A

Seminar-II Report

on

AN APPROACH FOR PRESERVING PRIVACY IN DATA MINING AND ITS TECHNIQUES

Submitted in Partial Fulfillment of
the Requirements for the Degree
of

Bachelor of Engineering

in

Computer Engineering

to

North Maharashtra University, Jalgaon

Submitted by

Puja Anil Naval

Under the Guidance of

Mr. Sandip S. Patil

DEPARTMENT OF COMPUTER ENGINEERING
SSBT's COLLEGE OF ENGINEERING AND TECHNOLOGY,
BAMBHORI, JALGAON - 425 001 (MS)
2016 - 2017

SSBT's COLLEGE OF ENGINEERING AND TECHNOLOGY, BAMBHORI, JALGAON - 425 001 (MS)

DEPARTMENT OF COMPUTER ENGINEERING

CERTIFICATE

This is to certify that the Seminar-II entitled *An Approach for Preserving Privacy in Data Mining and its Techniques*, submitted by

Puja Anil Naval

in partial fulfillment of the degree of *Bachelor of Engineering* in *Computer Engineering* has been satisfactorily carried out under my guidance as per the requirement of North Maharashtra University, Jalgaon.

Date: October 4, 2016
Place: Jalgaon

Mr. Sandip S. Patil
Guide

Prof. Dr. Girish K. Patnaik
Head

Prof. Dr. K. S. Wani
Principal

Acknowledgements

"No work can be accomplished unless it has evolved as a result of co-operating, assistance and understanding of some knowledgeable group of people". I take the opportunity to thank our Principal Prof. Dr. K. S. Wani and Head of Department Prof. Dr. Girish K. Patnaik for providing all the necessary facilities, which were indispensable in the completion of seminar. I would like to thank my guide Mr. Sandip S. Patil for providing to be a great help by giving us guidance through their vast experience and intellectual skills. I would like also thankful to all the staff members of the Computer Engineering Department. I would also like to thank the college for providing the required magazines, books and access to the internet for collecting information related to the seminar. Finally, I would like to thanks my parents and friends.

Puja Anil Naval

Contents

List of Tables

List of Figures

Abstract

The Privacy preserving Data mining(PPDM) has been among the important issues of current research that deals with preserving privacy of individuals 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. This seminar, include a unique concept of combining different PPDM techniques which provides high level security and integrity to confidential data. This seminar 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 seminar. It has been observed that slicing preserves better data utility and thus in this seminar tried to merge slicing with one of the best security mechanism that is Cryptography.

Keywords: Privacy-preserving data mining, data mining, research challenges, privacy preserving techniques, slicing.

Chapter 1

Introduction

Data mining aim to extract useful information from huge amount of data, whereas privacy preservation in data mining aims to preserve these data against disclosure or loss. The main objective of privacy preserving data mining is to develop algorithms for modifying the original data in some way, so that the private data and the private knowledge remain private even after the mining process.

Chapter is of 5 sections. Section 1.1 discuss data mining concept and its process. Goals of data mining discuss in the section 1.2. In section 1.3 discuss concept of privacy preserving data mining. Privacy preserving data mining framework is discuss in section 1.4. Section 1.5 gives summary.

1.1 Data Mining

Data mining coming from the rapid growth of information, is the process of extract useful information or patterns from large amount of data. It is also commonly known as Knowledge Discovery from Data(KDD).The data mining process shown in fig. 1.1.

Data mining is an emerging field which connects different major areas like databases, artificial intelligence and statistics. Data mining is a powerful tool that can investigate and extract previously unknown patterns from large amounts of data. The process of data mining requires a large amount of data to be collected into a central site. In modern days organizations are extremely dependent on data mining in results to provide better service, achieving greater profit, and better decision-making. For these purposes organizations collect huge amount of data. For example, business organizations collect data about the consumers for marketing purposes and improving business strategies, medical organizations collect medical records for better treatment and medical research. With the rapid advance of the Internet, networking, hardware and software technology there is remarkable growth in the amount of data that can be collected from different sites or organizations. Huge volumes of Data collected in this manner also include sensitive data about individuals. It is obvious that if a

