Journal of Computer Science 8 (6): 872-878, 2012
ISSN 1549-3636

Privacy Preserved Collaborative Secure Multiparty Data Mining

[1]Balamurugan, M., [2]J. Bhuvana and [3]S. Chenthur Pandian
[1]Department of IT, Selvam College of Technology, Namakkal, India
[2]Department of MCA, Selvam College of Technology, Namakkal, India
[3]Mahalingam College of Engineering and Technology, Pollachi, India

Abstract: Problem statement: In the current modern business environment, its success is defined by collaboration, team efforts and partnership, rather than lonely spectacular individual efforts in isolation. So the collaboration becomes especially important because of the mutual benefit it brings. Sometimes, such collaboration even occurs among competitors, or among companies that have conflict of interests, but the collaborators are aware that the benefit brought by such collaboration will give them an advantage over other competitors. **Approach:** For this kind of collaboration, data's privacy becomes extremely important: all the parties of the collaboration promise to provide their private data to the collaboration, but neither of them wants each other or any third party to learn much about their private data. One of the major problems that accompany with the huge collection or repository of data is confidentiality. The need for privacy is sometimes due to law or can be motivated by business interests. **Results:** Performance of privacy preserving collaborative data using secure multiparty computation is evaluated with attack resistance rate measured in terms of time, number of session and participants and memory for privacy preservation. **Conclusion:** Privacy-preserving data mining considers the problem of running data mining algorithms on confidential data that is not supposed to be revealed, even to the party running the algorithm.

Key words: Privacy preserving data mining, privacy preserving collaborative data mining, secure mutltiparty computation

INTRODUCTION

This research study is concerned with the study and analysis of preserving privacy in collaborative data mining in order to improve the efficiency and effectiveness of privacy preservation among the collaborative parties. Data mining is a process which uses different data analysis tools that discover patterns and relationships in data that can be used to make predictions (Sybil and Liebeman, 2001). Most existing data mining algorithms are carried out under the assumption that all the data could be available at single central site. While two or more parties, who don't have enough confidence in each other or even adversary individuals or organizations, have a common desire to extract knowledge from all of their private data, the privacy problems come up. As the data mining in the public and private sectors is increasingly used, privacy is becoming an important issue.

When common users are involved in data mining, all users need to send their data to trusted common centre to conduct the mining; however, in situations with privacy concerns, it is very difficult for a user to trust the other users and in such a situation, the process is called Privacy Preserving Collaborative Data Mining (PPDM) (Yasien, 2007) and the gap between the data mining and data confidentiality can be filled by the privacy preserving data mining. The trusted partners share information between themselves, by maintaining their privacy, with the secure cooperative computations. The shared information sent by the participant to a remote database also contains privacy data inferences which should not be disclosed to the receiving participant. In order to maintain privacy of individual's, when sharing data in public domain, needs to be secured. Few constraints and computations are required to maintain in the public domain during the information sharing happen between the participants.

One must know inputs from all the participants to conduct the constrained based security computations. However if nobody can be trusted enough to know all the inputs, privacy will become a primary concern. One of the solutions to this is Secure Multiparty Computation (SMC). This SMC is based on cryptographic functionality which plays a major role in the context of privacy preserving data between different participants in sharing authorized data. SMC is a computational system in which the value based on

Corresponding Author: Balamurugan, M., Department of IT, Selvam College of Technology, Namakkal, India

individually held secret bits of information that compute multiple parties wish to jointly. With exponential increase in usage of public network information sharing (internet), need for cooperative computation becomes more demanding along with the security of the user's private data. These secure cooperative computations could occur between trusted partners, between partially trusted partners, or even between competitors.

MATERIALS AND METHODS

The goal of methods for Secure Multi-party Computation (SMC) is to enable parties to jointly compute a function over their inputs, while at the same time keeping these inputs private. SMC refers to computational systems in which multiple parties wish to jointly compute some value based on individually held secret bits of information, but do not wish to reveal their secrets to one another in the process. SMC problem deals with computing any probabilistic function on any input, in a distributed network where each participant holds one of the inputs, ensuring independence of the inputs, correctness of the computation and that no more information is revealed to a participant in the computation than can be inferred from that participant's input and output. However success of privacy preserving data mining may depend on the ability to find new definitions that provide both the rigorous security guarantees that are necessary and can be met by highly efficient protocols.

