Volume 3, Issue 5, May 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Review of Privacy Preserving in Data Mining Using Homomorphic Encryption

Ekta Chauhan*, Sonia Vatta
School of Computer Science and Engineering
Bahra University, India

Abstract*— This paper provides an overview of the different techniques that are used for privacy preserving in data mining. There are many data mining applications which deal with privacy sensitive data. Data mining in such privacy sensitive domains is facing growing concerns. Therefore, we need to develop data mining techniques that are sensitive to the privacy issue. We also address the issue of the privacy preserving data mining. This paper provides an overview of the RSA encryption for the privacy preserving data mining. The aim of the RSA encryption is to encrypt the data so that the customer may not lose his\her personal or valuable data. It also provides an overview of the different techniques and how they are related to each other.

Keywords*— Data Mining, Homomorphic encryption, Client-Server Architecture.

I. INTRODUCTION

Data mining applications deal with privacy preserving data. Financial transactions, health records, network communication traffic etc. are some examples. The problem of privacy-preserving data mining has become more important in recent years because of the increasing ability to store personal data of the users, and the increasing sophistication of data mining algorithms to leverage this information. A number of techniques such as randomization and k-anonymity [1, 2, 3] 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. Generally, data mining (sometimes called data or knowledge discovery) is the process of analysing data from different perspectives and summarizing it into useful information. Information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analysing data. It allows users to analyse data from many different dimensions or angles, categorize it, and summarize the relationships identified.

Homomorphic encryption is a form of encryption which allows specific types of computations to be carried out on cipher text and obtain an encrypted result which decrypted matches the result of operations performed on the plaintext. For instance, one person could add two encrypted numbers and then another person could decrypt the result, without either of them being able to find the value of the individual numbers. We use the asymmetric encryption and RSA algorithm. RSA is an Internet encryption and authentication system that uses an algorithm developed in 1977 by Ron Rivest, Adi Shamir, and Leonard Adlemen. The RSA algorithm is the most commonly used encryption and authentication algorithm and is included as part of the Web browsers from Microsoft and Netscape. It's also part of Lotus Notes, Intuit's Quicken, and many other products. The encryption system is owned by RSA Security. The company licenses the algorithm technologies and also sells development kits. The technologies are part of existing or proposed Web, Internet, and computing standards.

The mathematical details of the RSA algorithm used in obtaining the public and private keys are available at the RSA Web site. Briefly, the algorithm involves multiplying two large prime numbers (a prime number is a number divisible only by that number and 1) and through additional operations deriving a set of two numbers that constitutes the public key and another set that is the private key. Once the keys have been developed, the original prime numbers are no longer important and can be discarded. Both the public and the private keys are needed for encryption /decryption but only the owner of a private key ever needs to know it. Using the RSA system, the private key never needs to be sent across the Internet. The private key is used to decrypt text that has been encrypted with the public key. Thus, if I send you a message, I can find out your public key (but not your private key) from a central administrator and encrypt a message to you using your public key. When you receive it, you decrypt it with your private key. In addition to encrypting messages (which ensures privacy), you can authenticate yourself to me (so I know that it is really you who sent the message) by using your private key to encrypt a digital certificate When I receive it, I can use your public key to decrypt it.

II. PROPOSED SYSTEM

Proposed system will consists of various modules. Each module uses different techniques and algorithms to perform its specific tasks. When a particular module completes its task, its output will become an input for the next module. In the end the combined effort of each module will be displayed.

- Privacy preserving data mining is an ongoing research area and there are lots of issues that need to be addressed.
- In our proposed system, we have implemented privacy preservation in data mining by using the homomorphic encryption to add security so that any data mining technique does not lose its valuable data.
- Here, we have assumed that the decryption occurs entirely at the Server. For real time applications with crucial time-constraints like biomedical applications, the keys for decryption can be distributed to the user for faster decryption and retrieval of data.

