International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 02, February-2015

An Enhanced Approach to Privacy-Preserving in Data Mining and its Techniques

Siddharth Pandey
Computer Engineering dept.
Thakur College of Engineering and Technology
Mumbai, India

Priya Singh
Computer Engineering dept.
Thakur College of Engineering and Technology
Mumbai, India

Rohan Patil
Computer Engineering dept.
Thakur College of Engineering and Technology
Mumbai, India

Harshali Patil
Assistant Professor
Computer Engineering dept.
Thakur College of Engineering and Technology
Mumbai, India

***Abstract*—The Privacy preserving Data mining (PPDM) has been among the important issues of current research that deals with preserving privacy of individual's data over a network. The major area of concern is that non-sensitive data even may deliver sensitive information, including personal information, facts or patterns. In this paper, we present a unique concept of combining different PPDM techniques which provides high level security and integrity to confidential data. This paper mainly highlights the improved results that can be obtained on merging the two different PPDM techniques. One of the latest concept of PPDM called Slicing has also been explained in our paper. It has been observed that slicing preserves better data utility and thus we have tried to merge slicing with one of the best security mechanism that is Cryptography.**

Keywords - Privacy-preserving data mining (PPDM), data mining, research challenges, privacy preserving techniques, slicing.

I. INTRODUCTION

Technology on the rise has always demanded ease of use, easy access, scalability and most importantly privacy of user data .Privacy has been a concern since the invent of internet. Every single process from booking tickets to making international economic transactions data is stored in electronic form. Electronic data is as vulnerable to risk as data in its physical form. Various algorithms past few years have been on the front to provide maximum protection to private data and to overcome the limitations of existing algorithms. Privacy preserving in data mining has been one of the fastest growing research area to improve the efficiency of existing techniques. PPDM provides different techniques and algorithms that try to protect data either by encrypting, changing or adding bogus data to make data more complex to be understood thus preventing its privacy. Anonymization, Cryptography, Perturbation based PPDM, Slicing, etc. are some of the PPDM techniques that have proved to be efficient to prevent data when it is being mined. However individually the methods have few drawbacks, thus a different approach for PPDM can be made. Combining the PPDM techniques provides more robust, stable and secure algorithm. Thus in our research we have mainly focused on combining two techniques of PPDM i.e. Slicing and Cryptography. Both techniques have proved to provide best privacy to user data, thus this motivates us to combine these techniques to generate a new algorithm that provides a simple yet an effective way of extracting data without risking the privacy of data.

The paper is organized as follows. In Section I, we give the basic concept of PPDM and its techniques. In Section II, we describe the related work done in the field of PPDM. In section III, we have mentioned the research challenges in PPDM. Section IV contains the proposed methodology that tries to overcome the flaws of PPDM. And finally we conclude in Section V.

II. RELATED WORK

Yehuda Lindell, Benny Pinkas [3], presented introduction to secure multiparty computation and its applicability to privacy-preserving data mining. The common errors that are established in the preserving data mining is implemented with secure multiparty computation techniques and the issues involved in the efficiency are discussed and also demonstrates the difficulties in constructing highly efficient protocols.

Sweety R. Lodha, S. Dhande [4], explained encryption algorithm implemented at three different levels in the paper Web Database Security Algorithms. In this paper Encryption is divided into three different levels i.e. Storage-level encryption, Database-level encryption, Application-level encryption. Storage-level encryption encrypts the data in the stored in subsystem and hence protects the static data stored. From a database point of view, storage-level encryption is transparent, thus any changes to existing applications is avoided easily. Database-level encryption provides security when data is being inserted or retrieved from database. Application-level encryption performs the encryption and decryption process at application level where the data is generated. Within the application that initiates the data into the system encryption is performed; the data is encrypted and then sent, thus naturally the data is stored and encrypted data is retrieved, which is finally decrypted again within the application.

Yuan-Hung Kao, Tung-Shou Chen and Jeanne Chen [5], proposed a novel hybrid protection scheme that protects the privacy of information and the clustering knowledge in data mining. The proposed scheme integrates the privacy – preserving data mining technique with that of knowledge-preserving anti-data mining technique. The given scheme allows user to adjust the amount of protection on personal level.

