Volume 3, Issue 8, August 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

An Emblematic Study of Different Techniques in PPDM

Ankita Shrivastava, U.Dutta
C.S Dept.,M.P.C.T,GWALIOR
India

Abstract- In topical ages, privacy-preserving data mining (PPDM) has been studied extensively, because of the extensive explosion of sensitive information on the internet. A number of algorithmic techniques have been analyzed for privacy-preserving data mining. This paper explores a metaphorical study of the different methods for privacy preserving data mining are Randomization, K-anonymization, Association rules, Cryptographic technique for information sharing and privacy. Data and knowledge hiding are two research advices that examine how the privacy of uncooked data, or information, can be maintained either before or after the course of mining the data. After survey of recent approaches that have been applied to knowledge hiding thread. Data mining services require precise input data for their results to be meaningful, but privacy concerns may require users to provide unauthentic information. This paper is based on the computational and theoretical limits associated with privacy-preservation.

Keywords- Privacy-preserving data mining (PPDM), Randomization, K-anonymization Distributed privacy-preserving data mining, Statistical Disclosure Control, Association rules.

I. Introduction

In modern years, data mining has been observed as a risk to privacy because of the extensive proliferation of electronic data maintained by organizations. This has led to increased worries about the privacy of the essential data. In recent years, a number of techniques have been proposed for modifying or transforming the data in such a way so as to preserve privacy. A survey on some of the techniques used for privacy-preserving data mining may be found. In this chapter, we will study an overview of metaphorical study of different techniques in privacy-preserving data mining. Data mining is used for retrieving intelligent information from huge databases. Presently these databases are distributed across the world. Distributed data must be retrieved from multiple locations in to the data warehouse, so there is a requirement for a secure transmission and maintaining confidentiality. The transmitted data may contain information which may be private to individual or corporate information which must be secured as shown in[3]. Also it contain Data perturbation technique which has different idea, that the distorted data does not reveal private information, and thus it is "safe" to use for mining. The key result is that the distorted data, and information on the distribution of the random data used to distort the data, can be used to generate an approximation to the original data distribution, without revealing the original data values as shown in[10].

II. Why We Need PPDM

Applications in commercial domains possess large datasets on individuals. This data includes private and sensitive information e.g. Patient diseases, bank account details, organization structural details etc. When data mining techniques are applied on these applications the private and sensitive information of the subjects will be revealed. However, it is necessary to share the information in such a way that the identities of the individuals are not revealed as shown in [7]. So it is necessary to anonymize the data. So it encompasses different techniques to anonymize the data. Also it has taxonomy of PPDM algorithms.

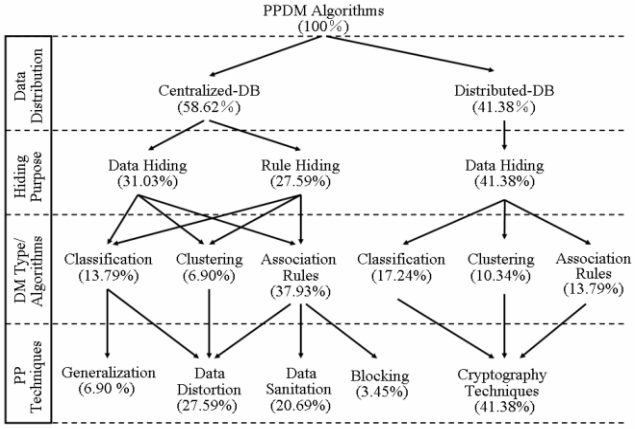

Fig. 1 Taxonomy of PPDM algorithms

III. Privacy Preserving Data Mining (PPDM)

Privacy Preserving Data Mining (PPDM) is one of the utmost fresh concepts of data mining research challenges. It refers to the area of data mining those efforts to protect sensitive information from disclosure. The problem with data mining output is that it also reveals some information, which is considered to be private and personal. Easy access to such personal data poses a threat to individual privacy. The actual anxiety of people is that their private information should not be misused behind the scenes without their knowledge. The real threat is that once information is unrestricted, it will be impractical to stop misuse as shown in [9]. There has been growing concern about the chance of misusing personal information behind the scene without the knowledge of actual data owner. Privacy preserving data mining technique gives new direction to solve this problem. PPDM gives valid data mining results without learning the underlying data values .The benefits of data mining can be enjoyed, without compromising the privacy of concerned individuals. The original data is modified or a process is used in such a way that private data and private knowledge remain private even after the mining process. The main purpose of privacy preserving data mining is to design efficient frameworks and algorithms that can extract relevant knowledge from a large amount of data without revealing of any sensitive information.

