ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 7, Issue 4, April 2019

Privacy Preserving Data Mining using Cryptography

Arati Dhake[1], Tejal Gharate[2], Komal Mali[3], Sapana Patil[4]

U.G. Student, Department of Computer Engineering,SSBT's COET BambhoriJalgaon, India[1]

Abstract: Now a day privacy preserving data mining is most important because Data mining is allows sharing sensitive data for analysis purpose.So people have become increasingly unwilling to share their data, often resulting in individuals either refusing to disclose their data or providing wrong data. Nowadays, privacy preserving data mining has been studied extensively, because of the wide proliferation of sensitive information on the internet. We discuss method for Perturbation, K-Anonymization, condensation, and Distributed Privacy Preserving Data mining. In this paper, we have given a review of the state-of-the-art methods for privacy and analyze the representative technique for privacy preserving data mining and point out their merits and demerits.

I. Introduction

Data mining research provides several techniques The need of privacy preserving data mining has become more significant 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 methods such as kanonymity, classification, association rule mining,clustering have been recommended in recent years in order to perform privacy preserving data mining. Furthermore,the problem has been discussed in multiple communities such as the database, the statistical disclosure control(SDC) and the cryptography. Data mining techniques have been developed successfully to extracts knowledge in order to support a variety of domain areas marketing, weather forecasting, medical diagnosis, and national security. But it is still a challenge to mine some kinds of data without violating the data owners 'privacy .For example, how to mine patients 'private data is an ongoing problem in health care applications. As data mining become more pervasive, privacy concerns are increasing. Commercial concerns are also concerned with the privacy issue. Most concerns gather details about individuals for their own particular needs. More often, different departments within an organization themselves may find it necessary to share information. In those cases, each organization or unit must ascertain that the privacy of the individual is not compromised or that sensitive business information is not divulged .Consider, for example, a government, or more specifically, one of its security branches interested in developing a system for determining, from passengers whose baggage has been checked, those who must be subjected to additional security measures. The data indicative of such need for further examination stems from a lot of sources like police records; airports; banks; general government statistics; and passenger information records that generally include personal information (such as name and passport number); demographic data (such as age and gender); flight information (such as departure, destination, and duration); and expenditure data (such as transfers, purchasing and bank transactions). In most countries, this information is considered as private and to avoid intentionally or unintentionally exposing confidential.Information about an individual, it is illegal to make such information freely available.

Though many types of preserving individual information have been developed, there are ways for circumventing these methods. For example, in order to preserve privacy, passenger information records can be de-identified before the records are shared with anyone who is not permitted directly to access the relevant data. This can be done by removing from the dataset unique identity fields, such as name and passport number. Even though if this information is deleted, there are still other forms of information both personal and behavioral (e.g. date of birth,
zip code, gender, number of children, number of calls, number of accounts) that, when connected with other available datasets, could easily recognize subjects. To avoid these types of violations, we require various data mining algorithms for privacy preserving. Weanalyze recent work on these topics, presenting general frameworks that we use to compare and contrast different approaches.

Explosive progress in networking, storage, and processor technology has led to the creation of ultra large databases that record unprecedented amount of transactional information. The main problem is that with the availability of non-

DOI: 10.15680/IJIRCCE.2019. 0704044

sensitive information or unclassified data, one is able to infer sensitive information that is not supposed to be disclosed. Despite its benefits in various areas such as marketing, business, medical analysis, bioinformatics and others, data mining can also pose a threat to privacy in database security if not done or used properly. Privacy preserving data mining, is a novel research direction in data mining and statistical databases, where data mining algorithms are analyzed for the side-effects they incur in data privacy. [1].in privacy preserving data mining (PPDM), the goal is to perform data mining operations on sets of data without disclosing the contents of the sensitive data. Since the results of the mining tell us something about the data, some information about the original data is leaked to the mining results. This leads to privacy loss. If the data is perturbed on the other hand for privacy concerns, it leads to information loss, which typically refers to the amount of critical information preservedabout the datasets after the perturbation [3].Thus we need to work towards minimizing both privacy loss and information loss. Many approaches emerged for privacy preserving data mining. The first approach involved perturbing the input before mining. Though it has the benefit of simplicity it does not provide a formal framework for proving how much privacy is guaranteed. Secure Computation technique [2] has the advantage of providing a well-defined model for privacy using cryptographic techniques and is also accurate. However it is a slower method.