III. LITERATURE REVIEW

There are several research communities whose work can contribute to privacy preserving distributed data mining. We first discuss privacy preserving work in the data mining community. Then related work from the cryptography and security communities and finally distributed data mining work.

Over the past few years, several approaches have been proposed in the context of privacy preserving data mining. Some of the main approaches include heuristic based approach, reconstruction based approach, and cryptographic approach [8]. The underlying concept of the heuristic based approach technique is: how to hide sensitive rules that can be mined from the original data while maximizing the utility of the released data. In the reconstruction based approach [4, 5], we first use some methods to distort the values of the original data and then release these distorted data. The third approach is Cryptography based approach [6, 7] which has been developed to solve the following problem: Two or more parties want to conduct a computation based on their private inputs, but neither party is willing to disclose its own output to anybody else. This problem is referred to as the Secure Multiparty Computation (SMC) problem, which requires that no more information be revealed to a participant in the computation than that participant's input and output. It is important to realize that data modification results in degradation of the database performance. In order to quantify the degradation of the data, we mainly use two metrics. The first one measures the confidential data protection, while the second measures the loss of functionality. Another important approach to Privacy Preserving Data Mining is the Access control based approach. In this approach, the groundwork to build an access control model over existing technologies was proposed called Multi-relational association rules (MRAR). This model is composed of three layers notably Authenticator, checker and the database server. In MRAR, the type of policy is Mandatory access control where the users are associated to mining levels. The addressed problem in MRAR is multilevel association rules. The major disadvantage of MRAR is that it is not always possible to assign clearances to users of commercial information systems and not always possible to assign sensitivity levels to data in case level contains another level.

This problem was overcome using the PRBAC model [1] which falls into the category of access control based approach; In Role based concept, the type of policy is Role based and the target system is Privacy preservation in data mining in the context of databases which can be built over existing database technologies. The idea of Cryptographic approach and PRBAC (Privacy Preserving Role based access control approach) has motivated us to provide a more secure approach to privacy preserving data mining by combining the benefits of these two techniques along with the idea of vertical fragmentation of the data for distributed storage. We illustrate this idea by identifying data as sensitive and non-sensitive objects and using cryptographic and vertical partitioning technique to securely store the data and taking into account the flexibility of role based access control models to access the stored data.

In 2000 Rakesh Agrawal, Ramakrishna Srikant describes the issue of privacy preserving data mining. Specifically, they consider a scenario in which two parties owning confidential databases wish to run a data mining algorithm on the union of their databases, without revealing any unnecessary information. Our work is motivated by the need to both protect privileged information and enable its use for research or other purposes.

V.S. Verykios, E. Bertino, I.N. Fovino, L.P. Provenza, Y. Saygin, and Y. Theodoridis in 2004 provide here an overview of the new and rapidly emerging research area of privacy preserving data mining. They also propose a classification hierarchy that sets the basis for analysing the work which has been performed in this context. A detailed review of the work accomplished in this area is also given, along with the coordinates of each work to the classification hierarchy. A brief evaluation is performed, and some initial conclusions are made.

W. Du and Z. Zhan, "Building Decision Tree Classifier on Private Data," Proc. IEEE Int'l Conf. Privacy, Security, and Data Mining, Dec. 2002. This paper studies how to build a decision tree classifier under the following scenario: a database is vertically partitioned into two pieces, with one piece owned by Alice and the other piece owned by Bob. Alice and Bob want to build a decision tree classifier based on such a database, but due to the privacy constraints, neither of them wants to disclose their private pieces to the other party or to any third party. We present a protocol that allows Alice and Bob to conduct such a classifier building without having to compromise their privacy. Our protocol uses an untrusted third-party server, and is built upon a useful building block, the scalar product protocol. Our solution to the scalar product protocol is more efficient than any existing solutions.