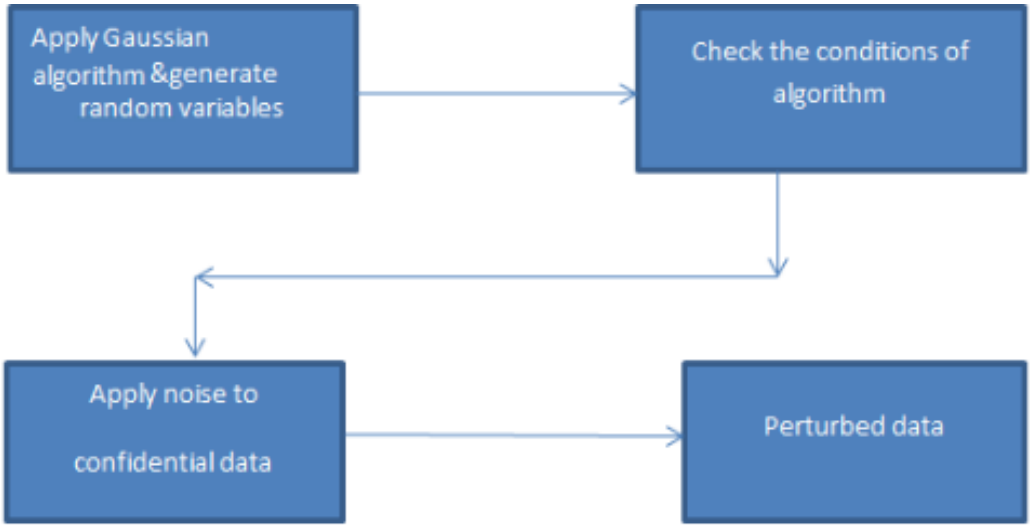

Fig. 2 Block diagram for implementing perturbation technique

The premise of this concept is that the data owner is unable to extract knowledge from a large repository of data including sensitive items. In the case of a trusted party as a data miner may be allowed to access the data in its original form. Yet, with increased data security threats across the communication channels, the private/sensitive data could require securing transformation from disclosure. Otherwise, the untrusted data miner could be provided with the encoded data for the protection of privacy or crucially sensitive information.

IV. Study of Different Techniques in Privacy-Preserving Data Mining

***A. RANDOMIZATION METHOD*:-** The randomization method is a technique for privacy-preserving data mining in which noise is added to the data in order to mask the attribute values of records. The noise added is sufficiently large so that individual record values cannot be recovered. Therefore, techniques are designed to derive aggregate distributions from the perturbed records. Subsequently, data mining techniques can be developed in order to work with these aggregate distributions.

1) Additive Perturbation: In this case, randomized noise is added to the data records. The overall data distributions can be recovered from the randomized records. Data mining and management algorithms re-designed to work with these data distributions.

2) Multiplicative Perturbation: In this case, the random projection or random rotation techniques are used in order to perturb the records. Privacy preserving data mining is used for secure mining from the data warehouse. Random perturbation technique is a method to convert raw data based on probability which has been discussed. Data distortion is achieved by changing the original data, in which some randomness value is added such that the original data is difficult to ascertain, while preserving global feature of a record as shown in [3]. In Fixed-data perturbation method the data is changed by adding a noise term e to the attribute X resulting in Y, $Y=X+e$, where e is drawn from some probability distribution. This method is called Additive Data Perturbation (ADP). In Multiplicative Data Perturbation (MDP) the value of e is multiplied with X to get Y the perturbed value, $Y=Xe$, where e has mean of 1.0 and a specified variance as shown in[3].

