A Review Paper on "Privacy Preservation of Data Mining Using Randomization Response Technique"

[1]Miss Acharya Dimple D., [2]Prof. Thacker Smit M

Research Scholar, Assistant Professor
Department of Computer Engineering,
HJD-ITER.
Gujarat Technological University

***Abstract*-The area of privacy do harm to rapid advances in recent years because of become greater in the ability to store data.Privacy preserving permit sharing of privacy sensitive data for a detailed examination of elements purposes so it is very popular technique. So, people have ready to share their data. It is used for protecting the privacy of the risky and sensitive data of data mining.In Randomization Response Technique random noise added to the original data to preserve privacy.The main goal of privacy preserving data mining is to develop algorithms for modifying the original data and securing the information to be apply to the wrong purpose, so that the private data and private knowledge remain as it is after mining process.**

***Keywords- Data Mining*, Privacy, Randomization Response Technique**

I. INTRODUCTION

Data mining refers to extracting or "mining" knowledge from large amounts of data[1].

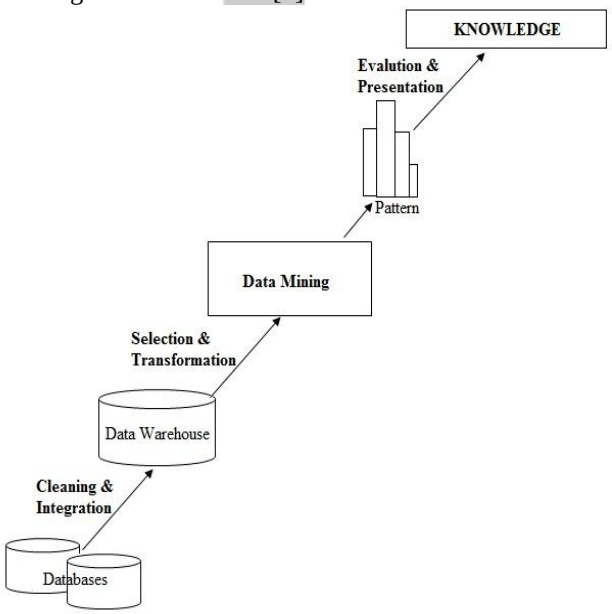

Figure 1 Data mining process [1]

Data mining consist of a repeat sequence of following step [1]:

1) Data cleaning (to remove noise and inconsistent data)

2) Data integration (where multiple data sources may be combined)

3) Data selection (where data relevant to the analysis task are retrieved fromthe database)

4) Data transformation (where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance)

5) Data mining (an essential process where intelligent methods are applied in order to extract data patterns)

6) Pattern evaluation (to identify the truly interesting patterns representing knowledge based on some interestingness measures

7) Knowledge presentation (where visualization and knowledge representation techniques are used to present the mined knowledge to the user)

A data warehouse is a repository of information collected from multiple sources, stored under a unified schema, and that usually resides at a single site [1].

Considering the rapid development in technology such as internet, data storage, data processing methods, we need to pay equal attention towards privacy preserving data mining. For secured public system we not only need to take care about the trimming of data but also the data inference. There are number of privacy preserving data mining methods have been proposed [2]. This paper gives a review of privacy preserving data mining using Randomization Response Technique and analyses the representative methods for privacy preserving data mining, as well as points out their advantages and disadvantages.

Privacy is an important issue in data mining and knowledgediscovery.The growth of the Internet makes it easy to perform data collection on a large scale. Collecting useful data for datamining is a great challenge: on one hand, such data collectionneeds to preserve customers' privacy; on the other hand,the collected data should allow one to use data mining to"mine" useful knowledge [3].

There are two types of privacy

1. B2B (Business to Business)
2. B2C(Business to Customer)

Business to Customer, which provides privacy at the point of data collection at the user site. Different people needs different level of privacy.

The rest of this paper is organized as follows. In section 2, introduction of various techniques of privacy preservation. In section 3, Details of randomization technique for privacy preserving data mining. In section 4, Method and major data

mining task which used in Randomization Response Technique. In section 5 contains the conclusions.

