Survey On Incentive Compatible Privacy Preserving Data Analysis Technique

Ms. Shital Gawhale[1], Prof.Rekha Jadhav[2]

[1,2.]Computer Department G.H.Raisoni Institute Engg & Technology Pune,India

Abstract

Privacy preserving is one of the most important research topics in the data security field and it has become a serious concern in the secure transformation of personal data in recent years. For example On line shopping used different credit card and d credit card company control may try to build better data sharing or publishing models for privacy protection through privacy preserving data mining techniques (PPDM). The Incentive model is very efficient to protecting the sensitive data in privacy preserving data sharing system because it provides the secrecy against not only semi-honest adversary model and also the malicious model. The incentive data are used to check user knowledge that is the processing user is correct user or not and if valid then proceed. Secure multi-party computation (SMC) has recently emerged as an answer to this problem but SMC model they had generate only one key for the each card and after some extent time there may be chance of hacking the password by the hackers. To overcome the above said problem in our proposed system we are using symmetric key cipher algorithm to increase the security measures. In this system it will encrypt the security code so if any hacker check for the key it will be in encrypted format so no one can hack the password. And to send the keys to the mail we are using Java Mail API directly to communicate with the Gmail server.

Keywords:- PPDM,SMC ,NCC etc

1. INTRODUCTION

In today's scenario we have E-Commerce, E- Governance and personal data is distributed online, privacy of data is become the most important issue. The information found in mining can be sensitive or it can be misuse by anyone. Involving parties are realizing that combining privacy preserving techniques are applied with data mining algorithm in order to protect the extraction of sensitive information during the knowledge finding. Main research objective of privacy preserving data mining (PPDM) is how to protect the sensitive information or private knowledge from leaking in the mining process, meanwhile obtain the accurate results of data mining. A PPDM is focus on protecting the sensitive data such as id, name, address and other sensitive information. Many privacy preserving techniques are using some form of transformation to achieve privacy. Privacy preserving is mainly focused on data distortion, data reconstruction and data encryption technology. The implementation of PPDM techniques has become the demand of the moment[2]. The goal of this paper is to present the review on privacy preserving techniques which is very helpful while mining process over large data sets with reasonable efficiency and preserve security.A wide variety of sources holds individuals private data such as banks(personal information name, birth-date , PAN police records(name, address, birth marks, physical appearance), airports (passport number departure, destination, duration, age and gender) expenditure data while purchasing or bank transaction. In most of countries sharing individual private data or exposing confidential information is against the law to share or make such information publicly available to others[3].

2. LITERATURE SURVEY

Variety of approaches has been proposed in the area of privacy preserving data mining. Some of the important approaches include cryptographic approach, heuristic approach and reconstruction based approach. The concept of the heuristic approach method is the way to hide sensitive rules which is used to be mined from the dataset while maximizing the outcome of the released data. The second approach is Cryptography based method, This approach has been developed to solve the problem such as SMC: If Two or more parties want to perform a computation based on their private inputs, but party is unwilling to disclose its own output to any other else. Such problem is referred to as the Secure Multiparty Computation (SMC) problem. Next approach is reconstruction based method ,In this approach they first used some methods to distort or twist the values of the original data and then release these twisted data[4[5]. Another important approach is the Access control based approach. It was built over existing technologies was proposed called Multi- relational association rules (MRAR). This model has three layers those are Authenticator, checker and the database server. MRAR is the type of policy where the users are associated to mining levels which is mandatory access control. Disadvantage of MRAR is that it is not always possible to assign sensitivity levels to data in case level contains another level[6].

Anonymization Method: This method is used to protect user's identities while releasing micro data. The k anonymity protects against identity disclosure. But it does not provide sufficient protection against field's disclosure and original data can be reconstructed.

Perturbation Method: Independent operation is performed on the different fields by this method. This method does not reconstruct the original data values, but only distribution, new algorithms have been developed which uses these reconstructed distributions to carry out mining of the data available.