In randomization perturbation approach the privacy of the data can be protected by perturbing sensitive data with randomization algorithms before releasing it to the data miner. The perturbed data version is then used to mine patterns and models. The algorithm is so chosen that combined properties of the data can be recovered with adequate accuracy while individual entries are considerably distorted. In this method privacy of confidential data can be obtained by adding small noise component which is obtained from the probability distribution .In a set of data records denoted by

$X = \{x_1 \ldots x_N\}$. For record x I ЄX, we add a noise component which is drawn from the probability distribution. Commonly used distributions are the uniform distribution over an interval $[-\alpha, \alpha]$ and Gaussian distribution with mean $\mu = 0$ and standard deviation σ. These noise components are drawn independently, and are denoted $y_1 \ldots y_N$. Thus, the new set of distorted records is denoted by $x_1 + y_1 \ldots x_N + y_N$. It is denoted by this new set of records $z_1 \ldots z_N$. In general, it is assumed that the variance of the added noise is large enough, so that the original record values cannot be easily guessed from the distorted data. One key advantage of the randomization method is that it is relatively simple, and does not require knowledge of the distribution of other records in the data.

B. K-ANONYMITY:- The k-anonymity model was developed because of the possibility of indirect identification of records from public databases. This is because combinations of record attributes can be used to exactly identify individual records. In the k-anonymity method, it reduces the granularity of data representation with the use of techniques such as generalization and suppression. This granularity is reduced sufficiently that any given record maps onto at least k other records in the data. An important method for privacy de-identification is the method of k-anonymity. The motivating factor behind the k anonymity technique is that many attributes in the data can often be considered pseudo-identifiers which can be used in conjunction with public records in order to uniquely identify the records. For example, if the identifications from the records are removed, attributes such as the birth date and zip-code can be used in order to uniquely identify the identities of the underlying records. The idea in k-anonymity is to reduce the granularity of representation of the data in such a way that a given record cannot be distinguished from at least $(k - 1)$ other records as shown in Table 1.

TABLE I
K-ANONYMOUS DATA

Age	Weight	Name
35	50	Ankita
60	55	Shweta
65	50	Nikhil

(a) Original Data

Age	Weight	Name
[35,45]	[50,65]	Ankita
[35,45]	[50,65]	Shweta
[55,65]	[50,65]	Nikhil

(b) K-Anonymous Data

C. CRYPTOGRAPHIC TECHNIQUES:- In many cases, multiple parties may wish to share aggregate private data, without leaking any sensitive information at their end. For example, different superstores with sensitive sales data may wish to coordinate among themselves in knowing aggregate trends without leaking the trends of their individual stores. This requires secure and cryptographic protocols for sharing the information across the different parties.

Cryptography, the science of communication and computing in the presence of a malicious adversary extends from the traditional tasks of encryption and authentication .In an ideal situation, in addition to the original parties there is also a third party called "trusted party". All parties send their inputs to the trusted party, who then computes the function and sends the appropriate results to the other parties. The protocol that is run in order to compute the function does not leak any unnecessary information. Sometimes there are limited leaks of information that are not dangerous. This process requires high level of trust.

Fig. 3 System using Semi trusted third party

D. CRYPTOGRAPHIC APPROACH:- In this novel framework the total process is divided into three components the customer, mediator and a group of service data providers. Initially there is no communication between customer and data provider. When the client sends a query, the mediator send the information to all the data holders and through exchange of acknowledgements, the mediator establishes the connection with data providers. Also, we have a cryptographic approach i.e. Elliptic Curve Cryptography (ECC).

E. ECC APPROACH:- Elliptic Curve Cryptography (ECC) is an attractive alternative to conventional public key cryptography, such as RSA. ECC is an ideal candidate for implementation on constrained devices where the major computational resources i.e. speed, memory is limited and low-power wireless communication protocols are employed. That is because it attains the same security levels with traditional cryptosystems using smaller parameter sizes as discussed in [2].Moreover, in several application areas such as person identification and e-Voting, it is frequently required of entities to prove knowledge of some fact without revealing this knowledge. Such proofs of knowledge are called Zero Knowledge Interactive Proofs (ZKIP) and involve interactions between two communicating parties, the Prover and the Verifier. In a ZKIP, the Prover demonstrates the possession of some information (e.g. authentication information) to the Verifier without disclosing it. In this review paper, we focus on the application of ZKIP protocols on resource constrained devices. In this paper we study well-established ZKIP protocols based on the discrete logarithm problem and transform them under the ECC setting.

