A Survey of Cryptographic and Non-cryptographic Techniques for Privacy Preservation

Bhawani Singh Rathore
Student, M. Tech, 6th Semester
Branch CSE
UIT
Barkatullah University

Anju Singh
Assistant Professor
Department of CS&IT
UTD
Barkatullah University

Divakar Singh
HOD,
Department of CSE
UIT
Barkatullah University

ABSTRACT

Cryptography is to become familiar with the requirement of large, complex, information rich data sets for it’s privacy preservation. The privacy preserving data mining has been generated; to go through the concept of privacy in data mining is hard. Several algorithms and approaches are being generated theoretically, but practically it is hard. Privacy in data mining can be achieved through several techniques such as Perturbation, Anonymization and Cryptographic. Here in this paper tries to reiterate several Privacy Preserving Data Mining (PPDM) techniques presently developed to cope up with the privacy issues in the process of data mining. In this paper there is a survey of various cryptographic & non-cryptographic techniques too.

Keywords

Privacy, Privacy Preservation, Cryptography, Anonymization, Data Mining, Randomization.

1. INTRODUCTION

Data mining techniques are being implemented in several researches and applications. But, Data mining techniques raises the problem of privacy too. Privacy refers as crucial factor in the information system. Because of, various works are being devoted to incorporate privacy preserving methods with data mining algorithms with the view to stop sensitive data at the time of knowledge discovery process. When a person transfers their database to their server, few sensitive patterns are concealing from its database regarding particular privacy policies. Recently, researchers focus on seriousness of the errors regarding privacy.

Privacy preserving refers as the crucial issue in the data mining techniques. The major challenge of present data mining algorithm is drawing the data at the time of maintenance of privacy datasets. Because of this concern, the privacy preserving data mining (PPDM) techniques are being adopted. The major concern in the privacy preserving data mining is referring as sensitive pattern mining. Privacy Preserving Data mining techniques implemented for modification of database with the help of insertion wrong information for concealing sensitive information. These problems at present are improved in data mining techniques but they also become expensive, computation and overhead will take place [1, 2]. The Privacy Preserving Data Mining shows privacy preserving data publishing model and multi analyze few associate technologies, k-anonymity, relational k-anonymity, 1-diversity and perturbation of privacy preserving. The privacy preserving techniques also have data sanitization, cryptography techniques and access control techniques. The data sanitization techniques refer as a process, which conceals sensitive information in their data sets. Here in this paper, we are disussing various methods to offer the privacy to their dataset. The access control methods are implemented to reduce and handle the access to the host system and their applications through communication links. The cryptography techniques are implemented to offer a tough security and privacy and also offer exact and practical implementation. The cryptography protocols permits secure communications through addressing their authentication [1].

There are various advantages and disadvantages of the cryptography of using techniques with privacy preservation data mining.

Advantages:

- Robust
- Sender and recipient authentication
- Anonymity
- Fairness
- Accountability
- Integrity in Storage

Disadvantages:

- Takes a long time to figure out the code.
- It takes long to create the code.
- If you were to send a code to another person in the past, it will take long to get to that person.
- Overall Cryptography It's a Long Process.

2. PRIVACY PRESERVATION

Privacy is an important concern while disclosing various categories of electronic data including business data and medical data for data mining. Privacy can be interpreted in two ways. For instance, privacy is so crucial with respect to medical data, since it contains sensitive information like type of disease. Especially for doing medical data mining the original data should be available for making accurate predictions otherwise lead to impractical solutions [3]. Any kind of disclosure related to the person- specific information leads to many problems including ethical issues. Therefore extra care should be taken to protect privacy of individuals before publishing such data. On the other hand, the privacy can be interpreted as preventing unwanted disclosure of information while performing data mining on aggregate results. Thus, privacy can be addressed at various levels in the process of data mining. For entire database security both privacy and security measures are needed.

2.1 Application of Privacy Preservation Data Mining

Privacy getting more attention to control terrorism, the government wanted examine, implementing data mining technology, high information of individuals to search unique disease outbreaks, financial fraudulent attitudes, network intrusions, etc. All these applications of data mining are beneficial towards society, on the other hand, their negative side of this technology due the reason that it can threaten the individuals' privacy. Overcome the "limitations" of the data mining methods also have the sectors like data security and privacy preserving data mining, these are active and growing research areas.

2.2 Purpose of PPDM

The PPDM algorithms mainly intended towards rule hiding and data hiding. Where in Data Hiding the sensitive data from its original database such as indentify name and address are linked, directly or indirectly towards individual person are concealed. Where in the rule hiding sensitive data (rule) from its original database after the application of data mining algorithm is eradicated. Mostly PPDM algorithms conceal sensitive patterns through modification data hiding.

