A Survey: Privacy Preserving Data Mining

Komal Kapadia
Dept. of Computer Science
Gujarat Technological University
P.I.T., Vadodara, Gujarat State India

Raksha Chauhan
Dept. of Computer Science
Gujarat Technological University
P.I.T., Vadodara, Gujarat State India

ABSTRACT

Privacy preserving data mining techniques are introduced with the aim of extract the relevant knowledge from the large amount of data while protecting the sensible information at the same time. The success of data mining relies on the availability of high quality data. To ensure quality of data mining, effective information sharing between organizations becomes a vital requirement in today's society. Privacy preserving data mining deals with hiding an individual's sensitive identity without sacrificing the usability of data. Whenever we are concerning with data mining, Security is measure issue while extracting data. Privacy Preserving Data Mining concerns with the security of data and provide the data on demand as well as amount of data that is required.

General Terms

Data Mining, Privacy Preserving.

Keywords

MHS algorithm, EMHS algorithm, CK secure sum and Randomized response technique.

1. INTRODUCTION

Data mining techniques have been widely used in many areas especially for strategic decision making. The main threat of data mining is to security and privacy of data residing in large data stores. Some of the information considered as private and secret can be bought out with advanced data mining tools. Different research efforts are under way to address this problem of privacy preserving and preserving security. The privacy term has wide range of different meanings. For example, in the context of the health insurance accountability and portability act privacy rule, privacy means the individual's ability to control who has the access to personal health care information. In organization, privacy means that it involves the definition of policies stating which information is collected, how it is used, how customers are involved and informed in this process. We can consider privacy as "Individual's desire and ability to keep certain information about themselves hidden from others." Privacy preserving data mining refers to the area of data mining that seeks to safeguard sensitive information from unsolicited disclosure. Historically, issues related to PPDM were first studied by the national statistical agencies interested in collecting private social and economical data, such as census and tax records, and making it available for analysis by public servants, companies, and researchers. Building accurate socio-economical models is vital for business planning and public policy. Yet, there is no way of knowing in advance what models may be needed, nor is it feasible for the statistical agency to perform all data processing for everyone, playing the role of a "trusted third party". Instead, the agency provides the data in a sanitized form that allows statistical processing and protects the privacy of individual records, solving a problem known as privacy preserving data publishing. There are many methods for preserving the privacy. In this survey many methods try to compute the answer to the mining without revealing any additional information about user privacy.

Progress in scientific research depends on the sharing and availability of information and ideas. But the researchers are mainly focusing on preserving the security or privacy of individuals. This issue leads to an emerging research area, privacy preserving data mining. For privacy preserving data mining, many authors proposed many technologies. The main aim of this paper is, to develop efficient methodology to find privacy preserving.

2. EXISTING WORK

We have studied some of the related work for the privacy preserving in horizontally partitioned databases. Existing work for privacy preserving in horizonatally partitioned database has different types of techniques.

2.1 Types of Privacy Preserving Techniques

2.1.1 Semi honest party

2.1.2 Without trusted party

2.1.3 With trusted party

In without trusted party each party will calculate their own partial support and add their own random number and sends the result to the next party in the ring so that the other party will never know the result of others and in last the initiator party will disclose the result that is global support.

In trusted party each party will calculate their partial support and send to the trusted party and add the own random number and send to the next coming site in the ring so that other party will never know the result of other parties after that trusted party will disclose the result and send to all sites that presents in the ring.

