Indian Journal of Science and Technology, Vol 9(48), DOI: 10.17485/ijst/2016/v9i48/108354, December 2016
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

A Review on Privacy Preserving Data Mining using Secure Multiparty Computation

U. Kumaran* and Neelu Khare

School of Information Technology and Engineering, VIT University, Vellore - 632014, Tamil Nadu, India;
kumaran.u@vit.ac.in

Abstract

With the widespread habituate of data mining technology in the entire available sectors (public and private) elevate concerns about the sensitiveness of data being mined. Data mining is an enormously powerful technology to extract information from raw data. With the growth of ease of handiness of digital data the possibility of misapply of the data and the mined information grows. A key challenge is to build up security and privacy methods suitable for data mining. This is the reason PPDM (Privacy Preserving Data Mining) has acquired a steam in recent times. In this paper we have addressed the issue of PPDM and moreover, we have considered a scenario where two different parties possesses confidential databases of their own and wish to run a data mining algorithm on the union of their databases, without disclosing any unnecessary information, where we have suggested different methodologies in order to preserve the privacy in the data mining process one among them is Secure Multiparty Computation which is a field of cryptography. PPDM looks at the job of applying data mining algorithms on secret (confidential) data i.e. not granted to be disclosed even to the trusted party who's running the algorithm.

Keywords: Cryptography, Data Mining Algorithm, Data Mining, Digital Data, PPDM (Privacy Preserving Data Mining), Secure Multiparty Computation

1. Introduction

Privacy Preserving Data Mining is one of the emerging fields in today's technological era as the data is generated in rapid amount because of the presence of automated devices such as credit cards, mobile phones, etc, which collects data automatically and services like e-mails, online transactions, etc, used in business and sciences are capable of generating huge amount of data like: Terabytes per hour. The data which is generated by all these devices and services can be very useful in order to make certain valuable organizational decisions example: Consider the data which is generated by credit card transactions and on the basis of most common transactions performed by a particular credit card we can recommend some offers which is beneficial for organization as well as the customer, this work is done by data mining algorithms. But here comes one more restriction, not all the data is easily available some confidential data is also there which is not meant for public usage like the databases of secret intelligence agencies and the databases of patients held by hospitals which is not meant to be disclosed to anyone but this data is very important for medical research which includes finding cures of diseases, finding the disease which is widespread in a particular region and then finding the reason behind it but this can't be done because the hospitals were abide by the law that they can't share the information of patients (due to the confidentiality of patient records) to any other hospitals but in order to find insights of medical research we need to combine the databases of all the hospitals i.e. the union of the databases then we need to run some data mining algorithms to get the results and taking care of the fact that none of the party can see other's database and this cannot be done

*Author for correspondence

by the classical data mining solutions, this scenario is a multiparty computation i.e. needed to be secured and that security is provided by Secure Multiparty Computation protocols.

2. Scenario

Here we consider that there are two hospitals one is H_1 and other is H_2 in Chennai and both of the parties (hospitals) has their own confidential databases D_1 and D_2 and they both wish to combine their databases and then run data mining algorithm in order to get the result while keeping their own inputs secret. Here the hospitals are trying to find the prone zone of a particular disease (example: Malaria) in one of the region in Chennai so in order to find it they have to combine their databases and then they need to check the patient's location who's suffering from a particular disease and then they need to run certain data mining algorithms to generate the result for these type of scenarios it is always simple to collect all the data in one place and then running data mining algorithm on the collected data. However, this is not what we wanted to do (H_1 and H_2 are not allowed to disclose their data out). Here comes a question, then how we are going to compute results without collecting data in one place (polling) in a manner that disclose nothing but the data mining computations final results. The scenario is visualized below:

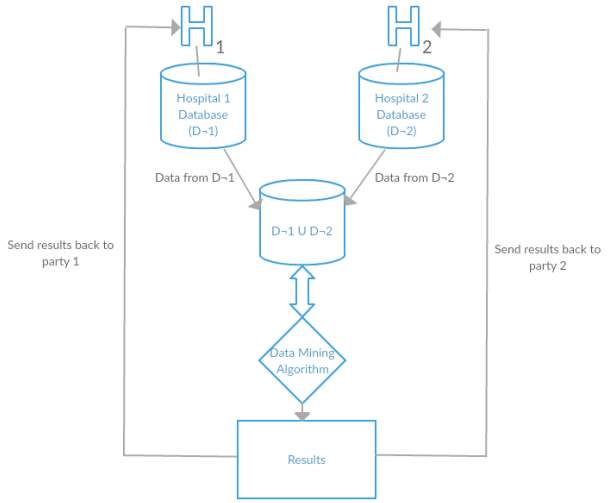


Figure 1. Scenario.

This question of PPDM is a special case of SMC (Secure Multiparty Computation).