Alexandre Evfimievski, Ramakrishnan Srikant, Rakesh Agrawal, Johannes Gehrke, "Privacy Preserving Mining of Association Rules," 2002. Here we present a framework for mining association rules from transactions consisting of categorical items where the data has been randomized to preserve privacy of individual transactions. While it is feasible to recover association rules and preserve privacy using a straightforward "uniform" randomization, the discovered rules can unfortunately be exploited to find privacy breaches. We analyze the nature of privacy breaches and propose a class of

randomization operators that are much more effective than uniform randomization in limiting the breaches. We derive formulae for an unbiased support estimator and its variance, which allow us to recover item set supports from randomized datasets, and show how to incorporate these formulae into mining algorithms. Finally, we present experimental results that validate the algorithm by applying it on real datasets.

Problem and misconceptions:

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, and the increasing sophistication of data mining algorithm to leverage this information. 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.

Need of Research Work:

The need for privacy is motivated by business interests. However, there are situation where the sharing of data can lead to mutual gain. A key utility of large databases today is research, whether it is scientific or economic and market oriented. Thus, for example, the medical field has much to gain by pooling data for research; as can even competing businesses with mutual interests despite the potential gain, this is often not possible due to the confidentiality issues which arise.

IV. Resources

The original and reference implementation Java compilers, virtual machines, and class libraries developed by Sun in 1991 and first released in 1995 were used. Java is an object-oriented language similar to C++. Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka's main user interface is the Explorer, but essentially the same functionality can be accessed through the component-based Knowledge Flow interface and from the command line. Various books on Data mining, network security and cryptography are also studied. Secondary sources such as academic literature and technical literature from the internet were studied for further classification of processes and techniques.

V. Conclusion

Research area of the data mining Privacy preserving is an ongoing research area and there are lots of issues that need to be addressed. In our approach, we have implemented privacy preservation in data mining by using the homomorphic encryption to add security so that any data mining technique does not lose its valuable data. We used the asymmetric encryption with RSA encryption. Here, we have assumed that the decryption occurs entirely at the Server. For real time applications with crucial time-constraints like biomedical applications, the keys for decryption can be distributed to the user for faster decryption and retrieval of data. If client encrypts the data he/she will be able to see it normally but the server or other clients will not be able to see its clear text format. In further work we can also use elliptical cryptography and compare the different cryptography technique. In the proposed system we can refresh the data only on the server site and it has the fixed length. But in the future we can refresh the data on client site and can also increase the length size.

References

[1] Anor F.A. Dafa-Alla, Eun Hee Kim, Keun Ho Ryu, *Yong Jun Heo "PRBAC: An Extended Role Based Access Control for Privacy.

[2] Preserving Data mining" In Proceedings of the Fourth Annual ACIS International Conference on Computer and Information Science (ICIS'05) of IEEE, 2005.

[3] Proceedings of the International Multi Conference of Engineers and Computer Scientists 2008 Vol I IMECS 2008, 19-21 March, 2008, Hong Kong.

[4] Yucel Saygin, Vassilios S.Verykios and Ahmed Elmagarmid K. Privacy preserving association rule mining, In Proceedings of the 12th International Workshop on Research Issues in Data Engineering pages 151.158.

[5] Rakesh Agrawal, Srikant. Privacy Preserving Data Mining. ACM SIGMOD, 2000.

6] C. Clifton,M. Kantarcioglu, J. Vaidya, X. Lin and M. Y. Zhu. "Tools for Privacy Preserving Distributed Data Mining". In SIGKDD Explorations, 4(2): 28-34 December 2002.

[7] Murat Kantarcioglu, Chris Clifton. "Privacy preserving Distributed Mining of association Rules on Horizontally partitioned Data. IEEE transactions on knowledge and data engineering, 2003.

[8] Privacy Preserving Data Mining Using Cryptographic Role Based Access Control Approach *Lalanthika Vasudevan , S.E. Deepa Sukanya, N. Aarthi** 2008 Vol IIMECS 2008, 19-21 March, 2008, Hong Kong.