In PRBAC technique, access to sensitive objects (SOBS) is based on roles [1].But it has the drawback of space complexity. i.e., as all data are stored in the Database Server, it leads to a large memory requirement. Also, risk of illegal access of data is not completely ruled out as the entire data is stored at one site. In our paper, we address these problems by applying vertical fragmentation and cryptographic techniques for data storage.Here, we propose a new approach to privacy preserving data mining based on cryptographic role based access control approach (PCRBAC) where we have 2 sets of object: Sensitive objects (SOBS) and Non sensitive objects (NSOBS). Using the data mining technique, users are allowed to mine different sets of data based on their roles. The data is first classified as sensitive objects and non- sensitive objects. Sensitive objects are encrypted and stored. The permitted user can access the sensitive objects only after decryption ensuring privacy

II. LITERATURE SURVEY

Detecting fake news on social media poses several new and challenging research problems. Though fake news itself is not a new problem

– nations or groups have been using the news media to execute propaganda or influence operations for centuries

– the rise of web-generated news on social media makes fake news a more powerful force that challenges traditional journalistic norms. There are several characteristics of this problem that make it uniquely challenging for automated detection. First, fake news is intentionally written to mislead readers, which makes it nontrivial to detect simply based on news content. The content of fake news is rather diverse in terms of topics, styles and media platforms, and fake news attempts to distort truth with diverse linguistic styles while simultaneously mocking true news. For example, fake news may cite true evidence within the in-correct context to support a non-factual claim. Thus, existing hand-crafted and data-specific textual features are generally not sufficient for fake news detection. Other auxiliary information must also be applied to improve detection, such as knowledge base and user social engagements. Second, exploiting this auxiliary information actually leads to another critical challenge: the quality of the data itself. Fake news is usually related to newly emerging, time

-critical events, which may not have been properly verified by existing knowledge bases due to the lack of corroborating evidence or claims. In addition, users social engagements with fake news produce data that is big, incomplete, unstructured, and noisy. Effective methods to differentiate credible users extract useful post features and exploit network interactions are an open area of research and need further investigations.

LITERATURE REVIEW

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. 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 thenrelease these distorted data.The third approach is Cryptography based approach [6, 7] which has been developed to solve thefollowing 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 techniques to securely store the data and taking into account the flexibility of role based access control models to access the stored data.

For a data set with a single confidential attribute, univariate microaggregation (UMA) involves sorting records by the confidential attribute, grouping adjacent records into groups of small sizes, and replacing the individual confidential values in each group with the group average.

Similar to SAN and MN, UMA causes bias in the variance of the confidential attribute, as well as in the relationships between attributes. Multivariate microaggregation (MMA) [3], differs from UMA in that it groups data using a clustering technique that is based on a multi-dimensional distance measure. As a result, the relationships between attributes are expected to be better preserved. However, this benefit comes with a higher computational time complexity, which could be inefficient for large data sets.

Xiao-Bai Li and SumitSarkar proposed a method, called perturbation trees [4] that uses a recursive partitioning technique to divide a data set into subsets that contain similar data. The partitioned data are perturbed using the subset average. Since the data are partitioned based on the joint properties of multiple confidential and nonconfidential attributes, the relationships between attributes are expected to be reasonably preserved. Further, the proposed method is computationally efficient.The algorithm is based on the kdtree technique.

Li Liu , Murat Kantarcioglu and BhavaniThuraisingham[5] proposed an individually adaptable perturbation model, which enables the individuals to choose their own privacy levels. This method enables users to choose different privacy levels without significant data mining performance degradation. The effectiveness of the new approach is demonstrated by various experiments conducted on both synthetic and real-world data sets.

Reconstruction is a very important step for the perturbation based PPDM approaches. It is found that when applied to real-world data sets reconstruction could be a problem. So some PPDM methods were proposed which skip this reconstruction step and compute the data mining results directly.