Randomized Response Method: This is very simple technique which can be easily implemented at the time of data collection. It is useful technique for hiding individual data in PPDM. This method results in high information loss. It is not suitable for multiple attribute databases.

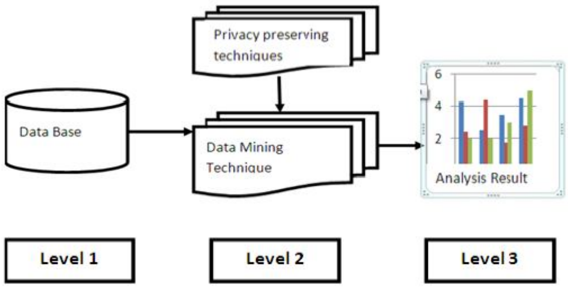

Fig 1 Architecture of Data Mining

There are various techniques for privacy preserving data mining.

Privacy preserving techniques can be classified based on following characteristics:

- Data Mining Scenario
- Data Mining Tasks
- Data Distribution
- Data Types
- Privacy Definition
- Protection Method

We describe these classifications characteristics as follows:

1) **Data Mining Scenario**: There are basically two major data mining scenario present. In the first one organization release their data sets for data mining and allowing unrestricted access to it. Data modification is used to achieve the privacy in this scenario. In the second one organization do not release their data sets but still allow data mining tasks. Cryptographic techniques are basically used for privacy preserving[7].

2) **Data Mining Task:** Data set contains various patterns. These patterns are taken out through different types of data mining tasks like classification, association rule mining, outlier analysis, clustering and evolution analysis [8]. Basically, all privacy preserving techniques should maintain data quality to support all possible data mining tasks and statistical analysis but it usually maintain data quality to support only a group of data mining tasks. Basis on that task we categorize the privacy preserving techniques.

3) **Data Distribution**: Data sets used for data mining can be either distributed or centralized. It is not depending on the physical location where data is stored but to the availability/ownership of data. The centralized data set is owned by a single party. It is either available at computational site or it can be sent to the site. However, distributed data set is shared between two or more parties which do not necessarily trust each other private data but interested to perform data mining on joint data. The data set can be heterogeneous means vertically partitioned where each party owns the same set of attributes but different subset of attributes. Alternatively it can be homogeneous means horizontally partitioned where each party owns the same set of attributes but different subset of records. In Fig. 2 we shows the classification based on distribution.

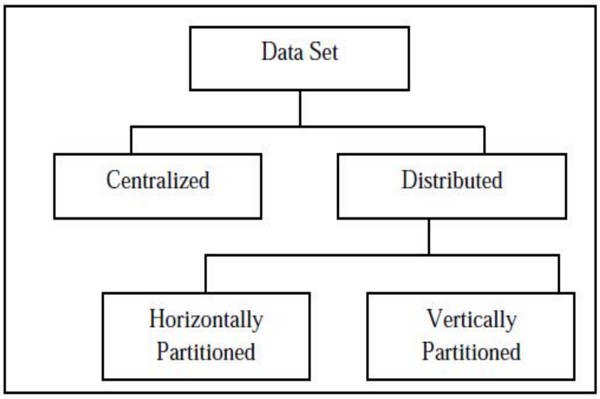

Fig 2 Classification of Different Dataset Based on Distribution

1) Data Types: There are basically two attributes in data set: Numerical and Categorical. Boolean data are the special case of categorical data which takes two possible values 0 and 1. Categorical values lack natural ordering in them. This is the basic difference between categorical and numerical values and its force the privacy preservation technique to take different approaches for them.

2) Privacy Definition: The definitions of privacy are different in different context. In some scenario individuals data values are private, whereas in other scenario certain association or classification rules are private.. Depend on the privacy definition we work on different privacy preserving techniques.

3) Protection Methods: Privacy in data mining is protected through different methods such as data modification and secure multiparty computation (SMC). On the basis of protection method we can also categorize the privacy preserving techniques. The classification is shown in Fig. 3.

