Privacy Preserving of Sensitive Data Using Encryption Techniques

Jyothi Mandala
Department of IT, GMRIT, Rajam, AP, India.

Abstract – Data Mining is a process of extracting useful information from large data repositories. Privacy-preserving Data Mining is a research direction in data mining where data mining algorithms can be applied without compromising with the privacy of the data. Data can be stored in centralized or distributed database. This paper includes privacy preserving data mining (PPDM) technique using encryption approach which would be efficient in providing confidentiality.

Index Terms – PPDM, data privacy, Data Mining, Encryption.

1. INTRODUCTION

Data mining research deals with the extraction of useful information from large collection of data. Organizations want to analyse their data by external agent. The data to be analysed may contain some sensitive individual information which may be exposed during the data mining process. Hence it is possible to learn lot of information about individuals from public data. Privacy preserving Data Mining (PPDM) can be applied on centralized or distributed data by providing privacy of the sensitive data. In paper [1] PPDM is defined as "getting valid data mining results without learning the underlying data values". PPDM achieves the dual goal by providing the privacy requirements and producing valid data mining results.

PPDM can be one of the three approaches: 1. Data hiding, in which sensitive raw data like identifiers, name, addresses, etc were altered, removed or cut out from the original database in order for the users of the data not to be able to compromise another person's privacy. 2. Rule hiding, in which sensitive data extracted from the data mining process is not used for use, because confidential information may be derived from the released knowledge and 3. Secure multiparty computation (SMC), where distributed data are encrypted before released or shared for computations, so that no party knows anything except its own inputs and the results.

PPDM algorithms are categorized as

1. Data Distribution: The data can be stored as centralized data or distributed data. Distributed data scenarios can be divided as horizontal data partition and vertical data partition. Horizontal distribution refers to these cases where different sets of records exist in different places, while vertical data distribution refers where all the values for different attributes reside in different places.
2. Data Modification**:** Data modification is used when there is change in the unique values of a database which are used for public data mining and in this way this guarantees high privacy protection. Methods of data modification include:

(a). Perturbation: This method changes the attribute value with a new value. This can be done by using Additive noise based technique [2] or multiplicative noise based technique [3].

(b). Blocking: which is the replacement of an existing attribute value with a "?".

(c). Swapping: This refers to interchanging values of individual record.

(d). Sampling: This refers to losing data for only sample of a population.

(e). Encryption: Many Cryptographic techniques are used for encryption.

3. Data Mining algorithms:

We can perform Association analysis, clustering and classification on data.

Based on these dimensions, different PPDM techniques may be classified into following five categories [4] [5]

1. Anonymization based PPDM

2. Perturbation based PPDM

3. Randomized Response based PPDM

4. Condensation approach based PPDM

5. Cryptography based PPDM

2. CRYPTOGRAPHY-BASED TECHNIQUES

A cipher is an algorithm that is used to encrypt plaintext into cipher text (encryption) and cipher text to plain text (decryption).

Ciphers are said to be divided into two categories: private key and public key.

Private-key (symmetric key) algorithms require a sender to encrypt a plaintext with the key and the receiver to decrypt the cipher text with the same key. A problem with this method is

that both parties must have an identical key, and somehow the key must be delivered to the receiving party.

Example algorithms are DES, AES.

A public-key (asymmetric key) algorithm uses two separate keys: a public key and a private key. The public key is used to encrypt the data and only the private key can decrypt the data. A form of this type of encryption is called RSA.

Cryptographic based technique is used if two or more parties want to perform data mining task on combined datasets. This problem is referred as Secure Multi-party Computation (SMC) problem[6]. By using cryptographic techniques we can perform privacy preserving classification [7], privacy preserving association rule mining [8] and privacy preserving clustering [9].

Secure multi-party computation has two models: A semi-honest participant will not deviate from the protocol but will only try to extract some extra information from the messages On the other hand; a malicious adversary can arbitrarily deviate from the protocol.

a) Public-key cryptosystems (asymmetric ciphers)

A classic example for PPDM is Yao's millionaire's problem: two millionaires want to find out who is richer without revealing toeach other how many millions they each own. In [10] a solution to the Yao's millionaire problem is given.

