Privacy-Preserving Statistical Quantitative Rules Mining *

(Work-in-Progress)

Weiwei Jing
Depart.of Comp. Sci. & Tech., USTC
NHPCC 416, East Campus
USTC, Hefei, 230026, PRC
Tel: 86-551-3602445
wwjing@mail.ustc.edu.cn

Liusheng Huang
Depart.of Comp. Sci. & Tech., USTC
NHPCC 416, East Campus
USTC, Hefei, 230026, PRC
Tel: 86-551-3602445
lshuang@ustc.edu.cn

Yifei Yao
Depart.of Comp. Sci. & Tech., USTC
NHPCC 416, East Campus
USTC, Hefei, 230026, PRC
Tel: 86-551-3602445
yaoyifei@mail.ustc.edu.cn

Weijiang Xu
Depart.of Comp. Sci. & Tech., USTC
NHPCC 416, East Campus
USTC, Hefei, PRC
Tel: 86-551-3602445
wjxu@mail.ustc.edu.cn

ABSTRACT

This paper considers the problem of mining Statistical Quantitative rules (SQ rules) without revealing the private information of parties who compute jointly and share distributed data. Based on several basic tools for Privacy-Preserving Data Mining (PPDM), including secure sum, secure mean and secure frequent itemsets, this paper presents two algorithms to accomplish privacy-preserving SQ rules mining over horizontally partitioned data. One is to securely compute confidence intervals for testing the significance of rules; the other is to securely discover SQ rules.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data mining; F.m [**Theory of Computation**]: Miscellaneous; E.3 [**Data**]: Data Encryption.

General Terms

Algorithms, Security, Theory

Keywords

Privacy-Preserving Data Mining, Statistical Quantitative rules

* Supported by the National Natural Science Foundation (No.60573171), the Science Foundation of Anhui Province (No.070412043), the Ph.D. Program Foundation of Ministry of Education of China (No.20060358014), and the China Postdoctoral Science Foundation (No.20060390700).

1. INTRODUCTION

Statistical Quantitative (SQ) rule plays an important and useful role in data mining. Centralized algorithms [1] have been presented for SQ rules mining. However, the algorithms cannot be easily adapted to mining SQ rules on distributed data, where privacy of parties becomes great concerns. Furthermore, privacy-preserving algorithms for mining other types of association rules on horizontally partitioned data (e.g., [2], [3]) are not applied to SQ rule. Thus, specific algorithms for privacy-preserving SQ rules mining need to be researched. SQ rules are useful in many domains. An example of SQ rule for mobile net companies would be sex = male $\vee$ profession = teacher $\Rightarrow$ mean of communication time per month with mobiles = x hours.

2. DEFINITIONS

We employ the form of rules by modifying the definition in [1]: the LHS(left-hand side) of an SQ rule is an itemset involving only categorical attributes and the RHS(right-hand side) is mean of values of the subset satisfying the LHS and we test significance of a rule based on confidence intervals.

When n parties, each of them has private data set but has the same attribute set, jointly mine for SQ rules, two privacy constraints need to be considered: 1) No party can learn any information of single record from other parties' data set; 2) No party can learn any information of mean of values from other parties' data subset.

The aim of our work is to present algorithms for mining SQ rules on horizontally partitioned data under these privacy constraints in the semi-honest model.

3. BUILDING BLOCKS

We employ three basic algorithms in our algorithm for privacy-preserving SQ rules mining. The three are: 1) *Secure_Sum* [2]; 2) *Secure_Mean: all* parties compute $Secure\_Sum(X_i)/N$; 3) *Secure_ Frequent_Itemsets* ([2] and [3]);

4. SECURE COMPUTING CONFIDENCE INTERVALS

The centralized algorithm in [1] is modified to satisfy both the security requirement and the multi-party computation setting. The new algorithm is called algorithm 1 and is outlined in *Table 1*.