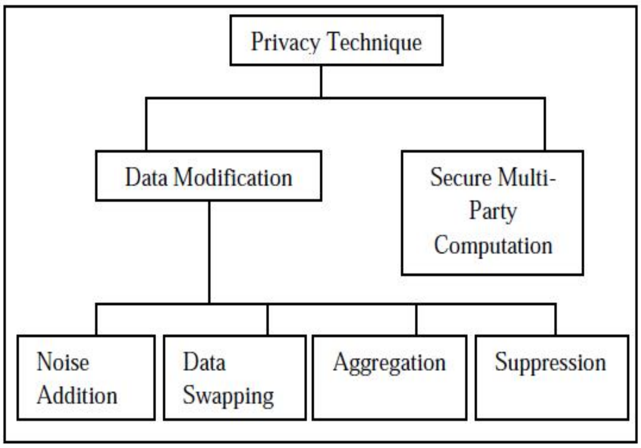

Fig 3 A Classification of Privacy Preserving Techniques

TECHNIQUES OF PRIVACY PRESERVING

A.Data Modification

Existing privacy preserving techniques method for centralized databases can be categorized in three main groups based on the approaches they take, such as query restriction, output perturbation and data modification [7]. From all these techniques data modification is a straightforward technique to

implement. In data modification before the release of a dataset for various data mining tasks and analysis, it modifies the data set for protection of individual privacy while the quality of released data remains high. After this modification of the data we can use any off the shelf software such as See to manage or analyze the data. It's not with the case of query restriction and output perturbation. The simplicity of this technique made it attractive and widely used in the context of statistical database in data mining. There are number of ways of doing data modification such as suppression, swapping, aggregation and noise addition. The basic idea of these techniques is given below[9].

1. Data Swapping: Data swapping technique were first introduced by Dalenius and Reiss in 1982, for categorical values modification in the context of secure statistical databases [10]. The main idea of the method was it keeps all original value in the data set, while at the same time makes the record re-identification very complex. This method actually replace the original data set by another one where some original values belonging to a sensitive attributes are exchanged between them. An introduction to existing data swapping technique can be found in [9], [10].

2. Aggregation: Aggregation is also known as generalization or global recording. It is used for protecting an individual privacy in a released data set by perturbing the original data set before its releasing. Aggregation change k no. of records of a data by representative records. The value of an attribute in such a representative record is generally derived by taking the average of all values, for the attributes, belonging to the records that are replaced. Another method of aggregation or generalization is transformation of attribute values.

3) **Suppression**: In this technique sensitive data value are deleted or suppressed prior to the release of a micro data. Suppression is used to protect an individual privacy from intruders attempt to accurately predict a suppressed value. To predict a sensitive value an intruder can use various approaches. An important issue in suppression is to minimize the information loss by minimizing the number of values suppressed. For some applications, such as medical, suppression is preferred mainly over noise addition in order to reduce the chance of having misleading pattern in perturbed data set. Suppression technique is also been used for association and classification rule confusion [12], [13].

C**. Secure Multiparty Computation (SMC)**

Secure Multi-party Computation (SMC) technique encrypts the data sets, while still allowing data mining operations. SMC techniques are not supposed to disclose any new information other than the final result of the computation to a participating party. These techniques are typically based on cryptographic protocols and are applied to distributed data sets. Parties involved in a distributed data mining encrypt their data and send to others parties. These encrypted data are used to compute the aggregate data, belonging to the joint data set, which is used for data mining purpose. Secure Multipart Computation was originally introduced by Yao in 1982 [17]. Basically, SMC is supposed to reveal to a party just the result of the computation and the data owned by the party. There are various SMC algorithms developed. Most of the algorithms make use of some primitive computations such as secure sum, secure set union, secure size of set intersection and secure scalar product.