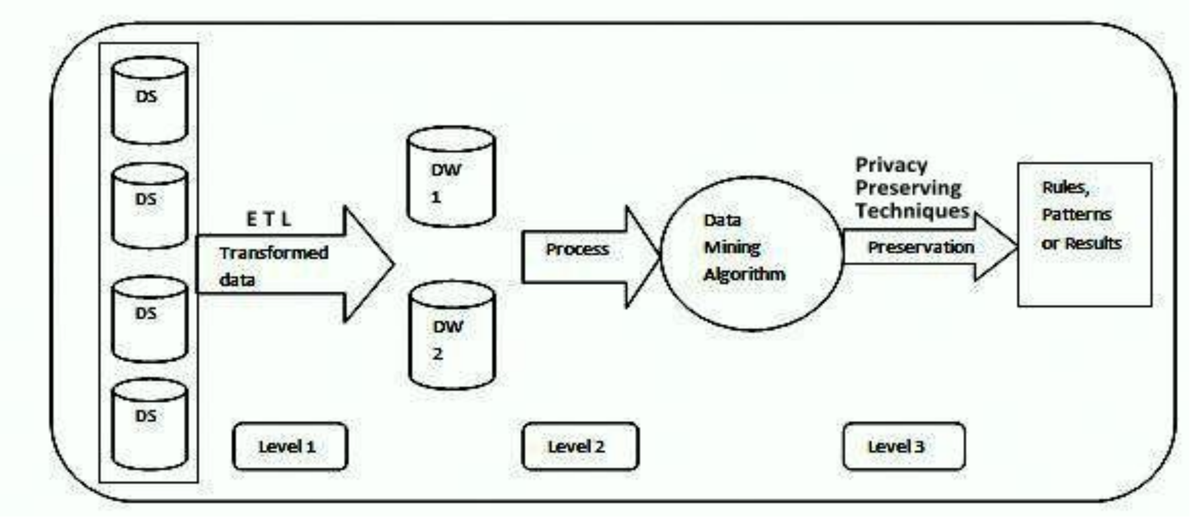

Fig 1: Framework of privacy preserving data mining [5]

3. SECURE MULTIPARTY COMMUNICATION

Approximately all Privacy Preserving data mining techniques rely on secure multi party communication protocol. Secure multi party communication is defined as a computation protocol at the last part of which no party involved knows anything else except its own inputs the outcome, i.e. the view of each party during the execution can be effectively simulated by the input and output of the party. Secure multi party communication has commonly concentrated on two models of security. The semi-honest model assumes that every party follows the rule of the protocol, but is free to later use what it sees during execution of the protocol. The malicious model assumes that parties can arbitrarily cheat and such cheating will not compromise moreover security or the outcome, i.e. the results from the malicious party will be correct or the malicious party will be detected. Most of the Privacy Preserving data mining techniques assume an intermediate model, Preserving Privacy with non-colluding parties. A malicious party May dishonest the results, but will not be able to learn the private data of other parties without colluding with another party.

4. MHS ALGORITHM FOR HORIZONTALLY PARTITION DATABASE

M. Hussein et al.'s Scheme (MHS) was introduced to improve privacy or security and try to reduce communication cost on increasing number of sites. Behind this main idea was to use effective cryptosystem and rearrange the communication path.

For this, two sites were discovered. This algorithm works with minimum 3 sites. One site acts as Data Mining Initiator and other site as a Data Mining Combiner. Rests of other sites were called client sites. This scenario was able to decrease communication time. Fig. shows MHS algorithm.

The working of the algorithm is as follows:

1. The initiator generates RSA public key and a private key. It sends the public key to combiner and all other client sites.
2. Each site, except initiator computes frequent Itemset and local support for each frequent Itemset using Local Data Mining.
3. All Client sites encrypt their computed data using public key and send it to the combiner.
4. The combiner merges the received data with its own encrypted data, encrypts it again and sends it to initiator to find global association rules.
5. Initiator decrypts the received data using the private key. Then it merges its own local data mining data and computes to find global results.
6. Finally, it finds global association rules and sends it to all other sites.