II. CLASSIFICATION OF PRIVACY PRESERVATION TECHNIQUES

The aim of privacy preserving data mining is to developdata mining methods without increasing the loss of data whichused to produce methods.The topic of privacy preserving data mining has been covering a large area of data mining community in recent years. A number of effective methods for privacy preserving data mining proposed earlier. Most methods use some form of transformation on the original data in order to perform the privacy preservation. The transformed dataset is made available for mining and must meet privacy requirements without losing the benefit of mining [4].We classifythem into the following categories:

A. The Randomization Method

Randomization Response (RR) techniques were developed inthe statistics community for the purpose of protecting surveyor'sprivacy.

Randomized Response technique was first introduced by Warner [5] in 1965 as a technique to solve the following survey problem: to estimate the percentage of people in a population that has attribute A, queries are sent to a group of people. Since the attribute A is related to some confidential aspects of human life, respondents may decide not to reply at all or to reply with incorrect answers. Two models (Related-Question Model and Unrelated-Question Model) have been proposed to solve this survey problem.

Randomization Response Technique will explained in section 3.

B. Anonymization Method

With the rapid growth in database, networking, and computing technologies, a large amount ofpersonal data can be integrated and analyzed digitally, leading to an increased use of data mining toolsto conclude trends and patterns.

Anonymization method aims at making the individual record be indistinguishable among a group records by using techniques of generalization and suppression [6]. Some of popular techniques such as k-anonymity, l-diversity and t-closeness, M-invariance, Personalized anonymity models. In order to preserve privacy, Sweeney proposed k-anonymity using generalization and suppression, Generalization involves replacing a value with a lessspecific but semantically reliable value. Forexample, the age of a person could be generalized to a rangesuch as youth, middle age and adult without specifyingappropriately, so as to reduce the risk of identification.Suppression involves reduction in exactness of applicationsand it doesn't liberate any information.By using this method itreduces the risk of detecting exact information [4].By using this method itreduces the risk of detecting exact information.

Table 1. Original patterns table [4]

S.No	Zip Code	Age	Disease
1	546177	39	Heart Disease
2	546102	32	Heart Disease
3	546178	37	Heart Disease
4	549105	53	Gastritis
5	549209	62	Heart Disease
6	549206	57	Cancer
7	546205	40	Heart Disease
8	546273	46	Cancer
9	546207	42	Cancer

Table 2. A 3-Anonymous version of Table1 [4]

S.No	Zip Code	Age	Disease
1	546***	3*	Heart Disease
2	546***	3*	Heart Disease
3	546***	3*	Heart Disease
4	549***	>=50	Gastritis
5	549***	>=50	Heart Disease
6	549***	>=50	Cancer
7	546***	4*	Heart Disease
8	546***	4*	Cancer
9	546***	4*	Cancer

C. Distributed Privacy Preservation Method

The key goal in most distributed methods for privacy preserving data miningis to allow calculate of useful aggregate such analyzed data over the entire dataset without compromising the privacy of the individual data sets within the differentparticipants.

The participants may wish to collaborate in obtaining aggregate results, but may not fully trust each other in terms of the distribution of their own data sets. For this purpose, the data sets may either be horizontally partitioned or be vertically partitioned. In horizontally partitioned data sets, the individual records are spread out across multiple entities, each of which have the same set of attributes. In vertical partitioning, the individual entities may have different attributes (or views) of the same set of records. Both kinds of partitioning pose different challenges to the problem of distributed privacy preserving data mining[7].

There are two classic settings for privacy preserving data mining. In first setting, the data is divided among two or more different parties; the aim being to run a data mining algorithm on the union of the parties' databases by not allowing any party to view another individual's private data. In the second setting, some statistical data that is to be released may contain confidential data; hence, it is modified first so that (a) the data does not compromise anyone's privacy, and (b) it is still possible to obtain meaningful results by running data mining algorithms on the modified data set[4].