D **Non-cooperative Computation**

Non Cooperative Computation, NCC is a game theoretic concept and specifically is couched in terms of mechanism design. In NCC the agents communicate their input (truthfully or not) to a trusted third party (center), which functions a commonly-known computation and distributes the results to the agents. The intersections of computer science and game theory have been studied extensively and are the recent research issue. So one of the closely related issue to our work is the algorithmic mechanism design and non-cooperative computation. The field of algorithmic mechanism design tries to explore how private preferences of many parties could be combined to find a global and socially optimal solution [15] (e.g., Vickrey-Groves-Clarke mechanisms [15]).NCC is a very broad framework. The technical results we give are specific to the setting in which each agent has a primary interest in computing the function and a secondary interest in preventing the others from computing it (properties called correctness and exclusivity).

The Incentive Compatible Model has been developed that to provocation the participating parties provide truthful input data [13]. The incentive compatible privacy preserving model has to interact with the participating parties to verify the transaction making use of the user's knowledge. The E-Shopping is a service oriented application, which provides a user interaction interface that provides more security for individual details transformation compared with the other privacy preserving models.

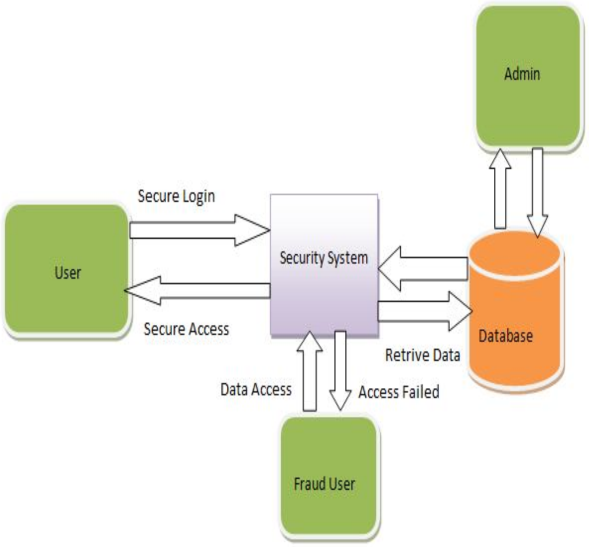

Figure 4 Privacy Preserving System Architecture

The figure 4 denotes the architecture of the privacy preserving system model. The architecture of a privacy preserving system gives the detailed explanation about the process of the security system in which it allows only the authorized person not others. Suppose, if any fraud user is trying to access the data security system will not allow the user and also the access will be denied for the particular user. Then the appropriate data are retrieved from the database according to the request given by the user.

Secure Code Computation Process

SCCP is providing incentive compatible secret code question for the NCC Model. The computation process theorem consists of following steps:

Step1: Select two fields from customer details from bank database as input for secure code computation process.

Step2: Here first field is constant and another one field is other information of customer details. For example (one field is username that is constant , another one field is other information like dob , accno , email id ,etc..).

Step3: Apply vertical partition on the first field data and attaching second field in the middle of partition data using Secure Sum Process technique.

Flowchart for Incentive Compatible System Model

The Incentive compatible model is a web service model for online shopping, online shopping and many online applications. The figure 5 represents the flowchart for incentive compatible system model. The system is used to protect the user details in data sharing such as payment processing. This model is built using data mining techniques such as association rules, Horizontal partitioning of the table and Vertical partition of the data. The system considers a distributed database like bank database that is used to construct the incentive marking. The incentive data are used to check the user knowledge that is the processing user is correct person or not.

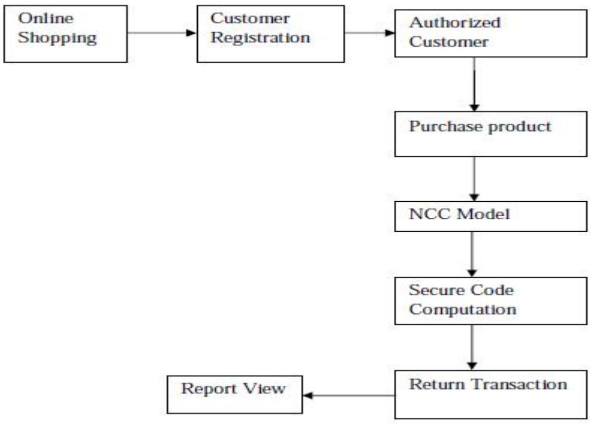

Fig 5 Flowchart for Incentive Compatible System Model

Methodology

Vertical partitioning (Heterogeneous Distribution) of data implies that though different sites gather information about the same set of entities, they collect different feature sets.

