T3	XY
T4	WXZ
T5	WY
T6	XY
T7	WXYZ
T8	WXY
T9	WX
T10	WZ
T11	XYZ
T12	WXYZ
T13	WXY
T14	WYZ
T15	Y

In the above example, Total number of items is 4 so, number of lattice itemset will be 16. Since in lattice total number of itemset is power set of number of items. 1-itemsets will be generated first. After generating 1-itemset, procedure runs for 2-itemset, 3-itemset and 4-itemset recursively.

After applying the algorithm, 1-itemset will be generated as shown in Table 2 and itemset lattice is shown in Figure 1.

Table 2. Items with their frequency count for database D1

Items	Frequency Count
W	10
X	11
Y	10
Z	8

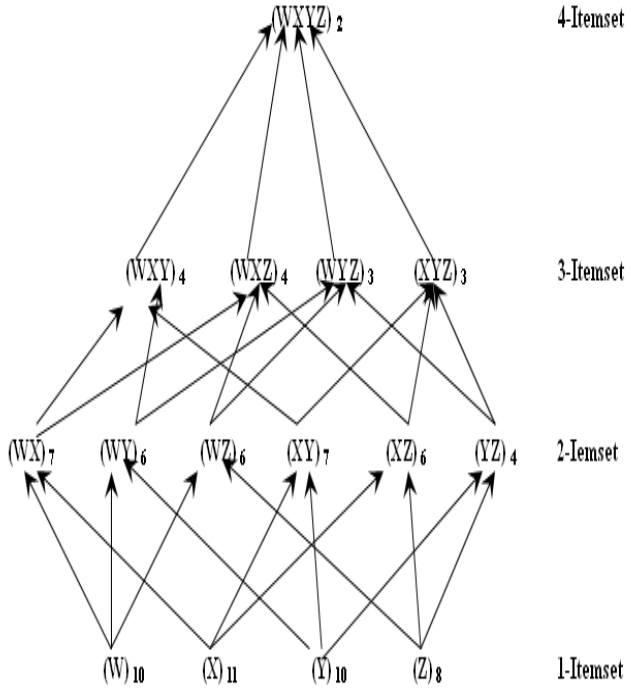

Fig 1: Itemset Lattice

In the lattice shown in figure 1, the number which is written in the subscript with the lattice items represents the number of transactions in which that itemset is present.

For example itemset "WXY" is present in 4 transactions i.e. transaction number T7, T8, T12, T13. If minimum support = 4 for the above database then { W , X , Y , Z , WX , WY, WZ , XY, XZ , YZ , WXY , WXZ } can be marked as "frequent itemsets".

5.2 Knowledge Sanitization

This phase takes the input Itemset Lattice and association rules to be hided. In this phase itemset lattice is modified to hide the sensitive association rules.

5.2.1 Algorithm K-Sanitization

Input: - Association rules and Itemset Lattice
Output: - Modified Itemset Lattice

1. Start
2. Generate all association Rules from frequent itemsets
3. for every sensitive rule B à A do

a) Make the support of lattice itemset BA to a value lower than MST and MCT, i.e. if the support of BA of is k, greater then MST and MCT, decrease r from its support to make it below MST and MCT.
b) Find all the subsets of BA, and decrease their support value by r.
c) Find all Supersets of BA and if any superset have support more than modified support of BA, then it violates the consistent relationship, so modify its support to satisfy the consistent relationship.
4. Exit

5.2.2 Data Reconstruction

In this phase Database is generate using inverse frequent set mining on modified itemset lattice.

6. EXPERIMENTAL RESULTS

Approach has been applied to many sample databases and also check with two popular association rule hiding algorithms ISL and DSR. Work applied on sample database 2 and sample database 3 shown in table 3 and table 5 respectively and found satisfactorily results as shown in table 4 and table 6..