2.1 Secure Multiparty Computation

SMC is a special field of cryptography which is meant for creating methods for parties to compute result from the union of their inputs by keeping their inputs private. Now one question comes into mind that how we can say that the computation is secured? One way of doing this is to list some security properties which should be satisfied while the transaction happens. When we talk about security of computation in Secure Multiparty Computation the properties that often comes in mind is: Input privacy and correctness.

2.1.1 Input Privacy

It refers to the inputs which we are going to combine in order to get the result from the union of it. By this property we mean that during the computation we may have some secure transactions which happen in order to calculate the final result and during these transactions the input of one party should not be visible to other party at any cost.

2.1.2 Correctness

It refers that the output should not be an incorrect result.

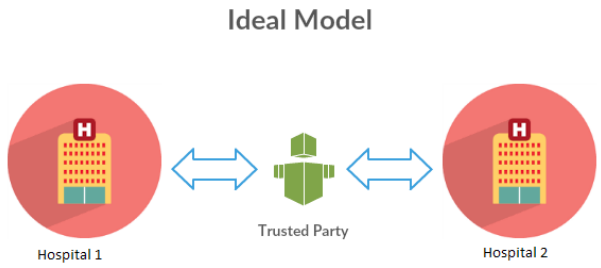
2.2 The Ideal/Real Model Paradigm

Due to the difficulties of providing security in multiparty computation the secure computations follow the approach of Ideal/Real model paradigm and this is one of the dominant paradigms from past few years in order to provide secure computation between multiparty. Here we consider two scenarios first is the ideal computing scenario and second is the real computing scenario and for both of them we follow some certain model for first scenario we follow ideal model and for second we follow real model.

2.2.1 Ideal Model

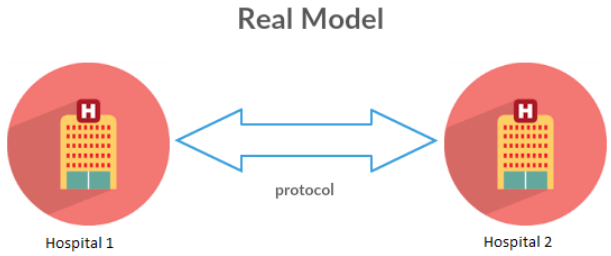
Now the solution of our main scenario i.e. securing the computation between two hospitals in order to find the result while keeping their databases secret. Ideally the solution of the problem is to have a trusted party (which is trusted by both the hospitals) to do the computation on the union of their databases i.e. Hospital H_1 will send its database to the trusted party and Hospital H_2 will send its database to the trusted party and then trusted party is

having the union of their databases then it will compute the results and send back the results to both the hospitals.



2.2.2 Real Model

In real world we can't trust anyone so Ideal Model won't work in real world especially in today's era. Rather, we can make use of some protocol approved by both the parties followed by some series of secured messages which won't reveal anything to anyone. Yet we can say in real world we use real model in order to emulate an ideal execution.



3. Conclusion

In this paper we have presented the use of Secure Multiparty Computation in order to preserve privacy in data mining by considering a scenario where two hospitals wants to compute a result from the union of their databases which is meant to kept private. Privacy preservation in data mining is comparatively a new field and with the continuous growth of e-data (digital data) the privacy issues will also increase and in this field not enough work has done. As digital data increases it will impose new security threats which are needed to be solved by continuous research in PPDM.

4. Future Scope

This field requires a continuous research in all the sectors, more algorithms are needed to be developed and analyzed. The growth of data from all domains imposes new problems on the privacy of data and this can only be solved by more researches. This field is growing and is open to get explored.

5. References

1. Saxena N, Gupta P, Singh O. A survey on security techniques in data mining. International Journal of Advanced Scientific Research and Management. 2016; 1(5):159–62.
2. Ouda MA, Salem SA, Ali IA, Saad ESM. Privacy Preserving Data Mining (PPDM) method for horizontally partitioned data. IJCSI International Journal of Computer Science Issues. 2012 Sep; 9(5):339–47.
3. Lindell Y, Pinkas B. Secure multiparty computation for Privacy-Preserving Data Mining. The Journal of Privacy and Confidentiality. 2009.
4. Agrawal R, Srikant R. Privacy Preserving Data Mining. IBM Almaden Research Center; 2000. p. 439–50.
5. Anju AP, Sasi Kala Rani K, Literature survey on privacy preserving cross domain optimization algorithms. IJCTT. 2015 Dec; 30(2):108–12.
6. Lindell Y. Secure multiparty computation for Privacy Preserving Data Mining. The Journal of Privacy and Confidentiality. 2009; 1(1):59–98.
7. Evfimievski A, Grandison T. Privacy Preserving Data Mining. IBM Almaden Research Center, USA.
8. Prakash VS, Shanmugam A. Privacy Preserving Data Mining using secure multiparty computation. VS-International Journal of Computer Science information and Engg Technologies.