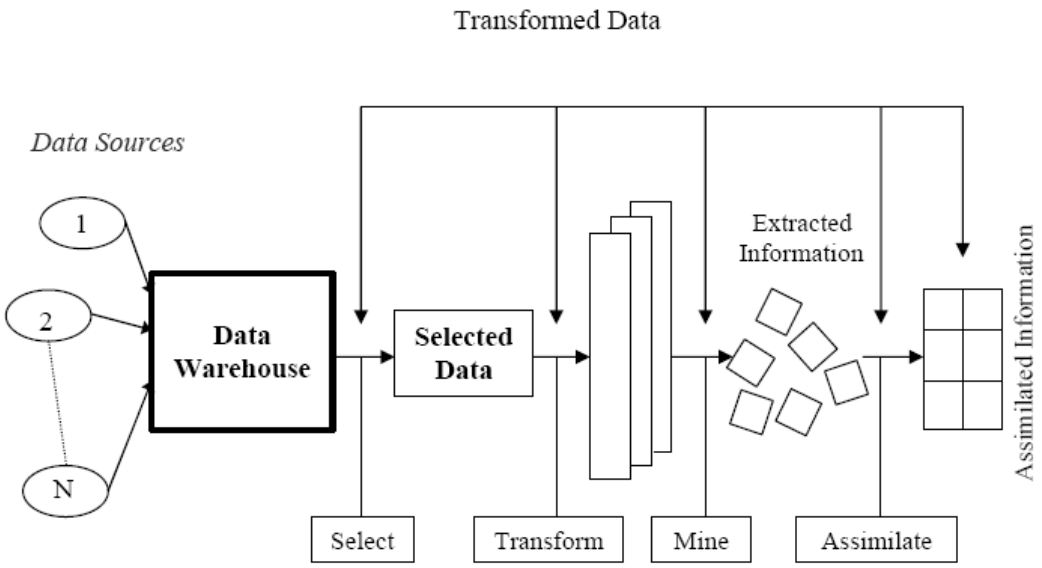

Figure 1.1: The Data Mining Process

data mining algorithm is run against the union of different databases, the extracted knowledge not only consists of discovered patterns and correlations that are hidden in the data but it also reveals the information which is considered to be private. Privacy is an important issue in many data mining applications that deal with health care, security, financial and other types of sensitive data.

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. Privacy can for instance be threatened when data mining techniques uses the identifiers which themselves are not very sensitive ,but are used to connect personal identifiers such as addresses, names etc., with other more sensitive personal information. The simplest solution to this problem is to completely hide the sensitive data or not to include such sensitive data in the database.

But this solution is not ideal and accurate because in many applications, like medicine research, DNA research etc. Different organizations or institutions wish to conduct a joint research on their databases because combining their data will definitely provide better results and mutual benefit to the organizations. In this scenario organizations want to share the data but neither of the institute or organizations want to disclose its database or private information about their clients due to privacy concern. In such a situation it is not only necessary to protect private and sensitive information but it is also essential to facilitate the use of database for investigation or for other purposes. Privacy preserving data mining is a special data mining technique which has emerged to protect the privacy of sensitive data and also give valid data mining results. In this seminar propose framework for preserve the

privacy at both at customer end and data owner site. The original data is systematically transformed using privacy preserving data mining technique.

1.2 Goals of Data Mining

The goals of data mining fall into the following classes:

- **Prediction:** Data mining can show how certain attributes within the data will behave in the future.
- **Identification:** Data patterns can be used to identify the existence of an item, an event, or an activity.
- **Classification:** Data mining can partition the data so that different classes or categories can be identified based on combinations of parameters.
- **Optimization:** One eventual goal of data mining may be to optimize the use of limited resources such as time, space, money, or materials and to maximize output variables such as sales or profits under a given set of constraints.

1.3 Privacy Preserving Data Mining

The knowledge discovered by various data mining techniques may contain private information about people or business. Preservation of privacy is a significant aspect of data mining and thus study of achieving some data mining goals without losing the privacy of the individuals. The analysis of privacy preserving data mining algorithms should consider the effects of these algorithms in mining the results as well as in preserving privacy. Privacy preserving data mining deals with hiding an individuals sensitive identity without sacrificing the usability of 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. Privacy preserving data mining 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 privacy preserving data mining, Slicing, etc. are some of the privacy preserving data mining 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 privacy preserving data mining can be made. Combining the privacy preserving data mining techniques provides more robust, stable and secure algorithm. Thus in our research we have mainly focused on combining two techniques of privacy preserving data mining 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 aim of privacy preserving data mining algorithms is to mined appropriate information from huge amounts of data while protecting at the same time thoughtful information.
The several approaches used by privacy preserving data mining can be summarized as below:

- The data is altered before delivering it to the data miner.
- The data is distributed between two or more sites, which cooperate using a semi-honest protocol to learn global data mining results without revealing any information about the data at their individual sites.
- While using a model to classify data, the classification results are only revealed to the designated party, who does not learn anything else other that the classification results, but can check for presence of certain rules without revealing the rules.