HillolKargupta et al.[6] proposed a methodology which attempts to hide the sensitive data by randomly modifying the data values often using additive noise. It is noted that random objects (particularly random matrices) have "predictable" structures in the spectral domain and it develops a random matrix-based spectral filtering technique to retrieve original data from the dataset distorted by adding random values. This paper illustrates some of the challenges that these techniques face in preserving the data privacy. It showed that under certain conditions it is relatively easy to breach the privacy protection offered by the random perturbation based techniques.

Reconstruction is the main and important part for the perturbation based approaches. Multiplicative data perturbation is the focus of the paper of Yingpeng Sang, Hong Shen, and HuiTian [7], they have proposed effective methods to reconstruct the original data from the perturbed data by random projections, reconstruction achieves a higher recovery rate .The results reveal the risks of employing random projections in the multiplicative data perturbation. Successful reconstructions essentially mean the leakage of privacy, so our work identifies the possible risks of RP when it is used for data perturbations. Distributed anonymous data perturbation method for privacy-preserving data mining is proposed by Feng LI†,

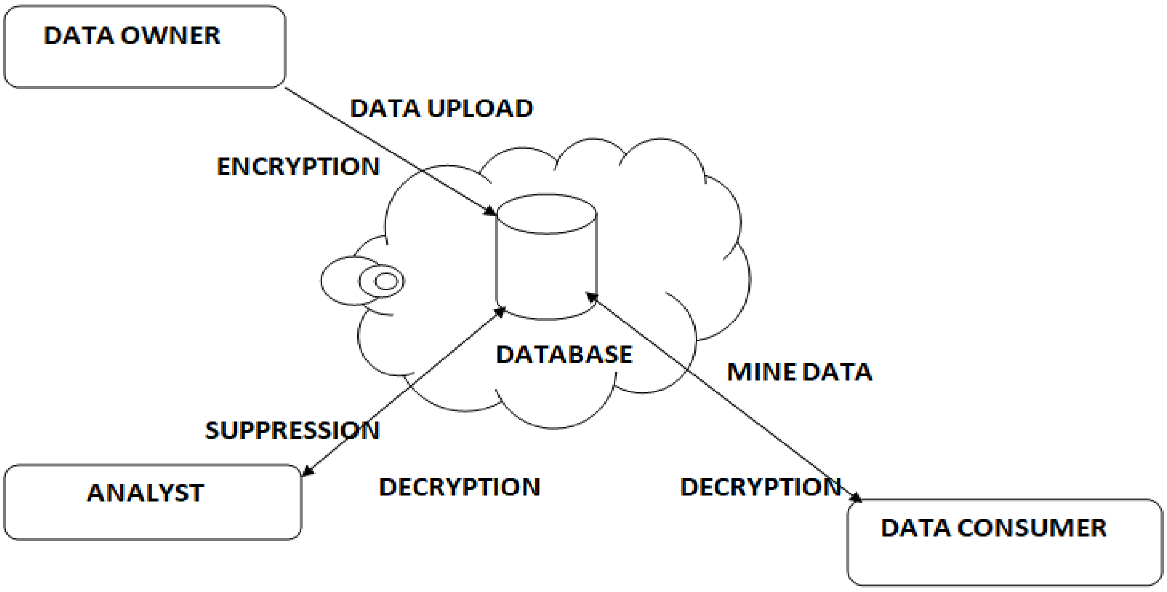
Jin MA and Jian-hua LI [8]. Cryptography-based secure multiparty computation is a main approach for privacy preserving. However, it shows poor performance in large scale distributed systems. Meanwhile, data perturbation techniques are comparatively efficient but are mainly used in centralized privacy-preserving data mining (PPDM). In this paper, a light-weight anonymous data perturbation method is proposed for efficient privacy preserving in distributed data mining. The privacy constraints for data perturbation based PPDM are defined in a semi-honest distributed environment.

Two protocols are proposed to address these constraints and protect data statistics and the randomization process against collusion attacks: the adaptive privacy-preserving summary protocol and the anonymous exchange protocol. Finally, a distributed data perturbation framework based on these protocols is proposed to realize distributed PPDM. Experiment results show that this approach achieves a high security level and is very efficient in a large scale distributed environment.