Table 3. Sample Database D2

T1	WXZ	1101
T2	X	0100
T3	WYZ	1011
T4	WX	1100
T5	WXZ	1101

Table 4. Results on sample database D2

Algorithms	No. of Sensitive Rules to be Hided	No. of Sensitive Rules Hidden	Ghost Rules Generated	No of Lost Rules
ISL	1	0	0	0
DSR	1	1	0	1
Proposed Framework	1	1	0	0

Table 5. Sample Database D3

Transaction	Items
T1	**WYZ**
T2	**Y**
T3	**WYZ**
T4	**WX**
T5	**YZ**
T6	**YZ**
T7	**XY**
T8	**WZ**
T9	**XYZ**
T10	**W**

Table 6. Results on sample database D3

Algorithms	No. of Sensitive Rules to be Hided	No. of Sensitive Rules Hidden	Ghost Rules Generated	No of Lost Rules
ISL	**2**	**1**	**1**	**2**
DSR	**2**	**1**	**0**	**4**
Proposed Framework	**2**	**2**	**0**	**0**

7. ANALYSES AND CONCLUSION

New Framework based on knowledge sanitization rather than transaction modification. Framework gives more satisfactory results than popular data sanitization algorithms like ISL and DSR. Approach applied to many datasets for testing its accuracy and it generates satisfactorily Results

8. REFERENCES

[1] 1. Shyue-Liang Wang, Yu-Huei Lee, Steven Billis, Ayat Jafari "Hiding Sensitive Items in Privacy Preserving Association Rule Mining" 2004 IEEE International Conference on Systems, Man and Cybernetics

[2] Vassilios S. Verykios, Ahmed K. Elmagarmid, Elisa Bertino, Yucel Saygin and Elena Dasseni"Association Rule Hiding", IEEE Transactions on Knowledge and Data Engineering, Vol. 16No. 4, April 2004.

[3] Yucel Saygin, Vassilios S. Verykios, Chris Clifton "Using Unknowns to Prevent Discovery of Association Rule" SIGMOD Record, Vol. 30, No.4, December 2001.

[4] Chris Clifton, Don Marks "Security and Privacy Implications of Data Mining", In Proceedings of the 1996 ACM SIGMOD Workshop on Data Mining and Knowledge Discovery.

[5] R. Agrawal and R. Srikant, "Privacy preserving data mining", In ACM SIGMOD Conference on Management of Data, pages 439450, Dallas, Texas, May 2000.

[6] Vi-Hung Wu, Chia-Ming Chiang, and Arbee L.P. Chen, Senior Member, IEEE Computer Society Hiding Sensitive Association Rules with Limited Side Effects IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 19, NO.1, JANUARY 2007

[7] R. Agrawal, T. Imielinski, and A. Swami, "Mining Association Rules between Sets of Items in Large Databases", In Proceedings of ACM SIGMOD International Conference on Management of Data Washington DC, May 1993.

[8] S. Oliveira, o. Zaiane, "Algorithms for Balancing Privacy and Knowledge Discovery in Association Rule Mining", Proceedings of 71 th International Database Engineering and Applications SYmposium (IDEAS03), Hong Kong, July 2003.

[9] Wu, Y.H., Chiang, C.M., and Chen, A.L.P. Hiding sensitive association rules with limited side effects. IEEE Transactions on Knowledge and Data Engineering, 2007,19(1):29-42.

[10] Fienberg, S. and Slavkovic, A. Preserving the confidentiality of categorical statistical data bases when releasing information for association rules. Data Mining and Knowledge Discovery, 11(2):155-180,2005.

[11] State-of-the-art in Privacy Preserving Data Mining Vassilios S. Verykios, Elisa Bertino, Igor Nai Fovino Loredana Parasiliti Provenza, Yucel Saygin, Yannis Theodoridisl SIGMOD Record, Vol. 33, No.1, March 2004, Pages: 50 - 57.

[12] C. Clifton and D. Marks, "Security and Privacy Implications of Data Mining," Proc. 1996 ACM Workshop Data Mining and Knowledge Discovery, 1996.

[13] C. Clifton, "Protecting against Data Mining through Samples," Proc. 13th IFIP WG11.3 Conf. Database Security, 1999.

[14] T. Johnsten and V.V. Raghavan, "Impact of Decision-Region Based Classification Mining Algorithms on Database Security," Proc. 13th IFIP WG11.3 Conf. Database Security, 1999.

[15] M. Atallah, E. Bertino, A. Elmagarmid, M. Ibrahim, and V. Verykios, "Disclosure Limitation Of Sensitive Rules," Proc. Knowledge and Data Exchange Workshop, 1999.

[16] Chen, X., Orlowska, M., and Li, X. A new framework for privacy preserving data sharing. In: Proc. of the 4th IEEE ICDM Workshop: Privacy and Security Aspects of Data Mining. IEEE Computer Society, 2004. 47-56.