Collaborative data mining: Advances in hardware technology have led to an increase in the capability to store and record personal data about consumers and individuals and this has led to concerns that the personal data may be misused for a variety of purposes. In order to alleviate these, a number of techniques have recently been proposed. These techniques for performing privacy preserving data mining are drawn from a wide array of related topics such as data mining, cryptography and information hiding. The progress of hardware technology has made it easy to store and process large amounts of transactional information in recent years. For example, even simple transactions of everyday life such as using the phone or credit-cards are recorded today in an automated way. Depending upon the nature of the information, users may not be willing to disclose the individual values of records. In particular, data mining techniques are considered a challenge to privacy preservation due to their natural tendency to use sensitive information about individuals (Keim, 2002).

While a large number of research work is carried out and now available in this field, many of the topics have been studied by different persons with different styles. At this stage, it becomes important to organize the topics in such a way that the relative importance of different research areas is recognized. Furthermore, the field of privacy-preserving data mining has been explored independently by the cryptography, database and statistical disclosure control communities. Due to the increase in the ability to store personal data about users, the problem of privacy-preserving data mining has become more important. A number of techniques such as randomization and k-anonymity have been suggested in recent years in order to perform privacy-preserving data mining. Furthermore, the problem has been discussed in multiple communities such as the database community, the statistical disclosure control community and the cryptography community. In some cases, the different communities have explored parallel lines of work which are quite similar (Agrawal and Aggarwal, 2001; Du and Atallah, 2001).

Data mining is the process of extracting patterns from the data and is an automated extraction of hidden predictive information from large data bases. Collaborative Data Mining is a data mining effort that is distributed to multiple collaborative agents that are either human or software. As multiple users/agents are involved, preserving privacy and transferring data/information without compromising user privacy is of great concern.

Secure multiparty computation: Recent advances in computing, communication have enabled large volumes of data to be accessible remotely across geographical boundaries. There is a great demand on collaborative data mining when compared to the distributed data stores to find the patterns or rules that benefit all of the participants. An important challenge for distributed collaborative data mining is how to protect each participant's sensitive information, while still finding useful data models. The techniques for performing privacy-preserving data mining are drawn from data mining, cryptography and information hiding. The service-oriented infrastructure for collaborative data mining of data distributed has become the most popular solution. Here the data providers are the collaborators who submit their own datasets to the required data mining service provider for discovering and mining the commonly interested models on the pooled data set. This model reduces the high communication cost associated with most cryptographic approaches (Chen and Liu, 2009).

The problem of privacy preserving data mining has become more important in recent years because of the increasing ability to store personal data about users. Number of techniques has been suggested in recent years in order to perform privacy preserving data mining. Multiparty secure computation allows N parties to share a computation, each learning only what can be inferred from their own inputs and the output of the computation. The core idea of the Secure Multiparty Computation (SMC) is secure and at the end of the computation, no participant has the knowledge except its own input and the results. The malicious adversary participant tries to capture the private information of authenticated participant or cause the computation made between the sharable participant's collaborative data to be incorrect.

The main contribution of this study is to securely compute multiparty information with privacy preservation in parallel by maintaining multiple sessions of participants. Each session is validated by trusted third party and the participants involved in it generate individual data independent rules and dependent sharable data rules to work effectively with collaborative data without losing individual's private data. A framework for secured multiparty computation process has been designed. An algorithm has been designed and developed for instance and dynamic rule generation for secure multiparty computation in collaborative data mining. Also, another algorithm for resisting internal adversaries in collaborative data mining has been designed and developed.

Secure two party computations were first investigated by Yao and were later generalized to multi-party computation. This research work uses a similar methodology: the functionality f to be computed is first represented as a combinatorial share and then the parties run a short protocol for every share. While this approach is appealing in its simplicity, the protocols it generates depend on the size of the share. This size depends on the size of the input which might be huge as in a collaborative data mining application. In multiparty collaborative data mining, participants contribute their own datasets and hope to collaboratively mine based on the pooled dataset. How to efficiently mine a quality model without interfering each party's privacy is the major challenge.