1.4 Privacy Preserving Data Mining Framework

The framework for privacy preserving data mining is shown in fig. 1.2. In data mining or knowledge discovery from databases (KDD) process the data (mostly transactional) is collected by single/various organization/s and stored at respective databases. Then, it is transformed to a format suitable for analytical purposes, stored in large data warehouse/s and then data mining algorithms are applied on it for the generation of information/knowledge [12].

The first level is raw data or databases where transactions exist in. The second level is data mining algorithms and techniques that ensure privacy. The third level is the output of different data mining algorithms and methods [11].

At level 1, the raw data collected from a single or multiple databases or even data marts is transformed into a format that is well suited for analytical purposes. Even at this stage, privacy concerns are needed to be taken care of. Researchers have applied different techniques at this stage but most of them deal with making the raw data suitable for analysis [12].

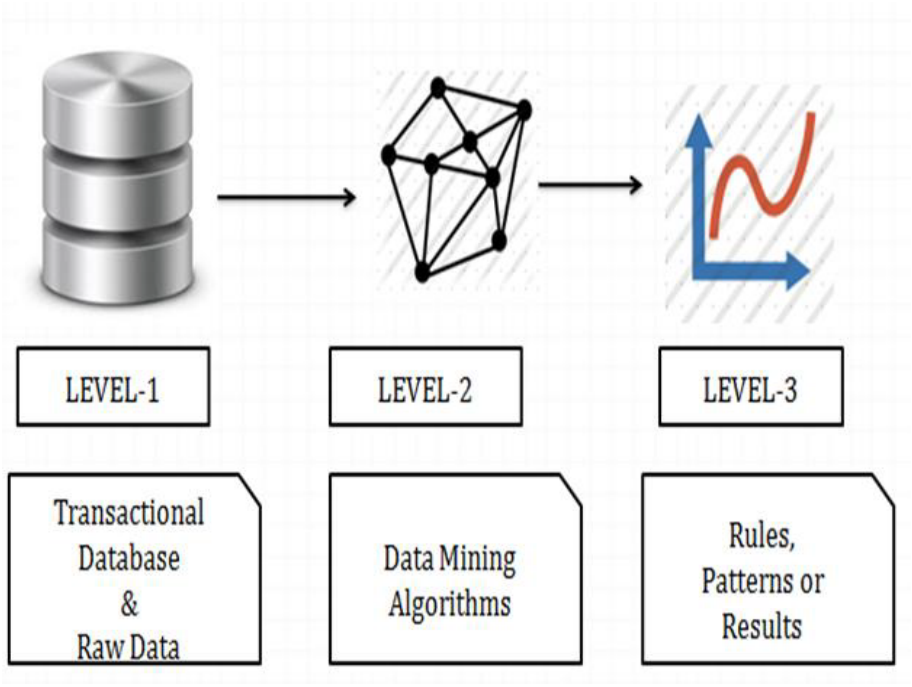

Figure 1.2: Privacy Preserving Data Mining Framework

At level 2, the data from data warehouses is subjected to various processes that make the data sanitized so that it can be revealed even to untrustworthy data miners. The processes applied at this stage are blocking, suppression, perturbation, modification, generalization, sampling etc. Then, the data mining algorithms are applied to the processed data for knowledge/information discovery. Even the data mining algorithms are modified for the purpose of protecting privacy without sacrificing the goals of data mining [12].

At level 3, the information/knowledge so revealed by the data mining algorithms is checked for its sensitiveness towards disclosure risks. The embedding of privacy concerns at three levels, but any combination of these may be used [12].

1.5 Summary

In this chapter, discuss the basic concept of privacy preserving data mining and its techniques. The rest of this seminar report is organized as follows. In chapter 2, describes about the related work done in the field of privacy preserving data mining and research challenges in privacy preserving data mining. The proposed methodology that tries to overcome the flaws of privacy preserving data mining describe in chapter 3. In chapter 4 discuss about uses and issues of privacy preserving data mining and finally conclusion and future work discuss in chapter 5.

Chapter 2

Literature Survey

Privacy preserving uses the data mining techniques to protect the users information from the intruders. It is essential to maintain a ratio between privacy protection and knowledge discovery. The goal is to hide sensitive item sets so that the adviser cannot extract the modified database.