V. An Overview of Zero Knowledge Protocols

Generally, a zero-knowledge protocol allows a proof of the truth of an assertion, while conveying no information whatsoever about the assertion itself other than its actual truth. Usually, such a protocol involves two entities, a prover and a verifier. A zero-knowledge proof allows the prover to demonstrate knowledge of a secret while revealing no information whatsoever of use to the verifier in conveying this demonstration of knowledge to others.

The zero-knowledge protocols to be discussed are instances of interactive proof systems and non-interactive proof systems. In the first category, a prover and a verifier exchange multiple messages (challenges and responses), typically dependent on random numbers which they may keep secret whereas in the second the prover sends only one message as shown in[2]. In both systems the prover's objective is to convince the verifier about the truth of an assertion, e.g. the claimed knowledge of a secret. The verifier either accepts or rejects the proof. A zero-knowledge proof must obey the properties of completeness and soundness. A proof is complete, if given an honest prover and an honest verifier, the protocol succeeds with vast probability and sound if the probability of a dishonest prover to complete the proof successfully is negligible. A typical example of zero-knowledge proof is known as Alibaba's cave problem. In this story, Annie has uncovered the secret word used to open a magic door in a cave. The cave is shaped like a circle, with the entrance on one side and the magic door blocking the opposite side, as shown in Figure 4. The left path from the entrance is labeled A and the right B. John states that he will pay her for the secret, but not until he's assured that she really knows it. Annie claims that she will tell him the secret, but not until she receives the money.

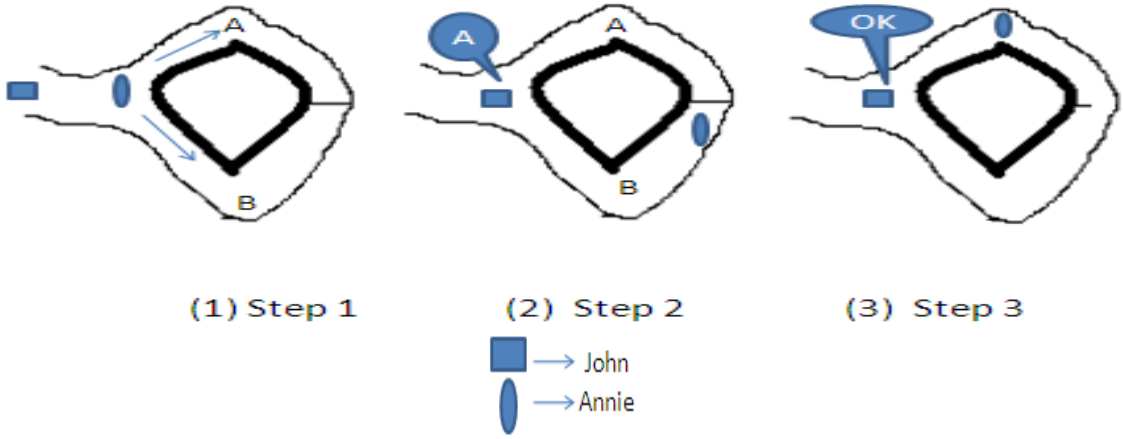

Fig. 4 Alibaba's Cave Problem.

John waits outside the cave as Annie goes in. Annie randomly takes either path A or B inside the cave John enters the cave and shouts the name of the path he wants her to use to return either A or B, chosen at random Annie does that using the secret word if needed to open the magic door. The above steps are repeated n times until John is convinced that Annie knows the secret word. Now, suppose that Annie does not know the secret word. Since John chooses path A or B at random, Annie has a 1=2 chance of cheating at one round. If the above steps are repeated for many rounds, Annie's chance of successfully anticipating all of John's requests would become vanishingly small. Thus, if Annie reliably appears at the exit John names, he can conclude that she is very likely to know the secret word. Other problems that involve zero-knowledge proofs are the square root of an integer modulo n, graph isomorphism, integer factorization and the discrete logarithm problem. On this paper we focus on zero-knowledge protocols based on the discrete logarithm problem devise a scheme by which Annie can prove that she knows the magic word without telling it to John. The scheme steps are now described.

(1) John waits outside the cave as Annie goes in.
(2) Annie at random takes either path A or B inside the cave.
(3) John enters the cave and shouts the name of the path he wants her to use to return either A or B, selected at random.
(4) Annie does that using the secret word if needed to open the magic door.
(5) The above steps are repetitive n times until John are convinced that Annie knows the secret word.