Collaborative data mining is distributed to multiple collaborate agents -human or software. The objective of the collaborative data mining it better when compared to the metric, with respect to those solutions that cannot be achieved using non collaborative data mining. The solutions require evaluation, comparison and approaches for combination. Collaboration requires communication and implies some form of community. The human form of collaboration is a social task and organizing communities in an effective manner is non-trivial and often requires well defined roles and processes. Based on the pooled data set, multiparty collaborative data mining participants contribute their own data sets and hope to collaboratively mine a comprehensive model.

Proposed algorithm: The proposed multiparty computation presented in this work is based on session framework which comprises of dependent and independent data objects that work cohesively and concurrently to fetch sharable data in collaborative data mining. The security concept introduced in this paepr preserves the participant's private data, on distribution of computing tasks, against an internal malicious adversary attacks on the collaborative data.

The framework of Session based Secured Multiparty Collaborative Data Computation (SSMCDM) comprises of participants, trusted third party, rule generation for private-sharable data, multi-sessions and instance generation. The participants are the users authenticated to involve in the data mining process and the trusted third party center is a common trust centre which provides session authentication to all participants. Rule generation deals with the formation of private data rules coined by the participants to preserve their privacy information (Bhuvana and Devi, 2011). Here, multi-session component that maintains more number of sessions for one or more participants in parallel performing various data mining tasks is considered in this study. The instance generation is dealt with, that cohesively handles both data dependency and independency rules to mine the authenticated sharable data of respective participant in any given session. The instance and multi-session components work together to preserve the private data of participant involved in single or multiple session of mining task simultaneously. In addition, the instance generation along with rule generation checks the internal malicious participant, who tries to invoke data in unauthorized sessions. The framework of secured multiparty computation process is presented in Fig. 1 where participant A requests the session key from the trusted center. The trusted center checks the session key status from the participant B after the session key verification and the session key is issued to participant A. Once, if the session key is generated for participant A, the privacy data rules are defined for participant A. Based on these, security is achieved during multiparty computation.

Rules generation: For every pair of participants which would like to engage in communication and share data between them, the privacy rule generator generates a corresponding rule and a session is established between the 2 participants.

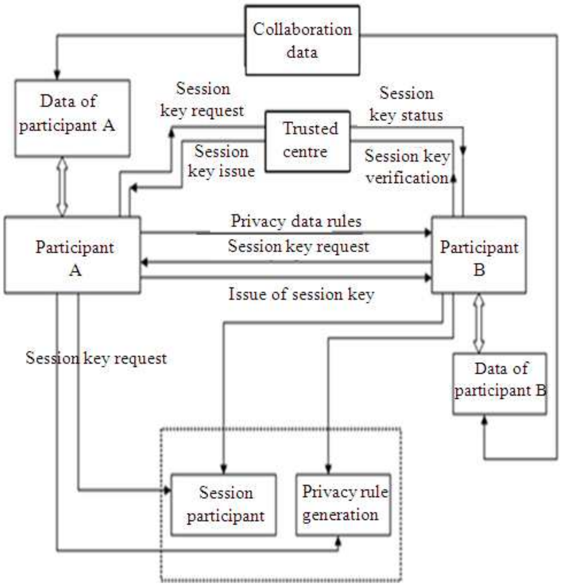

Fig. 1: Framework of secured multiparty computation process of collaborative data

The generated rules differ from different pair of participants. Let I = {I_1, I_2, . . . ,I_m} be a set of attributes, called items, let D be a set of transactions. Each transaction in D is a set T € I of items. An association rule is an implication of the form X → Y, where, X € I, Y € I and X∩Y= Ø. The rule X →Y has support's' in D if atleast s% of the transactions in D contain X U Y. The rule X→Y has confidence 'c' in D if atleast c% of the transactions in D that contain X also contain Y. The rules are generated to find all associated items between the participants involved in the session with support and confidence value with specified thresholds of minimum support and minimum confidence.

Data sharing: Data are collected from distributed databases using distributed network. The collected data, based on set of operations between distributed databases, can be shared among the parties during collaborative data mining and the information that can be shared is highly restricted when parties claim privacy on the data they hold. The secure distributed data sharing, aims to discover and share some information on the data distributed among different parties, while preserving each party's privacy on its data and not publishing them to everyone.