In this chapter, section 2.1 discuss the related work done in the field of privacy preserving data mining techniques. Research challenges in privacy preserving data mining discuss in section 2.2. Section 2.3 gives summary.

2.1 Privacy Preserving Data Mining Techniques

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 seminar 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 and one of its fundamentals i.e. Rotation in their paper. An orthogonal matrix is generated by the Walsh-Hadamard transformation, 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.

2.2 Challenges in Privacy Preserving Data Mining

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.

1. **Internal and External attacks, Cyber threats:**
   One of the major threats people face today is Cyber Crime 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 banks 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 Banks 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.

2. **Fraud in Credit Cards and Individuals 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. A similar and a more serious theft is identity theft. Here one pretends to be an identity of another person by obtaining that persons personal information and carrying out all types of transactions under the other persons 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.

3. **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:** The basic form of the data in a table consists of following four types of attributes: (i) Explicit Identifiers is a set of attributes containing information that identifies a record owner explicitly such as name, SS number etc.
     (ii) Quasi Identifiers is a set of attributes that could potentially identify a record owner when combined with publicly available data.
     (iii) Sensitive Attributes is a set of attributes that contains sensitive person specific information such as disease, salary etc.
     (iv) Non-Sensitive Attributes is a set of attributes that creates no problem if revealed even to untrustworthy parties.

     Anonymization refers to an approach where identity or/and sensitive data about record owners are to be hidden. K-anonymity is used for generalization and suppression for data hiding. 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. It even assumes that sensitive data should be retained for analysis. Its obvious that explicit identifiers should be removed but still there is a danger of privacy intrusion

when quasi identifiers are linked to publicly available data. Such attacks are called as linking attacks. For example attributes such as DOB, Sex, Race, and Zip are available in public records such as voter list.Such records are available in medical records also, when linked, can be used to infer the identity of the corresponding individual with high probability as shown in fig. 2.1.

Sensitive data in medical record is disease or even medication prescribed. The quasi-identifiers like DOB, Sex, Race, Zip etc. are available in medical records and also in voter list that is publicly available. The explicit identifiers like Name, SS number etc. have been removed from the medical records.

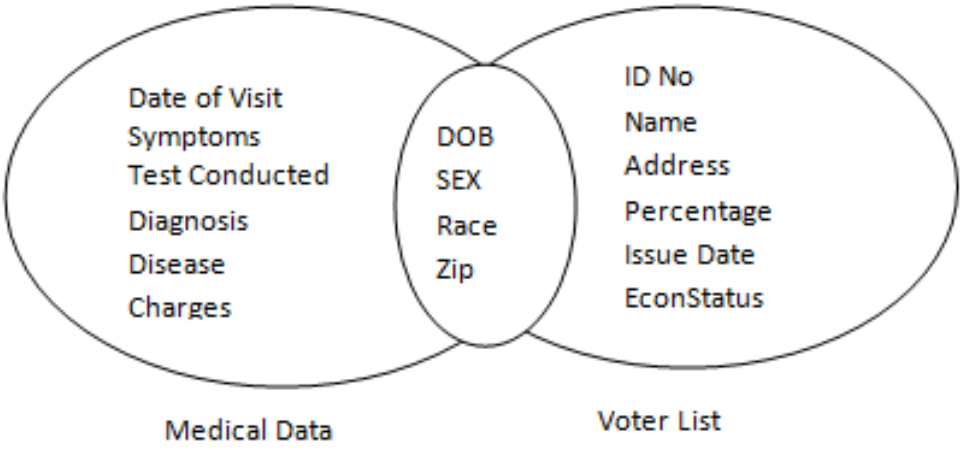

Figure 2.1: Linking Attack

Advantages:

- This method is protects identity disclosure when it is releasing sensitive information.

Disadvantages:

- It is prone to homogeneity attack and the background knowledge attack.
- Does not protect attribute disclosure to sufficient extent.
- It has the limitation of k-anonymity model which fails in real scenario when the attackers try other methods.

- **Cryptographic Technique:** The another method in Privacy preserving data mining is cryptography. This branch became famous for two reasons: First, cryptography provides a well-defined model for privacy, which has methodologies for proving and quantifying it. Secondly, there exists a huge tool set of cryptographic algorithms. However, recent work has pointed that cryptography does not defend the output of a computation. Instead, it prevents privacy leaks within the process

of computation [8]. Thus, for huge databases this algorithm does not proves to be a strong technique as this technique fails to protect the output of computation. Thus as a result mining the result may break the privacy of individuals record [9].