2.3 PPDM Techniques

PPDM techniques implemented through various categories Sanitation, there is a removal or modification of items for its database to decrease the support of few rapid used items sets and its sensitive patterns are not supposed to be mined. Blocking them can exchange specific attributes of data having a question mark [4]. Regarding this minimum support and its confidence going to be altered into lessen interval. Here in distort the support and the confidence of the sensitive rule depends on the confidentiality factor of data which is expected to be secured and called as data perturb action or data randomization too, in this case individual data records are changed from actual data, and rebuilt from its randomized data. This motive of this method to design for its distortion methods after that the actual value of individual record is hard to ascertain, but not vary for its danger data. In the generalization transforms and exchange each record value with its correspond generalized value.

2.4 Classification of Privacy Preserving Techniques

Mainly, these techniques are divided into two parts:

1. Cryptography
2. Non-Cryptography

2.4.1 Cryptographic Technique

Generally, these methods decrease the granularity of arranging so that it decreases the privacy. This results in some informational loss. It is a natural trade-off in between privacy and information loss. Various methods create of transformation on their actual data for offering the privacy reservation. The transformed dataset should be capable of mining and privacy needs in absence of losing their profits of mining.

Cryptographic techniques are ideally meant for such scenarios where multiple parties collaborate to compute results or share non sensitive mining results and thereby avoiding disclosure of sensitive information.

Cryptographic techniques find its utility in such scenarios because of two reasons:

1. It offers a well defined model for privacy that includes methods for proving and quantifying it.

2. Vast set of cryptographic algorithms and constructs to implement privacy preserving data mining algorithms are available in this domain. The data may be distributed among different collaborators vertically or horizontally [6].

Some Cryptography Methods:
a.Secure Multiparty Computation

All these methods are almost based on a unique encryption protocol called as Secure Multiparty Computation (SMC) technology. SMC used in distributed privacy preserving data mining made up of a set of protected sub protocols that are used in horizontally and vertically partitioned data: secure sum, secure set union, secure size of intersection and scalar product [19].

Advantages:

- Safe
- Secure
- Trust-worthy

Disadvantages:

- Communication complexity grows exponentially with n.

b. Public-key cryptosystems (asymmetric ciphers)

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 (discussed below), and is widely used for secured websites that carry sensitive data such as username and passwords, and credit card numbers.

Advantages:

- In asymmetric or public key, cryptography there is no need for exchanging keys, thus eliminating the key distribution problem.
- The primary advantage of public-key cryptography is increased security: the private keys do not ever need to be transmitted or revealed to anyone.
- Can provide digital signatures that can be repudiated

Disadvantages:

- A disadvantage of using public-key cryptography for encryption is speed: there are popular secret-key encryption methods which are significantly faster than any currently available public-key encryption method.

2.4.2 Non-Cryptography Techniques

2.4.2.1 Randomization

The randomization technique implements data distortion techniques for adding little noise in the actual data. Here its recovery of individual values from their records is more longer through adding noise to its actual data., but here only aggregate distribution are being recovered. The randomization method offer better balance in between privacy preservation and their knowledge discovery. There are sorts of perturbation are being possible through randomization method. This method also have random noise depended on perturbation and Randomized response scheme. Therefore it results effective and highly information method [5].

Advantages:

- Intended benefit, not the benefit actually received
- simple and easy to implement

Disadvantages:

- Very sensitive for sample data

2.4.2.2 The Anonymization method

Anonymization method [20] implements generalization and suppression methods to generate individual record in-distinguislable within group records. The techniques remains in k anonymity model is like various attributer in their data can after be taken quasi – identifiers. It built conjunction with its public records so that definitely identify the records. The anyonymization method is certain about transformed data is appropriate, but its information losses is also there, extent.

Advantages:

- High efficiency
- Leads to high-quality data.
- More flexible

Disadvantages:

- Failing to work in the Top-Down Specialization (TDS) approach.
- It does not address the complementary problem of reasoning about disclosure across multiple releases

3. LITERATURE REVIEW

3.1 Cryptography Based

In paper [6], Research in protected distributed computation, which takes place like a large body of research in the concept of cryptography, obtained great results. It provided non-trusting parties can connects to compute functions having distinct inputs at the time ensuring that no party seeks anything but provides output as a function. These results displayed through generic constructions which is applicable to any function and offer an efficient presentation like a circuit. Authors explain their results, focus on their efficiency, and analysis their relevance to privacy preserving calculation of data mining algorithms. Here they displayed some examples of protective computation of the data mining algorithms which implements these generic constructions.