Hanumantha Rao Jalla and P N Girija [6], proposed an algorithm that addresses the problem of individual customers related to their privacy issues. Authors proposed a transformation technique. This basis of this technique has been referred from Walsh-Hadamard transformation (WHT) and one of its fundamentals i.e. Rotation in their paper. An orthogonal matrix is generated by the WHT, it transfers entire data into new domain and also maintain the distance between the data records. Techniques which are statistical based can be used to reconstruct the records, so by applying Rotation transformation this problem is resolved. Inverse matrix is one these techniques

Sativa Lohiya and Lata Ragha [7], proposed a hybrid technique in which randomization and generalization is used in their paper. In this approach the data is first randomized and then generalization is performed on the modified or randomized data. This technique protects private data and reconstructs original data with better accuracy and with no information loss.

Tiancheng Li, Ninghui Li,Ian Molloy and Jian Zhang [1], presented a new approach called slicing. Slicing is used to preserve privacy of micro data. The limitations of generalization and bucketization are overcome by this method. Utility is preserved well in Insider Threats while protecting against threats related to privacy of data. Experiments show that data utility is much better preserved by slicing than generalization and its efficiency is more than bucketization in case of workloads that involve the sensitive attributes.

III. RESEARCH CHALLENGES

Now- a-days, Data Mining is used in many applications. There are certain areas where data mining if used without privacy may cause serious affects .These areas are the main research challenges and are mentioned below.

A. Internal and External attacks, Cyber threats.

One of the major threats people face today is Cyber Crime [9]. Since most of our information is stored on electronic media and a lot of data is also available on internet or networks. Attacks on such areas might be dangerous and devastating for an individual. For example, consider the Banking system. If hackers attack a bank's information system and empty the accounts, the bank could lose millions of dollars. Therefore security of information is a critical issue. There are two types of threats –Outsider or Insider. An attack on Information System from someone outside the organization is called outsider threat, such as hackers, hacking Bank's computer systems and causing havocs. A more critical problem is the insider threat. Insider threat can be due to an intruder present in the organization. Members of an organization have studied their policies and business practices and know every bit of the information so it can affect the organization's information assets.

B. Fraud in Credit Cards and Individual's Identity Theft

Another area which requires attention is detecting frauds and thefts. Frauds may be credit card frauds. These can be detected by identifying purchases made of enormous amounts [9]. A similar and a more serious theft is identity theft. Here one pretends to be an identity of another person by obtaining that person's personal information and carrying out all types of transactions under the other person's name. By the time, the owner finds out it is often far too late-the victims may already have lost millions of dollars due to identity theft.

C. Flaws in individual techniques

PPDM has a huge list of techniques with different approach and concept. However every individual technique in its own has some flaws which increases the challenge for designing a better algorithm, the individual flaws are stated below.

- Anonymization: Since Anonymization generates transformed data, its accuracy of applications on the data is reduced [8]. Available or unavailable attributes in external table are difficult to determine in *k*-anonymity model.
- Cryptographic Technique: For huge databases this algorithm doesn't proves to be a strong technique as this technique fails to protect the output of computation [8]. Thus as a result mining the result may break the privacy of individual's record [9].
- Data Perturbation: Preserving the original data becomes difficult in some perturbation approaches. Data mining technique is to be selected based on the method using which noise has been introduced in data [9].
- Randomization: Each records are treated individually irrespective of their local density [8].
- Generalization: A considerable amount of information is lost for high dimensional data in generalization [1].
- Bucketization: Membership disclosure is not prevented in this method and clear separation between sensitive attributes and quasi-identifying attributes is a must for this method else the method is inapplicable [1].

IV. PROPOSED METHODOLOGY

Our concept of merging different techniques aims on combining cryptographic technique and slicing. Cryptography has different approaches to provide privacy. Authentication, Encryption, key exchange, etc are some of the basic techniques which when modified provide a high level of security thus making it nearly difficult to break into an individual's privacy. Cryptography has been one of the

most used privacy prevention technique in multiparty data computation. This method prevents leakage of computations. Slicing was one of the techniques introduced that overcame the drawbacks of generalization and bucketization. Membership disclosure and preserving better data utility are the advantages of slicing. Slicing as the name says partitions the data set or attributes vertically as well horizontally.

Since cryptography aims at protecting leakage of private computation result and slicing aims at preserving better data utility each method holds some drawback. Thus our concept include different level authentication and database level slicing. Combining these two approaches ensures user level privacy and database level privacy. A robust algorithm is thus introduced in this paper.

