Cryptography in Data Mining

Dr. K. Venkatesh Sharma  
Professor, TKRCET  
Email: venkateshsharma.cse@gmail.com

Dr. DVSS Subrahmanyam  
Professor, MREC,  
Email: subrahmanyam.dvss@gmail.com

A. Sandhya Rani  
Asst. Professor, TKRCET

Abstract – One of the fresh aspects in data mining study is to develop techniques relating to the obsessions of privacy preserving, particularly, relating to the fact that techniques of data mining could be able to create sound models when the precise information of data is inaccessible. As a result of this research, numbers of data mining techniques with respect to privacy preserving are introduced in this study. One of these techniques - suggested in this paper - are to utilize methods of cryptography in data mining with respect to privacy preserving in distributed databases. We assume that data are stored in some private participants, and these participants agree upon a specific sort of estimating of data mining where the private characteristic of arrivals is preserved, and only the result of data mining is to be revealed. Variety of techniques has already been introduced in this field.

This paper is to analyze and assess techniques of privacy preserving, introducing a framework based on methods of cryptography in data mining with respect to the privacy preserving. Considering the prevailing application of data mining methods in distributed databases, the suggested classification can possibly be influential in opting for a proper approach.

Keywords – Cryptography, Distributed Data Mining, Privacy Preserving.

I. INTRODUCTION

Applications, but the violation of privacy is feasible in the absence of enough preservation, and private data may be used for other purposes. Data mining is a threat for privacy when the sensitive attribute is directly accessible or the detected models are able to divulge the personal and confidential information of individuals or organizations [1, 2 and 3]. Therefore, it is necessary to prevent the disclosure of not only the confidential and personal information, but also the critical science of a specific field. To handle such obsessions, the data mining researchers have embarked on practicing methods which makes it to mine data and preserve privacy concurrently [2, 3].

A wide variety of different techniques and approaches are recognizable in the field of “privacy-preserving data mining” (PPDM) and many different threat and trust models are taken into consideration [1, 4 and 5]. These techniques use different methods: in perturbation, noise is added directly to the database that is input to the algorithm in some occasions and to the output of queries in some other occasions to blur the values of sensitive attributes. In generalization, defining attributes have less specific values. In cryptography, joint computations between multiple parties are performed on encrypted data in order to hide inputs [5].

In the process of data mining, methods based on cryptography for the purpose of privacy preserving in distributed databases are used. As depicted in figure 1, this process comprises some stages which make it a tripartite-tiered architecture. The data providers and the data owners are often physically distributed at the bottom tier. The private data is submitted to the data warehouse by the data providers. Constituting the middle tier, this server supports online analytical data processing to make data mining easy. The data collected in disciplined physical structure are stored by the data warehouse server. These data include multidimensional data cube, and aggregate. In next stage, these data are pre-computed by warehouse in various forms including sum, average, min, and max. The actual data mining operation is carried out by the data mining server at the top tier. Data mining server serves different purposes: it possibly shares information with data mining servers from other systems; it constructs data mining models on its local data warehouse server. As depicted in the figure, sharing occurs in the top tier. In this tier, each data mining server holds its exclusive system of data mining model [6].

Following this paper, in section 2, the most important techniques to privacy preserving in data mining based on cryptography will be thoroughly presented. In section 3, the practical criteria for analysis and assessment of prevailing methods are introduced, and based on these criteria; the mentioned approaches will be assessed. To conclude, in section 4, the result of this study and its possible implications will be suggested.

II. PROPOSED FRAMEWORK

Analysis of techniques of privacy preserving in the process of distributed data mining (PPDDM) shows that these methods can be explicated from four different...

A. Secret Sharing

It characterizes the capacity of a secret to be distributed amongst a group of participants where each of which is marked by a share of the secret [7, 8]. The secret reconstruction is made feasible when shares are assembled; individual shares solely are deemed to be useless, i.e. Shamir secret sharing pattern [8].