This paper [8] was desired to discuss general ideas from a huge body of the cryptographic research on protective distributed computation, and its applications to the data mining. Authors explain in details definitions of security, and generic constructions for any two-party and multi-party scenarios. Authors displayed it is simple to design and usage depends on the constructions for their two-party state after that design one depends on the multi-party constructions. The major parameter which influences the feasibility of using a secure protocol depends on the generic constructions refer as the efficient combinatorial circuit which calculates the function which is to be discussed. The main computational bottleneck of the constructions is the oblivious transfer protocol, and any improvement in the overhead of this protocol should directly affect the overhead of secure computation.

The theoretical model was first proposed by Damgard and Ishai [7] where dedicated miner nodes are used to collect and process inputs form data donors.

S ˚ HAREMIND provides a right mix of cryptographic techniques and implementation techniques to get maximal efficiency. Indeed, the alternative multi-party computation frameworks SIMAP [9] are less optimized for large input sizes. The SEPIA framework [10] is comparable in the speed of multiplication, but it is slower in comparisons. Please note that both the running time and limiting cost were not available for all implementations. Also, benchmarks have been conducted in different settings and thus only magnitude differences are relevant. Frameworks for secure two-party computations: FAIRPLAY [11] and TASTY [12] are significantly slower, since they rely on public key cryptography. Public key operations are several orders of magnitude slower than share manipulation operations. SHAREMIND has some unique features that are not found in other secure computation implementations. First, it has a database for securely storing large datasets prior to aggregation. Second, the high-level SECREC algorithm language hides the details of cryptographic protocols. Competing languages such as SMCL represent cryptographic Framework Multiplication Less than or protocols instead of algorithms [13]. Third, SHAREMIND has strong support for vector and matrix operations which are executed as efficient parallel operations. All these features greatly simplify the development of data mining algorithms.

Paper [14] describes about the issue of privacy in cloud environment. This paper discusses about various privacy preserving techniques for the data stored in cloud environment. At the same time paper [15] gives briefs about the various privacy preservation techniques. This paper review progressive strategies for privacy. This paper gives main focus on organization, k-anonymization and distributed privacy-preserving data processing. These methods are mainly belongs to Non-Cryptographic.

The complete analysis can be shown as follows:

S. No.	**Method/ Paper**	Data Partition	Computation	Method Type
1	[6]	No Partition	Distributed	Cryptography
2	[7]	No Partition	-	Cryptography
3	[8]	No Partition	Distributed	Cryptography
4	[9]	Partitioned	Multi-Party Computation	Non-Cryptographic
5	[10]	No Partition	-	Non-Cryptographic
6	[11]	No Partition	Multi-Party Computation	Cryptography
7	[12]	No Partition	Multi-Party Computation	Cryptography
8	[13]	-	-	Cryptography
9	[14]	MIX	-	Non-Cryptographic.
10	[15]	MIX	-	Non-Cryptographic.

3.2 Non-Cryptography Based

There is some healthy literature on secrecy in statistical databases. An excellent Survey of work prior to the late 1980's was done by Adam and Wortmann. With the help of taxonomy, this work [16] goes down in category of output perturbation. Somehow, regarding our knowledge, the work which has vanish the opportunities for privacy naturally due to the reason that the exploit databases the original number of crises will be sub linear is Sect. 4 of [16] (joint work with D work). That approach is referred as single-attribute SuLQ databases.

Fanconi and Merola proposed an survey recently, with a concentrate on aggregated data released through web access [17]. Evfimievski, Gehrke, and Srikant, in the Introduction to [18], give a very better discussion of work in randomization of data, in Which data contributors (e.g., respondents to a survey) separately add noise with their personal responses. A special issue (Vol.14, No. 4, 1998) of the Journal of Official Statistics is dedicated to disclosure control in statistical data. A discussion of some of the trends in the statistical research, accessible to the non-statistician, can be found.

4. CONCLUSIONS & FUTURE WORK

Privacy preserving possesses high energy to reach and benefits of data mining technology. The major motive of this paper is to discuss several PPDM methods which are meant for controlling privacy issues in data mining. To offer accurate results in data mining, various PPDM methods are taken place. Since, no such techniques took place which works provide absolute privacy to data. Research in this direction can make important contributions. This paper discusses largely about cryptographic and non-cryptographic approaches for protection of distributed and centralized data, and their various applications to data mining.