In SSMCDM, once a rule is generated, a share key is given to each participant so that it knows which participant wants to communicate with which other participant. Normally, share key is different for every participant and share key indicates which participants can communicate with each other. In SSMCDM framework, participant A can communicate with participant B and that communication takes place only between participant A and participant B. At that time participant A cannot communicate with participant C whereas participant C can share data with participant D. In general, it refers to computational systems in which multiple parties can jointly compute some value based on individually held secret bits of information but do not wish to reveal their secrets to one another in the process.

Data dependency: The data dependency is a condition where the value of a calculation in the future is based upon a calculation being made now or conversely a value of a calculation being made now depends on a value calculated in the past. In data mining, data dependency on shared data becomes the key in collaborative data mining and data dependent rules generated on the session is formed between the common sharable data of the involving participants. Data dependencies cause attack feasibility for malicious adversary in parallel. The attack feasibility is thwarted by the generated instance which validates the authenticity of the participants to share the collaborative data.

Dynamic rule generation: The meaning of dynamic rule generating instances is to develop the rule dynamically for each and every instance and called as the privacy rules for session based collaborative data mining. The process starts with the creation of session between any two users, for example sessions between the user A and the user B, between user C and D and so on. After the creation of session, the rules are generated for each session, for example rule1 is generated between user A and user B and rule2 is generated between user C and D and so on and possibility of the cross session variance to be introduced.

In cross session variance, user A is communicating with user B and at the same time user A also communicating with user C where user C is communicating with user D and cross session variance gets introduced, also another occurrence stated as the data independency rule and data dependency rule. For example, User A communicates with user B and user C communicates with user D. In independency rule, user A sends x_1, x_2 to user C and at the same time user A sends x_2, x_3 to user B; and data dependency rule is generated by means of combining users private data i.e., x_1, x_2 of user A and y_1,y_2 of user B to form an appropriate session based data sharing rule set and the sharing rule is called as dependency rule.

The session based framework model presented in Fig. 1 need to maintain adversarial entity controls to avoid the subset of the parties to attack the multiparty computation. Secure computation should withstand adversarial attacks and prove that private data of the participant is secured. The main component involved in the security issues of multiparty computation is the privacy component which specifies that no party should get information other than its prescribed output. In correctness component, each participant is guaranteed that the output it receives is correct. Independence of Inputs component refers that the corrupted parties must choose their inputs independently of the authenticated participant inputs. The output of mining the shared data resists the malicious adversaries to provide polluted data in the collaborative data and restricts the internal malicious participant in getting information from the unauthorized share data.

Parameters used in this computation are share key of the participants involved in collaborative data mining, privacy data functions, session identifier and adversary queries. Privacy data function is a client instance that accepts when it gains sufficient information to compute a share key. It should be noted that the state of acceptance only appears in client instances and moreover, a client instance is accepted at any time and only once. The session identifier uniquely names a session and the identifier is used by a participant. The session identifier works on an individual instance in an execution of secured multiparty computation. The participant identifier names the participant with which a client instance affirms that it has just shared a session key.

RESULTS AND DISCUSSION

The experiment is conducted with two modes of privacy preservation with classical key model and session based secured multiparty collaborative data mining key model. The classical model is the traditional cryptographic Rivest, Shamir and Adleman (RSA) method. The secured multiparty computation allows explicit mutual authentication and the results obtained are listed in Table 1. The adversary resistance rate indicates the effect of secured multiparty computation in maintaining the privacy of the participants involved in mining the complex collaborative data. The model of session based secure multiparty collaborative data mining model shows better resistance rate compared to that of the classical key model. The communication cycles required for the two models show that less number of cycles is enough for the session based secured multiparty collaborative data mining model for preserving the private data of individual participants compared to that of the other model.