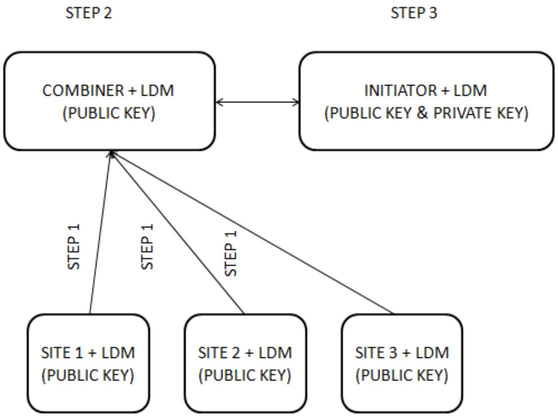
STEP 2
STEP 3
COMBINER + LDM (PUBLIC KEY)
INITIATOR + LDM (PUBLIC KEY & PRIVATE KEY)
STEP 1
STEP 1
STEP 1
SITE 1 + LDM (PUBLIC KEY)
SITE 2 + LDM (PUBLIC KEY)
SITE 3 + LDM (PUBLIC KEY)

Fig.2 MHS algorithm [11]

5. EMHS ALGORITHM FOR HORIZONTALLY PARTITION DATABASE

Enhanced M. Hussein et al.'s Scheme (EMHS) was introduced to improve privacy and reduce communication cost on increasing number of sites. This algorithm also works with minimum 3 sites. One site acts as Data Mining Initiator and other site as a Data Mining Combiner. Rests of other sites were called client sites. But this algorithm works on the concept of MFI (Maximal Frequent Itemset) instead of Frequent Itemset.

a. MFI (Maximal Frequent Itemset): A Frequent Itemset which is not a subset of any other frequent Itemset is called MFI. By using MFI, communication cost is reduced.

b. RSA (Rivest, Shamir, Adleman) Algorithm: one of the widely used public key cryptosystem. It is based on keeping factoring product of two large prime numbers secret. Breaking RSA encryption is tough.

6. MODIFIED EMHS ALGORITHM FOR HORIZONTALLY PARTITION DATABASE

In this technique, they used modified EMHS algorithm for improving its efficiency by using Elliptic curve cryptography. Here Elgamal cryptography technique is used which is of ECC for homomorphic encryption.

6.1 Elliptic Curve Cryptography

Elliptic curve cryptography provides public cryptosystem based on the descrete logarithm problem over integer modulo a prime. Elliptic curve cryptosystem requires much shorter key length to provide a security level same as RSA with larger key length. In this elgamal cryptography is used.

6.2 Elgamal Cryptography

a) A wishes to exchange message M with B[9].

b) B first chooses Prime Number p, Generator g and private key x.

c) B computes its Public Key Y = g^x mod p and sends it to A.

d) Now A chooses a random number k.

e) A calculates one time key K = Y^k mod p.

f) A calculates C1 = g^k mod p and C2 = M*K mod p and sends (C1,C2) to B.

g) B calculates K = $C1^x$ mod p

h) B calculates K^{-1} = inverse of K mod p i) B recovers M = K^{-1} * C2 mod P

j) Thus, Message M is exchanged between A and B securely.

In this system, Elgamal cryptography & paillier cryptosystem is used. Here, Elgamal cryptography is used for security purpose. Compared to EMHS algorithm here performance is better in terms of computation time.

7. RANDOMIZED RESPONSE TECHNIQUE

In this technique, here mainly focused on CK secure sum in randomized response technique for privacy preserving. Here, the multi party transaction data that discover frequent item sets with minimum support.

In the randomized response technique, consider the data sets I = {I1, I2, I3……In} and the random number or noise part are denoted by R= {R1, R2, R3 …Rn}, the new set of records are denoted by I1+R1, I2+R2 ….In+Rn and after that take a partial support Pij = {Pi1, Pi2………..Pin} so that partial support is

P ij=I+R[10]

I=Pij-R[10]

In Randomized response secure sum technique, secure sum each site will determine their own data value and send to predecessor site that near to original site and this goes on till the original site collects all the value of data after that the parent site will determine the global support.