Horizontal Partitioning

In Horizontal partitioning (Homogeneous Distribution), different sites collect the same set of information, but about different entities. An example of that would be grocery shopping data collected by different supermarkets (also known as market-basket data in the data mining literature).

Java Mail API:

The Java Mail API provides a platform-independent and protocol-independent framework to build mail and messaging applications. The Java Mail API provides a set of abstract classes defining objects that comprise a mail system. It is an optional package (standard extension) for reading, composing, and sending electronic messages.

Java Mail provides elements that are used to construct an interface to a messaging system, including system components and interfaces. While this specification does not define any specific implementation, Java Mail does include several classes that implement RFC822 and MIME Internet messaging standards. These classes are delivered as part of the Java Mail class package.

Following are some of the protocols supported in Java Mail API:

- **SMTP:** Acronym for Simple Mail Transfer Protocol. It provides a mechanism to deliver email.
- **POP:** Acronym for Post Office Protocol. POP is the mechanism most people on the Internet use to get their mail. It defines support for a single mailbox for each user. RFC 1939 defines this protocol.
- **IMAP:** Acronym for Internet Message Access Protocol. It is an advanced protocol for receiving messages. It provides support for multiple mailbox for each user, in addition to, mailbox can be shared by multiple users. It is defined in RFC 2060.

SMPT server

To send emails, you must have SMTP server that is responsible to send mails. You can use one of the following techniques to get the SMTP server.

•Install and use any SMTP server such as Postfix server (for Ubuntu), Apache James server (Java Apache Mail Enterprise Server)etc.

•Use the SMTP server provided by the host provider for eg: free SMTP provide by Jango SMTP site is relay.jangosmtp.net

•Use the SMTP Server provided by companies e.g. gmail, yahoo, etc.

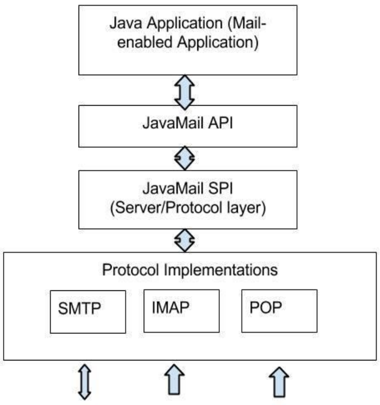
Java Application (Mail-enabled Application)
JavaMail API
JavaMail SPI (Server/Protocol layer)
Protocol Implementations
SMTP
IMAP
POP

Fig 6 Protocol Implementations

3.CONCLUSION

In this paper we survey on privacy preserving data mining and briefly review the techniques Data Modification and Secure Multiparty Computation. study on an incentive compatible privacy preserving data analysis technique. The incentive compatible privacy-preserving data analysis technique has been developed to motivate the participating parties to provide truthful inputs. The privacy preserving data analysis task that provides a new model called Incentive Compatible The main advantage of this model is that to reduce the number of False Positive transactions. It tries to find any anomalies transaction based on the data analysis model. In proposed system we are using symmetric key cipher algorithm to increase the security measures. In this system it will encrypt the security code so if any hacker check for the key it will be in encrypted format so no one can hack the password. And to send the keys to the mail we are using Java Mail API directly to communicate with the Gmail server**.**

REFERENCES

[1] Kirubhakar Gurusamy, Venkatesh Chakrapani, “ An assessment of Identity Security in Data Mining”, International Journal of Science and Modern Engineering (IJISME), Vol. 1, No. 7, PP. 29-31, 2013.