HillolKarGupta et al. [9] presented in the paper that random objects have predictable structures in the spectral domain and then it develops a random matrix-based filtering technique to retrieve original data set from data set distorted by adding random values. In many cases, it shows that random data distortion preserves very little privacy. So this paper addresses that under some certain conditions, it is relatively easier to breach the privacy protection offered by random perturbation methods.

III.PROPOSED SYSTEM

Architecture



Cryptographic Approach

Standard Encryption Algorithm:

```
Get the SOB;
Convert each character of the attribute
value into its corresponding ASCII value;
Convert it into binary value;
Perform NOT operation;
```

DOI: 10.15680/IJIRCCE.2019. 0704044

Add flag bits between two binary values;

Standard Decryption Algorithm:

```
Get the encrypted SOB;
Retrieve the binary values;
Perform NOT operation;
Convert it into corresponding ASCII value;
Retrieve the characters corresponding to the
ASCII value;
```

Data Suppression

A popular disclosure protection method is data suppression, which alters individual data in a way such that the summary statistics remain approximately the same. Problems in data mining are somewhat different from those in SDBs. A data mining technique, such as classification or numeric prediction, essentially relies on discovering relationships between data attributes. Preserving such relationships may not be consistent with preserving summary statistics. This study focuses on perturbing numeric data. Here a single confidential attribute is considered, although this method extends naturally to situations with multiple confidential attributes. Let X be a confidential attribute, and Y be the perturbed value of X. Traub et al. [1] proposed a simple additive noise method (SAN) as below:

$Y = X + e;$

Where the noise term e has a mean zero and a variance $p \times \sigma^2$ and p is a variance proportion parameter determined by the user. A drawback of this method is that the noise is independent of the scale of X. That is, the expected amount of noise added to X is the same no matter if X = \$20,000 or X = \$200,000. To overcome this problem, the multiplicative noise method (MN) was proposed [2], which can be written as:

$Y = X * e$

Where the noise e has a mean of one. The SAN and MN methods cause bias in the variance of the confidential attribute, as well as in the relationships between attributes. Another popular approach to data suppression is microaggregation (MA) [3]. MA perturbs data by aggregating confidential values, instead of adding noise.