Communication time: Communication Time for data transfer is calculated by measuring the time in (milliseconds) taken to transmit data from one participant to another participant through valid sessions.
Communication Cycle Measurement
Proposed work:

Cycle 1: Sender request to the trusted centre for obtaining session key
Cycle 2: Sender transfer the data along with session key to the receiver
Cycle 3: Verification of data authenticity by the receiver from the trusted centre

For Classical Cryptography:

Cycle 1: Authentication Request Made by the sender
Cycle 2: Sender transmit the crypt data to the receiver
Cycle 3: Acknowledgement by the receiver to sender
Cycle 4: Evaluation of crypt data by the receiver with crypt key

Adversary resistance rate: Measured in terms of restricted number of adversary participants trying to invade unauthorized sessions in a given time frame (say 10 min of Interval).

Adversary resistance rate in percentage: (No of resisted adversaries/ Total no of adversaries introduced) * 100.

In Table 1, the efficiency of the session based secured multiparty collaborative data model shows its minimum execution time required to mine the collaborative data as per the rule constraints set by individual session participants concurrently. In addition, from the table, the session based secured multiparty collaborative data model adapts smaller key size in the communication computation of session maintenance between various tasks.

The efficiency of the session based secured multiparty collaborative data mining model along with preservation of the privacy of all participants is shown in Table 2. The number of participants involved in the mining domain versus communication time for sharing the data shows the efficiency factor. Figure 2 shows that the secured multiparty communication model has minimal time compared to that of classical key model. In addition, as the number of participants increase, the execution communication time also increases and model shows better result than the classical key model.

Table 1: Performance results of session based secured multiparty collaborative data communication model on preserving privacy

Performance parameters	Classical key security model	Session based secured multiparty collaborative data model
Adversary Resistance Rate	83%	91%
Number of Communication cycles	4	3
Execution time for authenticated data sharing (Hundreds of Keys)	6 m sec	2 m sec
Share Key Length (in bits)	16	12

Table 2: Efficiency of secured multiparty communication in preserving the privacy compared to classical key model

	Communication time (milliseconds)	
No. of participants	Classical RSA key model	Session based secured multiparty collaborative data
10	5.20	2.60
20	5.70	3.00
30	5.99	3.32
40	6.32	3.60
50	6.57	3.80
60	6.88	4.14
70	7.21	4.30
80	7.50	4.65
90	7.77	4.90
100	7.90	5.20

Table 3: Effectiveness of session based secured multiparty collaborative data against classical key model

	Adversary resistance rate	
No. of participants	Classical RSA key model	Session based secured multiparty collaborative data
10	10.10	11.40
40	6.90	9.20
60	5.40	8.34
80	5.08	8.18
100	5.02	8.02

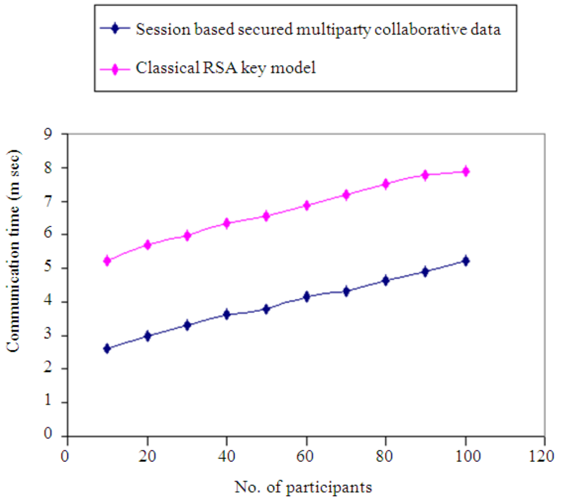

Fig. 2: Efficiency of secured multiparty communication in preserving the privacy compared to classical key model

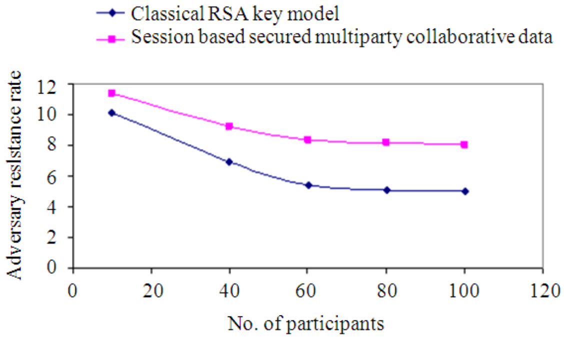

Fig. 3: Effectiveness of session based secured multiparty collaborative data against classical key model

The effectiveness of the session based secured multiparty computation is depicted in Table 3 by means of computing the session participant complexity against the adversary resistance rate. The session participant involved in the mining of collaborative data indicates the number of participants available in a specified session. It also contains the rule constraints applied by participants to maintain their respective data privacy. The adversary resistance rate is obtained during the collaborative data mining and it is measured as the rate of resistance against the internal or external adversaries to maintain the privacy element of its participant in mining the union of private databases.