This study motivates to work for the maintaining of privacy of the data through the cryptographic technique, called secure sum. Secure sum technique is specially meant for the horizontally partitions dataset.

5. REFERENCES

[1] Murat Kantarcioglu, Chris Clifton, "Privacy Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data", IEEE Transactions on Knowledge and Data Engineering, Vol. 16, No. 9, pp. 1026 – 1037, 2004.

[2] Shariq J. Rizvi, Jayant R. Haritsa, Maintaining Data Privacy in Association Rule Mining.

[3] SHEN Yangguang, SHAO Hui, YANG Li. Privacy Preserving c4.5 algorithm over vertically distributed datasets[[C]II Proc of the 2009 Int'! Conf. on Networks Security,2009(2): 446-448.

[4] P J MODI, P W T KIM. Classification of Examples by Multiple Agents with Private Features[C]II Proc of 2005 IAT, Washington DC, 2005:223-229.

[5] S Mukherjeea, M Banerjeea, Chen Zhiyuan, A privacy preserving technique for distance-based classification with worst case privacy guarantees[J]. Data & Knowledge Engineering,2008,66(2):264-288

[6] J VAIDYA,M Kantarcioglou,C CLIFTON. Privacy preserving naïve Bayes classification[J]. The VLDB Journal, 2007,17(4):879-898.

[7] Ivan Damgard and Yuval Ishai. Constant-round multiparty computation using a black- ˚ box pseudorandom generator. In Proc. of CRYPTO '05, volume 3621 of LNCS, pages 378–394. Springer, 2005.

[8] Martin Geisler. Cryptographic Protocols: Theory and Implementation. PhD thesis, Aarhus University, 2010.

[9] Peter Bogetoft, Ivan Damgard, Thomas Jakobsen, Kurt Nielsen, Jakob Pagter, and ˚ Tomas Toft. A practical implementation of secure auctions based on multiparty integer computation. In Proc. of Financial Cryptography '06, volume 4107 of LNCS, pages 142–147. Springer, 2006.

[10] Martin Burkhart, Mario Strasser, Dilip Many, and Xenofontas Dimitropoulos. SEPIA: privacy-preserving aggregation of multi-domain network events and statistics. In Proc. of USENIX conference on Security, pages 15–15. USENIX Association, 2010.

[11] Dahlia Malkhi, Noam Nisan, Benny Pinkas, and Yaron Sella. Fairplay - secure twoparty computation system. In Proc. of USENIX Security Symposium, pages 287–302, 2004.

[12] Wilko Henecka, Stefan Kogl, Ahmad-Reza Sadeghi, Thomas Schneider, and Immo ¨ Wehrenberg. TASTY: tool for automating secure two-party computations. In Proc. of CCS '10, pages 451–462. ACM, 2010.

[13] Janus Dam Nielsen. Languages for Secure Multiparty Computation and Towards Strongly Typed Macros. PhD thesis, Aarhus University, 2009.

[14] Amit Kumar Jha and Divakar Singh ,"A Survey of Cloud Computing Service and Privacy Issues," Advances in Computer Science and Information Technology (ACSIT), Volume 1, Number 2; November, 2014 pp. 4-8.

[15] Ashish Chouhan and Dr.Anju Singh ,"Privacy Preserving Data Mining: A Survey on Anonymity," International Journal of Electrical, Electronics and Computer Engineering 4(1): 82-92 (2015).

[16] A. V. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke. Privacy preserving mining of association rules. In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 217–228, 2002

[17]. L. Franconi and G. Merola, Implementing Statistical Disclosure Control for Aggregated Data Released Via Remote Access, Working Paper No. 30, United Nations Statistical Commission and European Commission, joint ECE/EUROSTAT work session on statistical data confidentiality, April, 2003, available at http://www.unece.org/stats/documents/2003/04/confidentiality/wp.30.e.pdf

[18]. A. V. Evfimievski, J. Gehrke and R. Srikant, Limiting privacy breaches in privacy preserving data mining, Proceedings of the Twenty-Second ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, pp. 211-222, 2003.

[19] X.-Y. Chang, D.-L. Deng, X.-X. Yuan, P.-Y. Hou, Y.-Y. Huang and L.-M. Duan," Experimental realization of secure multi-party computation in an entanglement access network," arxiv, 2015.

[20] Rohit Kutty, Srikanth Ravee, Sanchi Ubale and Amruta Sankhe," An Innovating Approach To Data Security," International Journal of Science, Technology & Management, Volume No.04, Issue No. 02, February 2015.