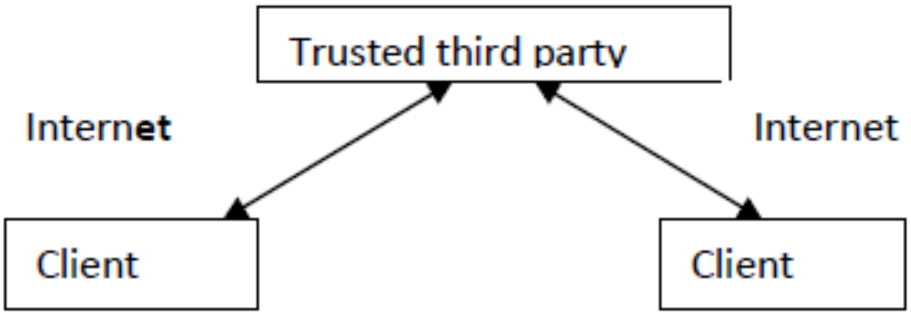

Figure 2.2: System Using Semi Trusted Third Party

Advantages:

- Cryptography offers a well-defined model for privacy for proving and quantifying it.
- There exit a vast range of cryptographic algorithms.

Disadvantages:

- It is difficult to scale when more than a few parties are involved.
- It does not guarantee that the disclosure of the final data mining result may not violate the privacy of individual records.

- **Data Perturbation:** The perturbation method has been extensively studied for privacy preserving data mining. In this method, random noise from a known distribution is added to the privacy sensitive data before the data is sent to the miner for data mining. Consequently, the data miner rebuilds an approximation to the original data distribution from the perturbed data and uses the reconstructed distribution for data mining purposes. Different individuals may have different approaches towards privacy based on cultures and customs. Unfortunately, recent privacy preserving data mining techniques based on perturbation do not allow the individuals to choose their desired privacy levels. This was a drawback as privacy was a personal choice. 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].

  Advantages:

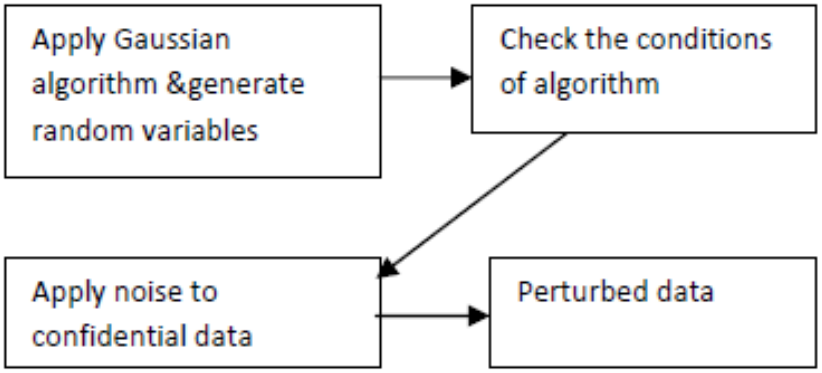

Figure 2.3: Block Diagram for Implementing Perturbation Technique

- Simple and efficient technique for building data mining models from perturbed data.
- As the distribution of the added noise is known, the data miner could rebuild the original distribution using various statistical methods and mine the rebuilt data.

Disadvantages:

- Recent privacy preserving data mining techniques based on perturbation do not allow the individuals to choose their desired privacy levels.
- As the noise is added, information loss versus preservation of privacy is always a trade off in the perturbation based approaches.

- **Randomization:** In Randomized response, the data is muddled in such a way that the central place cannot let know with probabilities better than a pre-defined threshold, whether the data from a customer contains truthful information or false information. The information received from each individual user is scrambled and if the number of users is significantly large, the aggregate information of these users can be predictable with good amount of accuracy. 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. The process of data collection in randomization method encompasses two steps. In the first step, the data providers transmit the randomized data to the data receiver. In second step, reconstruction of the original distribution of the data is done by the data receiver by employing a distribution reconstruction algorithm. Each records are treated individually irrespective of their local density [8]. The Model of Randomization is shown in fig. 2.4.

  Advantages:

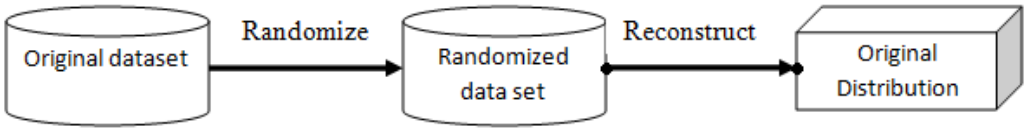

Figure 2.4: The Model of Randomization.