The session based secured multiparty computation model requires that the trusted centre and each participant pre-share a sequence of shared key pairs. The length of the shared key pairs is measured and key pairs need to be reconstructed by the trusted centre and the participant after each session of execution. The experimental results obtained clearly show that the proposed method for privacy preserving in secure multiparty collaborative data mining model yields better response time for multiple message communication between the user queries. This response time is a function of the number of keys and number of system participants. The performance analysis of proposed session based secured multiparty collaborative data mining model and comparative study of the same with the classical key models, are made in terms of communication time (shown Fig. 2) between the participants involved in various sessions and the effectiveness of number of participants and session participant against malicious adversaries of internal active and passive participants shown in Fig. 3. The proposed secure multiparty computation method for collaborative data mining environment is efficient than the classical RSA key model. The methodology proposed here can be applied to social networking communications such as face book and twitter where privacy is a key component.

CONCLUSION

The problem of privacy preserving data mining has become more important in recent years because of the increasing ability to store personal data about users. Number of techniques has been suggested in recent years in order to perform privacy preserving data mining. Secure multiparty computation allows N parties to share a computation, each learning only what can be inferred from their own inputs and the output of the computation. The core idea of the Secure Multiparty Computation (SMC) computation is secure and at the end of the computation, no participant has the knowledge except its own input and the results. The malicious adversary participant tries to capture the private information of authenticated participant or cause the computation made between the sharable participant's collaborative data to be incorrect. The privacy rules generated in the proposed collaborative multiparty computation verifies the authenticity of participants and restricts or avoids the internal malicious adversary participant to involve in the mining.

A software prototype has been developed to show the concepts and the algorithm for secure multiparty computations. The software prototype has been designed using a top-down approach and programs have been developed using a bottom-up approach. The prototype consists of 5 modules: session key generation module, session id generation module, authentication module, message transmit module and resist malicious adversary module. The software prototype has been tested with test data. The results clearly show that there is an improvement of 8-10% efficiency in communication time for a range of 10-100 users and an improvement of 8% effectiveness in adversary resistance rate. Thus, the proposed concepts, algorithms and software prototype to secure multiparty computation in collaborative data mining are successful.

REFERENCES

Agrawal, D. and C.C. Aggarwal, 2001. On the design and quantification of privacy preserving data mining algorithms. Proceedings of the 20th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, May 21-23, ACM, Santa Barbara, California, USA., pp: 247-255. DOI: 10.1145/375551.375602

Bhuvana, J. and T. Devi, 2011. Performance of secure multiparty computation for preserving privacy in collaborative data mining. Int. J. Res. Rev. Comput. Sci., 2: 463-469.

Chen, K. and L. Liu, 2009. Privacy-preserving multiparty collaborative mining with geometric data perturbation. IEEE Trans. Parallel Distribut. Syst., 20: 1764-1776. DOI: 10.1109/TPDS.2009.26

Du, W. and M. Atallah, 2001. Privacy-preserving cooperative statistical analysis. Proceedings of the 17th Annual Computer Security Applications Conference, Dec. 10-14, ACM, IEEE Computer Society Washington, DC, USA., pp: 102-102.

Keim, D.A., 2002. Information visualization and visual data mining. IEEE Trans. Visualizat. Comput. Graphics, 8: 1-8. DOI: 10.1109/2945.981847

Sybil, S. and H. Lieberman, 2001. Intelligent profiling by example. Proceedings of the 6th International Conference on Intelligent user Interfaces, Jan. 14-17, ACM, Santa Fe, NM, USA., pp: 145-151. DOI: 10.1145/359784.360325

Yasien, A.H., 2007. Preserving privacy in association rule mining. Ph.D Thesis, University of Griffith.