[2] Manish Sharma, Atul Chaudhary, Manish Mathuria, Shalini Chaudhary, “ A Review Study on the Privacy Preserving Data Mining Techniques and Approaches” , International Journal of Computer Science and Telecommunications, Vol. 4, No. 9,PP.42-46, 2013.

[3] Ekta Chauhan, Sonia Vatta,“ Review of Privacy Preserving in Data Mining Using Homomorphic Encryption”, International Journal of Advanced Research in Computer Science and Software Engineering , Vol. 3, No. 5, PP. 1431-1433, 2013.

[4] Mohammad Reza Keyvanpour, Somayyeh Seifi Moradi,“Classification and Evaluation the Privacy Preserving Data Mining Techniques by using a Data Modification based Framework”,International Journal on Computer Science and Engineering (IJCSE), Vol. 3, No. 2, PP. 862-870, 2011.

[5] K. Sathiyapriya, Dr. G. Sudha Sadasivam,“A SURVEY ON PRIVACY PRESERVING ASSOCIATION RULE MINING”, International Journal of Data Mining and Knowledge Management Process, Vol.3, No.2, PP. 119-131, 2013.

[6] Murat Kantarcioglu, Chris Clifton, “Privacy-Preserving Distributed Mining Of Association Rules On Horizontally Partitioned Data”, Knowledge And Data Engineering,IEEE Transactions,Vol. 16, No. 9, 2004.

[7] BENJAMIN C. M. FUNG, KE WANG, RUI CHEN, PHILIP S. YU,“Privacy-Preserving Data Publishing: A Survey of Recent Developments ” , ACM Computing Surveys, Vol. 42,No. 4, 2010. Supriya S. Borhade et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 1560-1563.

[8] Tamir Tassa, “Secure Mining Of Association Rules In Horizontally Distributed Databases”,IEEE Transactions On Knowledge and Data Engineering , Vol. 1, No. 99, PP. 1-14, 2013.

[9] Murat Kantarcioglu , Wei Jiang, “Incentive Compatible Privacy- Preserving Data Analysis”, IEEE Transactions On Knowledge And Data Engineering, Vol. 25, No. 6, PP. 1333-1335, 2013.

[10] Sowmyarani C N, Dr. G N Srinivasan, “Survey on Recent Developments in Privacy Preserving Models”, International Journal of Computer Applications, Vol. 38, No.9, PP. 18-22,2012.

[11] Bin Zhou, Yi Han, Jian Pei, Bin Jiang, YufeiTao,YanJia,“Continuous Privacy Preserving Publishing of Data Streams”, EDBT , P. 24,S26, 2009.

[12] Madhan Subramaniam, Senthil R,“ An Analysis on Preservation of Privacy in Data Mining”,(IJCSE) International Journal on Computer Science and Engineering,Vol. 02, No. 05, PP.1696-1699, 2010.

[13] Dr.K.P.Thooyamani, Dr.V.khanaa, “Privacy-Preserving Updates to Anonymous and Confidential Database”,International Journal of Data Mining Techniques and Applications, Vol. 01,PP. 2278-2419,2012.

[14] W. Jiang and B.K. Samanthula, “N-Gram Based Secure Similar Document Detection,” Proc. 25th Ann. WG 11.3 Conf. Data and Applications Security and Privacy (DBSec ’11), July 2011.

[15] M. Kantarcioglu and O. Kardes, “Privacy-Preserving Data Mining in the Malicious Model,” Int’l J. Information and Computer Security,vol. 2, pp. 353-375, Jan. 2009.

[16] M. Kantarcioglu and R. Nix, “Incentive Compatible Distributed Data Mining,” Proc. IEEE Int’l Conf. Soc. Computing/IEEE Int’l Conf. Privacy, Security, Risk and Trust, pp. 735-742, 2010.

[17] M. Kantarco_glu and C. Clifton, “Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data.