- It is relatively simple useful for hiding information about individuals.
- Better efficiency compare to cryptography based privacy preserving data mining technique.

Disadvantages:

- It is not required for multiple attribute databases.
- It results in high information loss.

- **Generalization:** Generalization involves replacing a value with a less specific but semantically reliable value. For example, the age of the person could be generalized to a range such as youth, middle age and adult without specifying appropriately, so as to reduce the risk of identification. Suppression involves reduce the exactness of applications and it does not liberate any information. By using this method it reduces the risk of detecting exact information. Data Generalization is the process of creating successive layers of summary data in an evaluation database. The original table is shown in fig. 2.1 and Generalization table in fig. 2.2 With the help of semantically consistent value generalization is applied on the quasi-identifiers and replaces a quasi-identifiers value. More records will have the same set of quasi-identifier values display as a result. We define an equivalence class of a generalized table to be a set of records that have the same values for the quasi-identifiers. Three types of encoding schemes have been introduced for generalization:

  - Global Recording
  - Regional Recording
  - Local Recording

  The property gifted to global recoding is that the generalized value can be replaced with the multiple occurrences of the same value. Regional record partitions the domain space into non- intersect regions and data points in the same region are represented by the region they are in. Regional record is also called multi-dimensional recoding. Local recoding allows different occurrences of the same

Table 2.1: Original Table

Name	Age	Gender	Zipcode	Disease
ALEX	20	F	12345	AIDS
BOB	24	M	12342	FLU
CARY	23	F	12344	FLU
DICK	27	M	12344	AIDS
ED	35	M	12412	FLU
FRANK	34	M	12433	CANCER
GARY	31	M	12453	FLU
TOM	38	M	12455	AIDS

Table 2.2: Generalization Table

Age	Gender	Zipcode	Disease
[20-38]	F	12***	AIDS
[20-38]	M	12***	FLU
[20-38]	F	12***	FLU
[20-38]	M	12***	AIDS
[20-38]	M	12***	FLU
[20-38]	M	12***	CANCER
[20-38]	M	12***	FLU
[20-38]	M	12***	AIDS

value to be generalized differently and does not have the above constraints. Generalization consists of substituting attribute values with less precise but semantically consistent values. Generalization maintains the correctness of the data at the record level. Generalization may also results in less specific information that may affect the accuracy of machine learning algorithms applied on the k-anonymous dataset. A considerable amount of information is lost for high dimensional data in generalization [1].

Disadvantages:

- Due to the curse of dimensionality generalization fails on high-dimensional data.
- Due to the uniform distribution assumption, generalization causes too much information loss.

- **Bucketization:** Bucketization partitions tuples in the table into buckets and then

Table 2.3: Bucketization Table

Age	Gender	Zipcode	Disease
[20-27]	*	1234*	AIDS
[20-27]	*	1234*	FLU
[20-27]	*	1234*	FLU
[20-27]	*	1234*	AIDS
[35-38]	*	124**	FLU
[35-38]	*	124**	CANCER
[35-38]	*	124**	FLU
[35-38]	*	124**	AIDS

separates the quasi identifiers with the sensitive attribute by randomly permuting the sensitive attribute values in each bucket. A set of buckets with permuted sensitive attribute values called as anonymized data. In particular, bucketization is used for anonymizing high dimensional data. Its main aim is to separation between quasi identifiers and sensitive attribute. The Bucketization table is shown in fig. 2.3 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].

2.3 Summary

In this chapter, section 2.1 described related work in the privacy preserving data mining techniques. Challenges in privacy preserving data mining is described in section 2.2. In the next chapter, describes about propose methodology that tries to overcome the flaws of privacy preserving data mining.

Chapter 3

Methodology

The proposed framework includes concept of merging different techniques aims on combining cryptographic technique and slicing. Cryptography has been one of the most used privacy prevention technique in multiparty data computation. This method prevents leakage of computations. A new technique is developed for privacy-preserving is known as Slicing. Slicing was one of the techniques that overcame the drawbacks of generalization and bucketization.

In this chapter, section 3.1 methodology that tries to overcome the flaws of privacy preserving data mining. Section 3.2 gives summary.