Ashraf B. El-Sisi and Hamdy M. Mousa [11] proposed a cryptographic approach for PPDM which uses a semi-honest model. This employs a public-key cryptosystem algorithm on horizontally partitioned data among three or more parties. The approach is as follows:

Consider three parties A, B and C

- A generates the public key KPA. This KPA is known to B and C.
- Now A, B and C encrypts their dataset DB_i with KP_A key. Encryption is applied on each row of the dataset. This encryption is denoted as $KP_A(DB_i)$ as shown in Fig 1. Only A can perform decryption on these datasets as A only knows his private key.

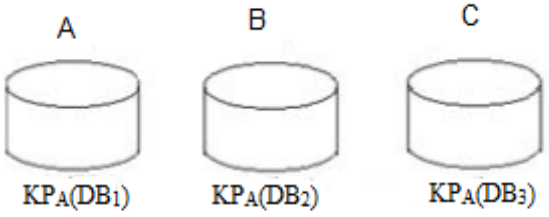

Fig 1: A, B and C encrypts their data sets

- A passes his encrypted dataset i.e. $KP_A(DB_1)$ to B.
- Now B performs random shuffle of $KP_A(DB_1)$ and $KP_A(DB_2)$ and forwards the resultant dataset to C as shown in Fig 2.

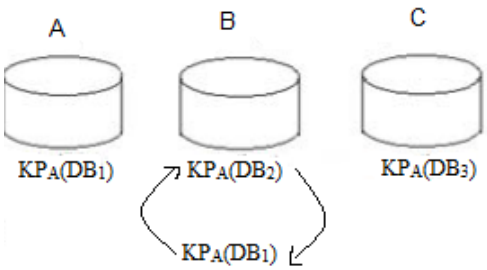

Fig 2: B shuffles the data sets transactions.

- C adds and shuffles hid dataset transactions $KP_A(DB_3)$ to the transactions received from B as shown in Fig 3.

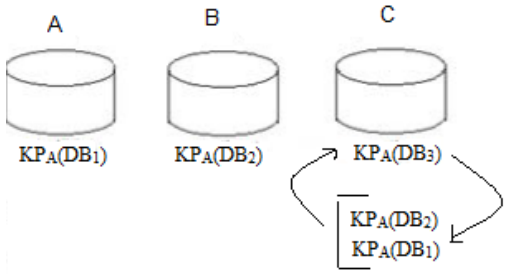

Fig 3: C shuffles the datasets transactions

- C forwards these transactions back to A.
- A decrypts the entire dataset with his secret private key as shown in Fig 4. A can identify his own transactions. However, A is unable to link transactions with their owners because transactions are shuffled.
- Finally A publishes the transactions to all other parties.

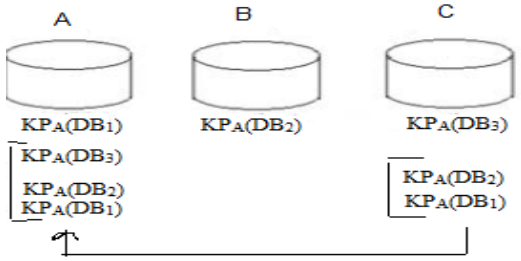

Fig 4: A performs the decryption.

Using the above approach the information that is hidden is what data records where in the possession of which party.

b) Murat Kantarcioglu and Chris cliftion [12] proposed another approach for PPDM using cryptographic techniques. This approach uses commutative encryption for

privacy preserving association rule mining on horizontally distributed data. Commutative encryption means the order of encryption does not matter. If a plaintext message is encrypted by two different keys in a different order, it will be mapped to the same cipher text. Formally, commutatively ensures that Ek1 (Ek2(x)) = Ek2 (Ek1(x)). To determine global candidate itemsets the approach is as follows:

Each party encrypts its own frequent itemsets along with enough "fake" itemsets. The encrypted itemsets are the passes to other parties until all parties have encrypted all itemsets. These are passed to a common party to eliminate duplicates and to begin decryption. This set is then passed to each party and each party decrypts each itemset. The final result is the common itemsets. Fig 5 shows an example of this approach where ABC and ABD are common item sets.