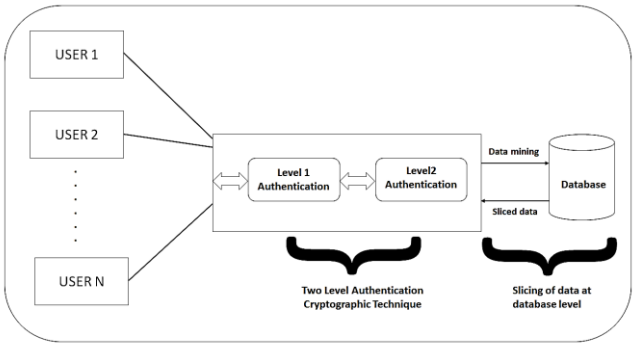

Fig. 1. Basic framework of the proposed approach

Figure 1 depicts the entire framework of the proposed approach. The process involves multiple parties trying to gain access to some private data. Thus as a measure of validation the users are authenticated using a two level authentication process.
The two level authentication process involves exchange of private unique ids. Every time a user trying to access data has to go through the two level authentication process.

The two level authentication ensures that private data is shared only with the party that has requested that data. Cryptography is used at first two stages for authentication since cryptography is one of the best known technique for multiparty data computation. The system is never confined to a single user thus the role of cryptography at initial stage is to make a secure gateway for the users to make clear demand for data and get the precise data requested.

Data entered in database may have unknown dimensionality. Thus in order to handle all dimensions of data especially high –dimensional data slicing acts as a major support factor. Since slicing slices the dataset horizontally and vertically, it aims at breaking association across the column but at same time preserving the association within each column. Slicing ensures database level security as sliced data may have least association with other records thus reducing the risk of leaking additional private data which is not requested. As shown in figure 1, a request from user is processed and thus the data highly correlated with the data requested are grouped together using the slicing algorithm.

Thus cryptography ensures user level privacy whereas slicing ensures database level privacy. Cryptography and Slicing form a robust hybrid technique for privacy preserving in data mining.

V. CONCLUSION

This paper has introduced a robust, stable and effective method for preserving data privacy at different platform. Since data online are the most vulnerable one this hybrid technique can be used over internet. Implementation of the algorithm guarantees security to a higher extent. However further research may make this technique much more unpredictable and difficult to break. The two level authentication proves to be an impact factor as a fresh approach of key exchange and authentication are used at the same time. Further research may reduce the overhead on the two level authentication algorithm. Slicing at the basic level supports cryptography in the given approach thus plays an important role at database level. Hybrid techniques have always proved to be a better approach for PPDM, thus overcoming different flaws and providing a better mean for preserving data privacy.

REFERENCES

[1] Tiancheng Li, Ninghui Li, Jian Zhang, and Ian Molloy, “Slicing: A New Approach for Privacy Preserving Data Publishing”, in proceedings of IEEE Transactions on Knowledge and Data Engineering, Vol. 24, No. 3,pp. 561-574, Mar. 2012.

[2] Anand Sharma and Vibha Ojha, “Implementation of Cryptography for Privacy Preserving Data Mining”, in International Journal of Database Management Systems (IJDMS), Vol.2, No.3, Aug. 2010.

[3] Yehuda Lindell and Benny Pinkas, “Secure Multiparty Computation for Privacy-Preserving Data Mining” ,The Journal of Privacy and Confidentiality, Number 1, pp. 59-98,2009.

[4] Sweety R. Lodha and S. Dhande, “Web Database Security Algorithms”, in International Journal of Advance Research in Computer Science and Management Studies (ijarcsms), Volume 2, Issue 3, pp. 293-299, Mar 2014.

[5] Tung-Shou Chen, Jeanne Chen and Yuan-Hung Kao, ”A Novel Hybrid Protection Technique of Privacy Preserving Data Mining and Anti-Data Mining”, in Information Technology Journal, Volume 9, Issue 3,pp. 500-505,2010.

[6] Hanumantha Rao Jalla and P N Girija, “An Efficient Algorithm For Privacy Preserving Data Mining Using Hybrid Transformation”, in International Journal of Data Mining & Knowledge Management Process (IJDKP) Volume 4, Number 4, July 2014.

[7] Savita Lohiya and Lata Ragha, “Performance Analysis of Hybrid Approach for Privacy Preserving in Data Mining”, in proceedings of Int. J. on Recent Trends in Engineering and Technology, Volume 8, Number. 1, Jan. 2013.

[8] Dharmendra Thakur, Prof. Hitesh Gupta, “An Exemplary Study of Privacy Preserving Association Rule Mining Techniques”, in proceedings of International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE) , Volume 3, Issue 11, pp. 893-900, Nov. 2013.

[9] Shweta Taneja, Shashank Khanna, Sugandha Tilwalia, Ankita, “A Review on Privacy Preserving Data Mining: Techniques and Research Challenges”, in proceedings of International Journal of Computer Science and Information Technologies (IJCSIT), Volume 5,Issue 2 ,pp. 2310-2315, 2014.