3.1 Method that Overcome the Flaws of Privacy Preserving Data Mining

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 individuals privacy. Membership disclosure and preserving better data utility are the advantages of slicing. Slicing preserves more attribute correlations with the sensitive attribute than bucketization. It can handle high-dimensional data and data without a clear separation of quasi-identifying and sensitive attributes. It can be effectively used based on the privacy requirement of l-diversity for preventing attribute disclosure.An efficient algorithm is developed for computing the sliced table that satisfies l-diversity. This algorithm partitions attributes into columns, then applies column generalization,and partitions tuples into buckets. The associations between uncorrelated attributes are broken; the provides better privacy as the associations between such attributes are less frequent and potentially identifying. Attributes that are highly correlated are in the same column; this preserves the correlations between such attributes. Slicing as the name says partitions the data set or attributes vertically as well horizontally. The sliced data table is shown in Table 3.1. Many algorithms like generalization, bucketization

Table 3.1: Sliced Data Table

(Age,Gender,Disease)	(Zipcode,Disease)
(20,F,AIDS)	(12345,AIDS)
(24,M,FLU)	(12342,FLU)
(23,F,FLU)	(12344,FLU)
(27,M,AIDS)	(12344,AIDS)
(35,M,FLU)	(12412,FLU)
(34,M,CANCER)	(12433,CANCER)
(31,M,FLU)	(12453,FLU)
(38,M,AIDS)	(12455,AIDS)

have tried to preserve privacy however they exhibit attribute disclosure. So to remove this problem an algorithm called slicing is used. Slicing algorithm consists of three phases:

- Attribute Partitioning
- Column Generalization
- Tuple Partitioning

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 seminar.

Figure. 3.1 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.

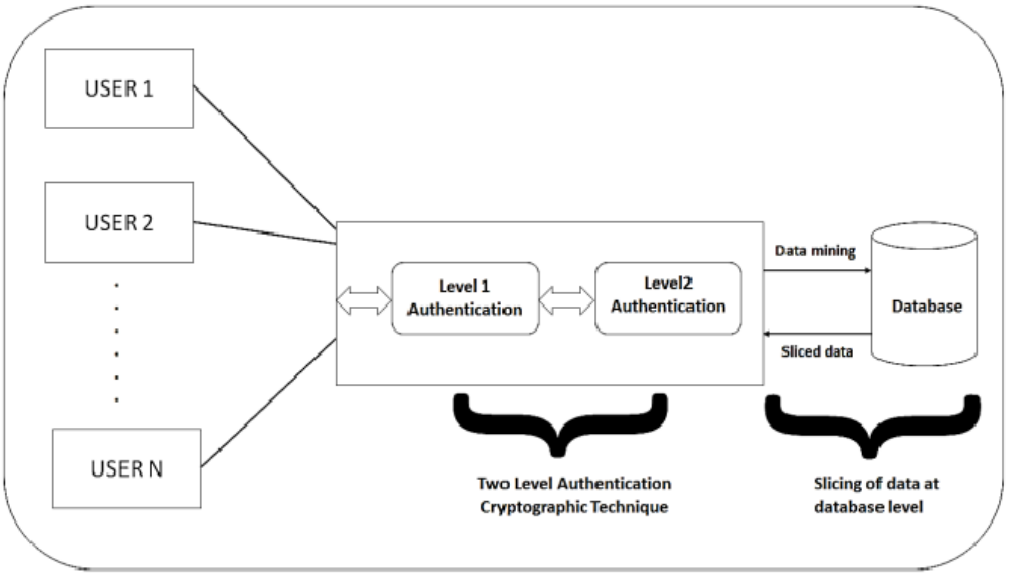

Figure 3.1: Basic Framework of the Proposed Approach

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. 3.1 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.

3.2 Summary

In this chapter, section 3.1 described methodology that tries to overcome the flaws of privacy preserving data mining. The next chapter describes uses of privacy preserving data mining, advantages of privacy preserving data mining and privacy preserving data mining issues.

Chapter 4

Discussion

In this chapter, section 4.1 will discuss uses of privacy preserving data mining. Advantage of privacy preserving data mining discuss in section 4.2. In section 4.3 discuss issues of privacy preserving data mining. Section 4.4 gives summary.

4.1 Uses of Privacy Preserving Data Mining

Data mining involves the extraction of implicit previously unknown and potentially useful knowledge from large databases. Data mining is a very challenging task since it involves building and using software that will manage, explore, summarize, model, analyses and interpret large datasets in order to identify patterns abnormalities. Privacy preserving in data mining techniques are being used increasingly in wide verity of application.