Experimental Setup

Our experimental setup includes three users: data owner, analyst & data consumer

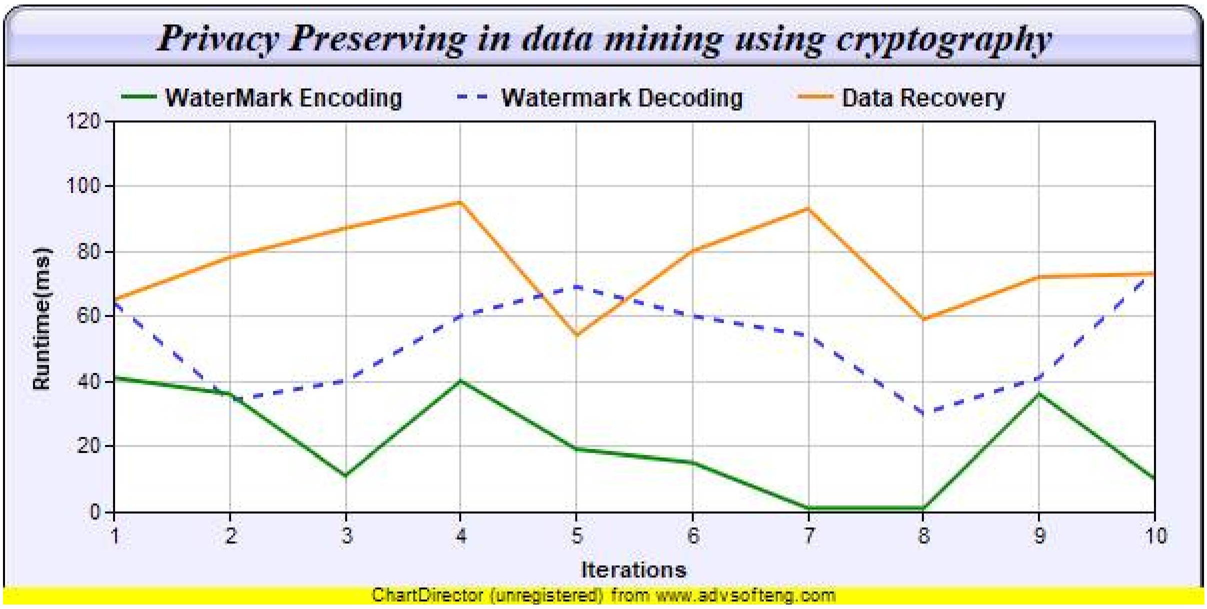


Fig:Graph

VI.CONCLUSION AND FUTURE WORK

We developed privacy preserving in data mining by using the Cryptographic Technique. Here we assumed decryption occurs entirely at the server . This project can be designed for any type of dataset but we are developed for employee dataset using AES-256 and suppression algorithm by considering the modules such as user, analyst, and data consumer. In this we have successfully preserved the privacy. In future the project can be work with multiple domain for protecting privacy also combination of algorithms can be implemented to improve security using AES-256 algorithm.

REFERENCES

1. J.F. Traub, Y. Yemini, and H. Wozniakowski, “The Statistical Security of a Statistical Database,” ACM Trans. Database system vol. 9, no. 4, pp. 672-679, 1984.
2. N.R. Adam and J.C. Wortmann, “Security-Control Methods for Statistical Databases: A Comparative Study,” ACM Computing Surveys, vol. 21, no. 4, pp. 515-556, 1989.
3. J. Domingo-Ferrer and J.M. Mateo-Sanz, “Practical Data-Oriented Microaggregation for Statistical Disclosure Control,” IEEE Trans.Knowledge and Data Eng., vol. 14, no. 1, pp. 189-201, 2002
4. Xiao-Bai Li and SumitSarkar, “A Tree based Data Perturbation Approach for Privacy-Preserving Data Mining”,IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 18, NO. 9, SEPTEMBER 2006
5. Murat Kantarcioglu, BhavaniThuraisingham ,”The applicability of the perturbation based privacy preserving data mining for real-world data”, IEEE, Li Liu, 2007
6. HillolKargupta , SouptikDatta, Qi Wang and KrishnamoorthySivakumar , “ On the Privacy Preserving Properties of Random Data Perturbation Techniques”,ICDM2003.Third IEEE International conference on 19-22,Nov 2003
7. Yingpeng Sang, Hong Shen, and HuiTian, “Effective reconstruction of data perturbed by random projections”, IEEETRANSACTIONS ON COMPUTERS, VOL. 61, NO. 1, JANUARY 2012
8. Feng LI†, Jin MA, Jian-huaLI , “Distributed anonymous data perturbation method for privacy preserving data mining”, Journal of Zhejiang University SCIENCE A
9. HillolKarGupta and SouptikDatta,Qi Wang and KrishnamoorthySivakumar,”Random Data Perturbation Techniques and Privacy Preserving data Mining”,IEEE International Conference on
data Mining 2003
10. C.C. Aggarwal and P.S. Yu, “A Condensation Approach to Privacy Preserving Data Mining,” Proc. Ninth Int’l Conf. Extending Database Technology, pp. 183-199, 2004. 11. P.Samarati, “Protecting respondent's privacy in micro data release”, IEEE Transaction on knowledge and Data Engineering,pp.010-027,2001.
12. L. Sweeney, “k-anonymity: a model for protecting privacy ", International Journal on Uncertainty, Fuzziness and Knowledge based Systems, pp. 557-570, 2002.
13. Laur, H. Lipmaa, and T. Mieli' ainen,"Cryptographically private support vector machines". In Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 618-624,2006
14. Ke Wang, Benjamin C. M. Fung1 and Philip S. Yu, "Template based privacy preservation in
Classification problems", In ICDM, pp. 466- 473, 2005
15. Yang Z., ZhongS.Wright R.” Privacy-Preserving Classification of Customer Data without Loss of Accuracy” SDM Conference, 603-610, 2006