D. Cryptographic Technique

This technique became hugely popular for two main reasons: Firstly, cryptography offers a well-defined model for privacy, which includes methodologies for proving and determine the quantity. Secondly, there exists a vast toolset of cryptographic algorithms and constructs to implement privacy-preserving data mining algorithms [8].

The aim of secure multiparty computation is to enable parties to carry out distributed computing tasks in a secure manner [9].

E. Soft computing Technique

Soft computing techniques include fuzzy logic, neural networks, genetic algorithms, and rough sets. Fuzzy sets provide a natural framework for the process in dealing with uncertainty.It makes it possible to model imprecise and qualitative knowledge as well as the transmission and handling of uncertainty at various stages. Neural Networks are widely used for classification and rule generation. Genetic algorithms are adaptive, robust, efficient and global search methods, suitable in situations where the search space is large. Rough set is a mathematical tool for managing uncertainty that arises from indiscernibility between objects in a set [8].

F. Encryption Technique

Encryption technique solves the problem of dataprivacy easily. Use of encryption techniques makes easy toconduct data mining among mutual untrusted parties, oreven between competitors. In distributed data miningencryption technique is used due to its privacy concern.Neglecting the efficiency of Encryption, it is used in bothapproaches of distributed data mining i.e. horizontallypartitioned data and that on vertically partitioned data [2].

III. Randomization Response Technique

Randomization response techniques are mainly designed for protecting privacy in a distributed database by using encryption techniques.The main goal of privacy preserving data mining is privacy preserving data mining algorithm should not only prevent the discovery of sensible information but also should be resistant to the various data mining techniques [10].

Figure 2 Model of Randomization [10]

In the randomized response technique [10], consider the data sets

$$I = \{I1, I2, I3\ldots\ldots In\}$$

And the random number or noise part are denoted by

$$R= \{R1, R2, R3 \ldots Rn\},$$

The new set of records are denoted by

$$I1+R1, I2+R2 \ldots.In+Rn$$

And after that take a partial support

$$Pij = \{Pi1, Pi2\ldots\ldots\ldots..Pin\}$$

So thatpartial support is

$$Pij=I+R$$

$$I=Pij-R$$

Randomization technique basically contains two steps to convert the data set into the original data sets. The first step is randomize the data and transmit their data into the data receiver. And second step the data receiver estimates their original data by applying the distribution reconstruction algorithm.

Algorithm [10]:

Step1:-Consider parties

$$P1, P2, P3\ldots\ldots\ldots\ldots Pn. (n>=4)$$

Step2:-Each party will generate their own random number

$$R1, R2\ldots\ldots\ldots RN$$

Step3:-Connect the parties in the ring (P1, P2, P3……PN) and let P1 is a protocol initiator.

Step4:-Let RC=N, and Pij=0 (RC is round counter and Pij is partial support)

Step5:-Partial support P1 site calculating by using following formula

$$Psij = Xij.support – Min support * |DB| + RN1 - RNn$$

Step6:-Site P2 computes the PSj for each item received the list using the formula

$$PSij= PSij + Xij. Support – minimum support * |DB| +Rn1-Rn (i-1)$$

Step7:-While RC! =0begin for j=1 to N do

begin for I=1 to N do begin each site will calculate their partial support Pij send to the next site that is neighbor to the current site.

Step8:-P1 exchange its position to P (j+1) mod N

end

RC=RC-1

end

Step9:-Party P1 allowance the result Pij

Step10:-End

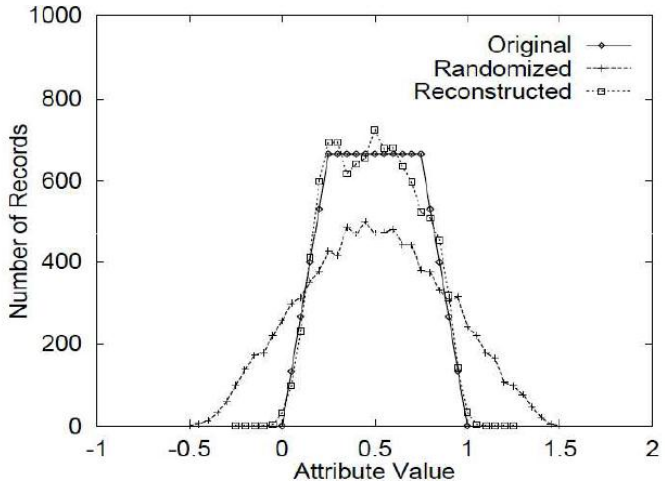

Figure 3 Variance of original result, Applying Randomization Response Technique on original result and then getting result and then reconstruction of result [12].