- **Privacy-Preserving Data Publishing:**
  These techniques tend to study different transformation methods associated with privacy. These techniques include methods like randomization, k-anonymity, and l-diversity. Another related issue is how the perturbed data can be used in conjunction with classical data mining methods such as association rule mining. Other related problems include that of determining privacy preserving methods to keep the underlying data useful (utility-based methods), or the problem of analysing the different definitions of privacy, and how they compare in terms of effectiveness in different scenarios.

- **Changing the results of Data Mining Applications to preserve privacy:**
  In many cases, the results of data mining applications such as association rule or classification rule mining can compromise the privacy of the data. This has spawned a field of privacy in which the results of data mining algorithms such as association rule mining are modified in order to preserve the privacy of the data. A classic example of

such techniques is association rule hiding methods, in which some of the association rules are suppressed in order to preserve privacy.

- **Query Auditing:**
  Such methods are akin to the previous case of modifying the results of data mining algorithms. Here, we are either modifying or restricting the results of queries.

- **Cryptographic Methods for Distributed Privacy:**
  In many cases, the data may be distributed across multiple sites, and the owners of the data from these different sites may wish to compute a common function. In those cases, a variety of cryptographic protocols may be used in order to communicate among the different sites, so that secure function computation is possible without revealing sensitive information.

- **Theoretical Challenges in High Dimensionality:**
  Real data sets are usually extremely high dimensional, and this makes the process of privacy preservation extremely difficult both from a computational and effectiveness point of view. It has been shown that optimal k-anonymization is NP-hard. Furthermore, the technique is not even effective with increasing dimensionality as the data can typically be combined with either public or background information to reveal the identity of the underlying record owners.

4.2 Advantages

- PPDM is very advantageous in development of various data mining techniques.
- It allows sharing of large amount of privacy sensitive data for analysis purposes.
- It has a ability to track and collect large amounts of data with the use of current hardware technology.

4.3 Privacy Preserving Data Mining Issues

The increases in digital data have raised concerns about information privacy on a global basis. Their research laid the foundation for future research that addresses privacy issues within a data mining context. The Internet has made data collection and data storage much easier, but the potential for misuse has also risen significantly. Data mining results can show models of aggregate data, but the models accuracy depends on the quality of data. The authors raise the concern that any changes to data affect the accuracy and output of data mining

models. Their approach to this problem allows the consumer to provide a perturbed value for sensitive attributes. This allows consumers to participate in the process and hopefully gives the consumer a sense of control over his or her own information. A major drawback of this approach is that output accuracy is lost during data mining activities. However, the authors maintain that small drops in accuracy are an acceptable trade-off for privacy.

A drop in the accuracy of data mining output may not be acceptable for applications where accuracy of results is significantly important. In privacy preserving data mining research, there are tradeoffs. For example, an increase in privacy preservation will result in lower accuracy in the data mining model. Privacy preserving data mining research attempts to control these drops in accuracy while still preserving individual privacy at the aggregate level. Early privacy preserving data mining research suggests that privacy and data accuracy cannot coexist in data mining activities.

One of the major problems of privacy preserving data mining is the abundant availability of personal data. Many technologies exist for supporting proper data handling, but much remains, and some barriers must be overcome in order for them to be deployed.

The privacy issues occur in data mining and that this is a generalization of the inference problem.The inference problem refers to an issue when a user can infer new knowledge by executing successive queries against a database. Since data mining techniques are designed to help the user discover new knowledge, the results of data mining can raise the likelihood of an inference to occur. This may cause ethical issues based on how the information is going to be used.

The above discussed privacy preserving data mining techniques are remarkably good, but there is always extent for more enhancements.

4.4 Summary

In this chapter, section 4.1 described uses of privacy preserving data mining. Advantage of privacy preserving data mining discribed in section 4.2. In section 4.3 described issues of privacy preserving data mining. In next chapter conclusion and future Work is present in area of privacy preserving data mining.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

This study 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.

5.2 Future Work

Further research 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 privacy preserving data mining, thus overcoming different flaws and providing a better mean for preserving data privacy.

Bibliography

[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.

[10] Ghinita.G,Tao.Y, and Kalnis.P, OnThe Anonymization of Sparse High Dimensional Data, Proc. IEEE 24th Intl Conf. Data Eng. (ICDE), 2011.

[11] Charu C. Aggarwal, Philip S. Yu Privacy-Preserving Data Mining Models and algorithmadvances in database systems 2008 Springer Science, Business Media, LLC.

[12] Ahmed HajYasien. Thesis on PRESERVING PRIVACY IN ASSOCIATION RULE MINING in the Faculty of Engineering and Information Technology Griffith University June 2007.