Theorem 1(Security)**:** *Algorithm 1 securely computes the confidence intervals.*

Proof sketch: We construct a simulator S to simulate the view of all parties on the known input and output. Note that the distribution of random variable R_{id} that records the number of *ids* received by party *i* in Step 5 is a binomial distribution. S invokes the S_{mean} for views of *Secure_Mean* and outputs all views. □

Table 1: Algorithm 1(secure computing confidence intervals)

Input: Database D_i , $1 \le i \le n$, where *n* parties have their own database D_i respectively, N_{dist} , N_{perm} , α , N where N is the total number of records.
Output: N_{dist} distributions and significance thresholds.
1 For each iteration *dist* of the loop for computing N_{dist} distributions
2 { Part 1 do: $N_{sample} = dist / N_{dist} \times N$;
3 For each sampling of N_{perm} sampling processes
4 { Part 1 generates N_{sample} id numbers at random and sends them to the corresponding party, respectively;
5 Party *i* samples randomly N_i records from D_i where N_i is the number of ids received from party 1;
6 All parties compute *Secure_Mean*(S_i , N_{sample});}
7 Party 1 sorts N_{perm} means and computes the confidence interval based on significance level α : *lowerCI*[*dist*], *upperCI*[*dist*];}
8 Party 1 outputs N_{dist} distributions and significance thresholds.

Complexity: the total number of rounds in algorithm 1 is $O(N \times N_{perm} \times N_{dist})$, whereas the number of bits exchanged in each round is not too large. For every party, the computation complexity of algorithm 1 is $O(N \times N_{perm} \times N_{dist})$ that is the same as the centralized algorithm in [1], i.e., the secure algorithm for confidence intervals does not bring additive time cost.

5. PRIVACY-PRESERVING SQ RULES MINING

The algorithm for privacy-preserving SQ rules mining is outlined in *Table 2* and is called algorithm 2.

Theorem 2(Security)**:** *Algorithm 2 securely computes SQ rules.*

Proof sketch: We consider the view of party *i* step by step. We can construct a simulator S' which can simulate the view of party *i* in algorithm 2 by invoking S_{SFI} and S_{mean} , which are the simulator for *Secure_ Frequent_Itemsets* and the simulator for *Secure_Mean,* respectively. □

Complexity: Let the round complexity of secure frequent itemsets be $O(N_{SFT})$, and then we have the round complexity $O(N_{SFT} + |L| \cdot n)$ of algorithm 2. Likewise, the time complexity of algorithm 2 is $O(t_{SFT} + |L| \cdot N)$, where t_{SFT} denotes the computation cost of the algorithm for secure frequent itemsets.

Table 2: Algorithm 2(privacy-preserving SQ rules mining)

Input: Database D_i , $1 \le i \le n$, N_{dist} , N , the threshold for frequent itemsets *minsup*, and significance thresholds.
Output: A set of SQ rules.
1 Parties do: (L, *support* []) = *Secure_Frequent_Itemsets* (D_i , *minsup*);
2 For every frequent itemset $x \in L$
3 { Each party compute the sum S_i' of values of the quantitative attribute of records satisfying the itemset;
4 Parties do: $mean_x$ = *Secure_Mean* (S_i' , *support*[*x*]);
5 Based on *support*[*x*], party 1 selects confidence interval and tests the significance;
6 Party 1 holds the rule if its RHS is outside the interval;}
7 Party 1 outputs all rules.

6. CONCLUSIONS

This paper concerns the security of distributed statistical quantitative rules mining. We have presented two algorithms for mining SQ rules without revealing private information of parties on horizontally partitioned data in the semi-honest model. Based on several basic tools of secure multi-party computation, algorithm 1 is proposed for securely computing confidence intervals which are used to test the significance of SQ rules. Algorithm 2 is presented to mine for SQ rules without disclosure of private information of parties. In both algorithm 1 and algorithm 2, any information of single record from local dataset of parties and any information of mean of values from local data subsets of parties are not revealed. In addition, the security and the complexity of the proposed algorithms are analyzed. More detailed discussion, including more detailed description of algorithms, more detailed analysis and experiments, will be our future work.

7. REFERENCES

[1] Hong Zhang, Balaji Padmanabhan, and Alexander Tuzhilin. On the discovery of significant statistical quantitative rules, In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining KDD '04*. ACM Press, New York, NY, USA, 2004, 374 - 383.

[2] M. Kantarcioglu, and C. Clifton. Privacy-preserving Distributed Mining of Association Rules on Horizontally Partitioned Data, *Transactions on Knowledge and Data Engineering. Volume 16, Issue 9*, IEEE Computer Society Press, Los Alamitos, Sept 2004, 1026 - 1037.

[3] Weiwei Jing, Liusheng Huang, Yonglong Luo, Weijiang Xu, and Yifei Yao. An Algorithm for Privacy-Preserving Quantitative Association Rules Mining, In *Proceedings of the 2nd IEEE International Symposium on Dependable, Autonomic and Secure Computing (DASC'06)* (Indianapolis, USA, September 29-October 1, 2006). IEEE Press, 2006, 315 - 324.