C(t, n) secret sharing scheme is a set of two functions S and R. the function S is a sharing function and takes a secret x as input and creates n secret shares: S(x) = (x₁, …, xₙ). The two functions are selected in a model that, for any union Ic{1… n} of t indices R(I, x₁, …, xₙ) = x, in addition we require that it is impossible to recover x from a union of t - 1 secret shares [9].

B. Public Encryption

In the process of encrypting data, secret data is translated utilizing an algorithm to a format being unreadable to everyone but those who have access to some information which is specialized being used to decrypt it (see figure 3) [10, 11].

A collection of probabilistic polynomial time algorithms for key generation, encryption, and decryption produce a public key cryptosystem P, taking the first letter being G, E, and D. The key generation algorithm G produce a private key being SK, and a public key being PK with a size which comes to be specified. The operation of decryption is granted to everyone where everyone is given the right to encrypt the message and read it, but the operation of decryption is solely granted to the holder of a private key where he or she has the privilege to decrypt and read the message in actuality. The encryption E takes an input, namely plaintext m, a random value, namely r, and a public key, namely PK, and subsequently the cipher text E_PK (m, r) is output correspondingly. The decryption algorithm D takes an input a cipher text, namely c, a private key, namely SK and a plaintext D_PK(c) is output. It is required, therefore; D_PK (E_PK (m, r)) [11, 12].

1) Homomorphic Encryption: A public-key cryptosystem is defined to be homomorphic when one is given the space to perform an operation which is specifically algebraic on the plaintext by performing a (possibly different) operation in algebraic fashion on the cipher text. Needless to have a secret key, this distinct attribute permits a party to add or multiply plaintext by means of a couple of simple computations with cipher text [11]. This kind of encryption is defined as follows [13, 14]:

F(m, k) = (g^m x mod p, g^k mod p) = (G, H) = M (1)

y = g^p is a public key; x is a private key in the range [1, p – 2] and p is a large random prime number;

m is the message to be encrypted;

k is chosen uniformly at random in the range [1, p – 2];

g is a generator for a cyclic group G whose discrete logarithm is hard to compute;

G, and H are the pair of cipher texts M (encrypted texts).

2) Communicative Encryption: when a cryptosystem composition of the encryption with two non-similar keys resembles each other, ignoring the order of encryption, it’s called communicative; that is, the encryption algorithm E which takes as input plaintext M for any non-similar encryption keys pk₁, pk₂, … and any permutations of i, j, the same cipher text would be the result [11, 15, 16].

E_PK₁ (… E_PK₃ (M) …) = E_PK₂ (… E_PK₃ (M) …) (2)

∀ M₁, M₂ ∈ M such thatM₁ ≠ M₂ and for given pk ∈ 1/2^pk

Pr (E_PK₁ (… E_PK₃ (M) …) = E_PK₂ (… E_PK₃ (M) …)) < ε (3)

C. Randomization

This method is characterized by the additional noise to the original data for hiding its real value in order to protect privacy of the data sets. In other words, randomization is a process where the input data perturbs to distributed data mining algorithms; therefore, the values of data in individual entities are protected to be revealed [8, 17 and 18].

In this method, the data collection process consists of two steps [19]. The step 1 is for the data providers to randomize their data and transmit the randomized data to the data receiver. In the step 2, the data receiver estimates the original distribution of the data by employing a distribution reconstruction protocol (see figure 4).
D. Oblivious Transfer Protocol (OT)

A secure protocol existence is defined between two sections. One of the sections possesses one secret b bit, and the second section can know this b bit with probability of ½ or know nothing about it at the end of computation. The main point is that the server has no information of which of these two events have occurred, and the user does not know anything about other bits in database. This mood is an instance of 1-out-of-2 oblivious transfer protocol. Other transfers of this protocol (oblivious transfer protocol 1-out-of-N, the distributed oblivious transfer protocol, and so forth) are recognizable to resolve security problems [8, 20]. This protocol comprises two sections of the sender and the receiver. The sender input is a couple (x₀, x₁) and the receiver input is σ ∈ {0, 1}. At the end the receiver protocol does not know anything but x₀, and the sender knows nothing. In other words, if we apply symbols (input₁, input₂) → (output₁, output₂) to define the result of a function, the oblivious transfer is the function ((x₀, x₁), σ) → (λ, x₀) where λ is the empty output [21].