VI. Conclusion

The ever increasing ability to recognize and gather large amounts of data, analyzing the data using data mining process and decision on the results gives prospective benefits to organizations.The principle of zero-knowledge proof to facilitate fast, distributed,trustworthy yetsecure public verification. But, such repositories also contain private and sensitive information and releasing the personal information can cause significant damage to data owner. Reliable trust is built in anonymous environment without revealing any sensitive identifier. Hence there is increased need to discover and distribute the databases, without compromising the privacy of the individual's data. So in this paper zero knowledge protocol (zkp) seems effective concept for de-identification our secret information without any knowledge to data miner. This protocol is most recent concept in cryptography techniques. A zero-knowledge proof allows to reveal knowledge of a secret while enlightening no information whatsoever of use to the verifier in assigning this demonstration of knowledge to others. To the best of our knowledge, this is the first attempt of implementing and evaluating ZKIP protocols with emphasis on low-end devices. This work's results can be used from developers who wish to achieve certain levels of security and privacy in their applications. We obtain enhanced privacy because the result obtained is in perturbed form, so the privacy of original data is retained giving valid data mining result.

References

[1] Anand Sharma and Vibha Ojha*"Implementation of Cryptography for Privacy Preserving Data Mining"* International Journal of Database Management Systems (IJDMS) Vol.2, No.3, August 2010.

[2] Ioannis Chatzigiannakis, Apostolos Pyrgelis, Paul G. Spirakis, Yannis C. Stamatiou *"Elliptic Curve Based Zero Knowledge Proofs and Their Applicability on Resource Constrained Devices"* University of Patras Greece, arXiv: 1107.1626v1 [cs.CR] 8 Jul 2011.

[3] Kiran P, S Sathish Kumar and Dr Kavya *"A Novel Framework using Elliptic Curve Cryptography for Extremely Secure Transmission in Distributed Privacy Preserving Data Mining",*An International Journal (ACIJ), Vol.3, No.2, March 2012.

[4] Muthu Lakshmi and Dr. K Sandhya Rani," *Privacy Preserving Association Rule Mining Without Trusted Party for Horizantally Partioned Databases"* International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.2, No.2, March 2012.

[5] Benny Pinkas *"Cryptographic techniques for privacy-preserving data mining"* Published in SIGKDD Explorations, Volume 4, Issue 2, January 10, 2003.

[6] Umesh Kumar Singh, Bhupendra Kumar Pandya, Keerti Dixit*" An Overview on Privacy Preserving Data Mining Methodologies"* International Journal of Engineering Trends and Technology- Sep to Oct Issue 2011 ISSN: 2231-5381.

[7] Nathani sushma, Priyanka Kanaparthi*," Multidimensional Techniques for Privacy Preservation in Datasets"* International Journal of Computer Science and technology (IJCST) Vol. 2, Issue 4, Oct- Dec. 2011.

[8] Lambodar Jena, Ramakrushna Swain*," A Comparative Study on Privacy Preserving Association Rule Mining Algorithms"* International Journal of Internet Computing, Volume-I, Issue 1, 2011.

[9] Archana Tomar, Vineet Richhariya, Mahendra Ku. Mishra*," A Improved Privacy Preserving Algorithm Using Association Rule Mining in Centralized Database",*International Journal of Advanced Technology & Engineering Research (IJATER) ISSN NO: 2250-3536 Volume 2, Issue 2, March 2012.

[10] Ashraf El-Sisi*," Fast Cryptographic Privacy Preserving Association Rules Mining on Distributed Homogenous Database",*The International Arab Journal of Information Technology, Vol. 7, No. 2, April 2010.

[11] Fuad Al-Yarimi, Sonajharia Minz*," Multilevel Privacy Preserving in Distributed Environment using Cryptographic Technique"*Proceedings of the World Congress on Engineering 2012 Vol I .WCE 2012, July 4 - 6, 2012, London.

[12] P.Kamakshi , Dr.A.Vinaya Babu*," Preserving Privacy and Sharing the Data in Distributed Environment using Cryptographic* Technique *on Perturbed data"* Journal of Computing, Volume 2, Issue 4, April 2010.