7.1 CK Secure Sum Algorithm

```
Step1:-Consider parties P1, P2, P3…………Pn.
Step2:-Each party will generate their own random number R1, R2……….RN
Step3:-Connect the parties in the ring (P1, P2, P3………PN) and let P1 is a protocol initiator.
Step4:-Let RC=N, and Pij=0 (RC is round counter and Pij is partial support)
Step5:-Partial support P1 site calculating by using following formula
Psij = Xij.support – Min support * |DB| + RN1 – RNn
Step6:-Site P2 computes the PSj for each item received the
List using the formula,
PSij= PSij + Xij. Support – minimum support * |DB| +Rn1-Rn (i-1)
Step7:-While RC! =0 begin for j=1 to N do
          begin for I=1 to N do
Step8:-P1 exchange its position to P (j+1) mod N and
          RC=RC-1
          end
Step9:-Party P1 allowance the result Pij
Step10:-End
```

In ck secure sum technique, mainly focused on for computing global support within a scenario of homogeneous database and provides the high security to the database and hacking of data is zero.

8. CONCLUSIONS

In this paper, they reviewed five privacy preserving technique in horizontally partitioned database. In MHS algorithm RSA cryptography is used. In EMHS algorithm, by using MFI approach accuracy is high compared to MHS. Modified EMHS algorithm used elgamal technique so privacy is high than EMHS technique. Randomized response technique provides high security to the database compared to other techniques. In future they can compute less number of rounds instead of n number of rounds. Here, they can use encryption technique for encrypting random number and sends it to the predecessor.

9. ACKNOWLEDGMENTS

Authors are grateful to referees for their valuable comments.

10. REFERENCES

[1] Neelamadhab Padhy, Dr. Pragnyaban Mishra & Rasmita Panigrahi. “The Survey of Data Mining Applications and Feature Scope.” 2012 IJCSEIT.

[2] Xinjun qi, Mingkui zong. “An overview of privacy preserving data mining.” 2011 ICESE.

[3] Kishori pawar, Y.B. gurav. “Overview of privacy in horizontally distributed databases.” 2014 IJIRAE.

[4] Manish Sharma, Atul chaudhary , Manish mathuria & Shalini chaudhary. **“**A review study on the privacy preserving data mining techniques and approaches.”. 2013 IJCST.

[5] Shweta taneja, shashank khanna, sugandha tilwalia, ankita. “A review on privacy preserving data mining: techniques and research challenges.” 2014 IJCSIT.

[6] Jayanti dansana, Raghvendra kumar & Jyotirmayee rautaray. “Techniques for privacy preserving association rule mining in distributed database.” 2012 IJCSITS.

[7] Xuan canh nguyen, Tung anh cao. “An enhanced scheme for privacy preserving association rules minig on horizonatally distributed databases.” 2012 IEEE.

[8] Manish Sharma, Atul chaudhary, Manish mathuria, Shalini chaudhary & Santosh kumar. “An efficient approach for privacy preserving in data mining.” 2014 IEEE.

[9] Rachit v. Adhvaryu, Nikunj h. Domadiya. “Privacy preserving in association rule mining on horizontally partitioned database.” 2014 IJARCET.

[10] Jayanti Dansana , Raghvendra Kumar , Debadutta Dey. “Privacy preservation in horizontally partitioned databases using randomized response technique.” 2013 IEEE.

[11] Rachit v. Adhvaryu, Nikunj h. Domadiya, “Research Trends in Privacy Preserving in Association Rule Mining (PPARM) On Horizontally Partitioned Database”. 2014 IJEDR.

[12] Agrawal D. Aggarwal C. C. On the Design and Quantification of Privacy-Preserving Data Mining Algorithms.ACM PODS Conference, 2002.

[13] D.W.Cheung,etal.,Ecient Mining of Association Rules in Distributed Databases, "IEEE Trans. Knowledge and Data Eng., vol. 8, no. 6, 1996,pp.911-922.