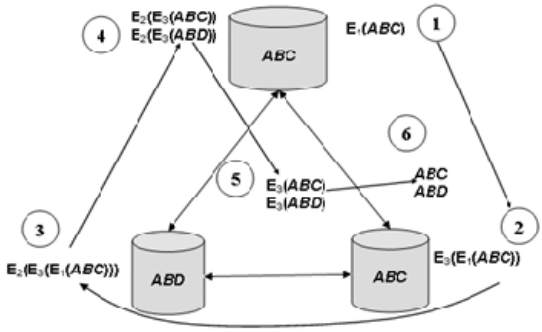

Figure 5: Determining global candidate itemsets.

3. CONCLUSION

While applying data mining methods encryption methods can be used to preserving privacy of the datasets . In this paper we have surveyed different privacy preserving data mining techniques using encryption techniques. And also we presented usage of cryptographic algorithms for privacy preserving data mining. In summary, it is possible to mine globally valid results from distributed data without compromising the privacy of the datasets.

REFERENCES

[1] C. Clifton, M. Kantarcioglu, and J. Vaidya. Defining Privacy For Data Mining. In Proc. of the National Science Foundation Workshop on Next Generation Data Mining, pages 126-133, Baltimore, MD, USA, November 2002.

[2] R. Agarwal and R. Srikanth. Privacy- Preserving Data mining. In Proc. of the 2000 ACM SIGMOD International Conference on Management of Data, pages 439-450, Dallas, Texas, May 2000.

[3] S. R. M. Oliveira and O.R. Zaiane. Achieving Privacy Preservation When Sharing Data for Clustering. In Proc. of the Workshop on Secure Data Management in a Connected World (SDM'04) in conjuction with VLDB'2004, pages 67-82, Toronto, Ontario, Canada, August 2004.

[4] S.V. Vassilios , B. Elisa, N.F. Igor, P.P. Loredana, S. Yucel and T. Yannis, 2004, "State of the Art in Privacy Preserving Data Mining" Published in SIGMOD Record, 33, 2004, pp: 50-57

[5] Wang P, "Survey on Privacy preserving data mining", International Journal of Digital Content Technology and its Applications, Vol. 4, No. 9, 2010.

[6] W. Du and M. J.Atallah. Secure Multi-party Computation Problem and their Applications: A Review and Open Problems. In Proc. of Data Warehousing and Knowledge Discovery DaWak-99, Florence, Italy, August, 1999.

[7] M. Kantarcioglu and J. Vaidya. Privacy Preserving Naïve Bayes Classifier for Horizontally Partitioned Data. In Proc. of the IEEE ICDM Workshop on Privacy Preserving Data Mining, Pages 3-9, Melbourne, FL, USA, November 2003.

[8] J. Vaidya and C. Clifton. Privacy Preserving Association Rule Mining in Vertically Partitioned Data. In Proc. of the 8th ACM SIGKDD Intl. Cong. On Knowledge Discovery and Data Mining, Pages 639-644, Edmonton, AB, Canada, July 2002.

[9] J. Vaidya and C. Clifton. Privacy Preserving K-Means Clustering Over Vertically Partitioned Data. In Proc. of the 9th ACM SIGKDD Intl. Cong. On Knowledge Discovery and Data Mining, Pages 206-215, Washington, DC, USA, August 2003.

[10] O.Goldreich, "Secure multi-party computation", (working draft). [online]. Available: http://www.wisdom.weizmann.ac.il/oded/pp.html

[11] Ashraf B. El-Sisi and Hamdy M. Mousa "Evaluation of Encryption algorithms for privacy preserving Association Rules Mining",International Journal of Network Security,vol 14, No.5,sep 2012.

[12] Murat kantarcioglu and chris Clifton," Privacy Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data", IEEE Transactions on Knowledge and Data Engineering , vol. 16 No. 9, sep 2004