III. ASSESSMENT OF PRIVACY PRESERVING TECHNIQUES

In this section, we evaluate privacy preserving techniques according to main functional criteria. Our assessment is summarized in Table 1.

A. Proposed Criteria

To assess aforementioned techniques, the following criteria are taken into account, and ranking is done in three different tiers – low, medium and high.

- Cost of encryption: the cost of applied algorithms for encryption of data and sensitive information participants are unlikely to disclose them.
- Rate of input change: the degree of change in input data of participants in the process of data mining.
- Performance: to employ algorithms by appropriate coding.
- Privacy Preserving: the degree of privacy preservation, applying methodologies to curb the disclosure of participants’ exclusive information in the process of data mining.
- Accuracy of the results of data mining: it is the level of validity and accuracy of the results in the end of apply data mining protocols.
- Scalability: if the performance of a process won’t decrease with increasing the amount of participant parts in the data mining, this process will be scalable.

- Complexity: it is the total of computation and communication cost. Communication cost are dependent on the amount of sent message between different sites in the data mining process and the computation cost are calculated according to the amount of the encryption and decryption process which are done.

Table I: Evaluation framework for the privacy preserving approaches based on cryptography in ppdm

<table>
<thead>
<tr>
<th>Privacy preserving Technique</th>
<th>Cost of Encryption</th>
<th>Rate of Input Change</th>
<th>Performance</th>
<th>Privacy Preserving</th>
<th>Accuracy</th>
<th>Scalability</th>
<th>Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Public Key Enc.</td>
<td>Hom</td>
<td>med</td>
<td>high</td>
<td>high</td>
<td>high</td>
<td>high</td>
<td>med</td>
</tr>
<tr>
<td>Com</td>
<td>high</td>
<td>med</td>
<td>high</td>
<td>high</td>
<td>high</td>
<td>high</td>
<td>med</td>
</tr>
<tr>
<td>OT Protocol</td>
<td>high</td>
<td>high</td>
<td>med</td>
<td>high</td>
<td>low</td>
<td>high</td>
<td>low</td>
</tr>
<tr>
<td>Secret Sharing</td>
<td>low</td>
<td>low</td>
<td>high</td>
<td>high</td>
<td>high</td>
<td>med</td>
<td></td>
</tr>
<tr>
<td>Randomization</td>
<td>high</td>
<td>low</td>
<td>high</td>
<td>low</td>
<td>high</td>
<td>low</td>
<td></td>
</tr>
</tbody>
</table>

IV. CONCLUSION

In this paper, privacy preserving techniques in data mining process in distributed databases are introduced. These techniques are proposed on Secret Sharing, Public Key Encryption, Randomization and Oblivious Transfer Protocol. According to the proposed framework, the place of each approach with respect to criteria with which the approach is assessed, is analyzed. Therefore, providing the possibility of appropriate application from privacy preserving techniques based on needs is of influential findings of this study. Challenges to privacy preserving techniques, high cost of encryption and high complexity of computation and communication, can possibly be enumerated.

Nonetheless, the permanence of current research as an incentive to mount up balance amongst privacy preserving and complexity of algorithm can be a vigor brought up as still rooms for further works.

REFERENCES


[18] A. Gurevich, E. Gudes, “Privacy preserving Data Mining Algorithms without the use of Secure Computation or Perturbation”, In: 10th International Database Engineering and Applications Symposium, pp. 121-128, 2006