Two kinds of perturbation are possible with the randomization methods [10]:

- **Additive Perturbation:**In this case, randomized noiseis added to the data records. The overall datadistributions can be recovered from the randomizedrecords. Data mining and management algorithmsdesigned to work with these data distributions.
- **Multiplicative Perturbation:** In this case, either therandom projection or random rotation techniques isused in order to perturb the records.

The Randomized Response (RR) was firstly proposed by Warner [4]. The RR scheme is a technique originally developed such analyzed data community to collect sensitive information from individuals in such a way that survey interviewers and those who process the data do not know which of two alternative questions the respondent has answered.

Agrawal and Srikant [4] proposed a scheme for privacypreserving data mining using random perturbation anddiscussed how the reconstructed distributions may be used fordata mining. In their randomization scheme, random number isadded to the value of a sensitive attribute. For example, if x isthe value of a sensitive attribute, x+r rather than x will appearin the database, where r is a random noise drawn from certaindistribution. After Agrawal and Srikant, Kargupta et al. proposed a random matrix-based spectral filtering technique to obtain the original data from the perturbed data. Huang et al. further proposed two other data reconstruction methods: Principle compound analysis (PCA) and Maximum like hood estimation (MLE).

Randomization was first proposed in on numerical data. For a data distribution X. Value distortion is performed on data record xi by adding to it a random value r from a predefined random distribution R. All such values of xi + r are used to recreate the original distribution of data using a Bayesian reconstruction procedure to perform data mining operations mining operations do not always need individual data.Having the distribution of data will be sufficient to run mining operations.

Challenges in Randomization Response Technique [11]:

There are several challenges in Randomization Response Technique. Two significant challenges are:

- Develop algorithms for comparing the two (original and randomized) versions of the data. This can be considered as privacy metric.
- Develop algorithms for estimating the impact that certain modifications of the data have on the statistical significance of individual patterns obtained by data mining algorithms.

The first challenge said that the developing algorithm which shows variance of two result the original result and we get result after applying randomization response technique. And the second challenge [11] is to measure the impact the modification of data values has on a discovered pattern's Accuracy. Creating a universal method irrespective of approach to tackle these two Challenges are difficult. Based on the distortion method used, a unique approach is needed for measuring privacy and accuracy. A number of surveys conducted on users have led to the conclusion that multiple privacy levels are needed [11]. The logical basis existence, each user may have a different Level of sensitivity when answering a question or presenting some data. If the user is not satisfied with the privacy level offered by the system, the user may not answer the question. Offering the user with multiple levels of privacy would help the user to have freedom or authority to act according to one's judgment when needed without the system having to compromise on accuracy where unnecessary.

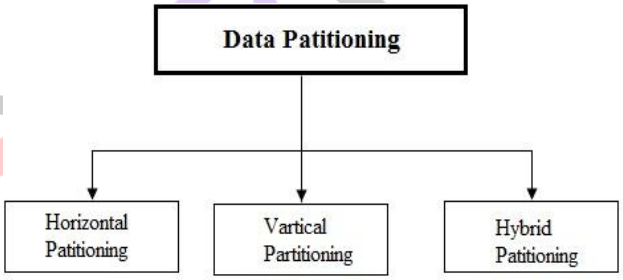

Figure 4 Database Partitioning [10]

Figure 3 described follow [10]:

- **Horizontal Partitioning:**Horizontal Partition divides the table into multiple tables that contains smaller number rows refers to these cases where dissimilar database records reside in dissimilar places.
- **Vertical Partitioning:** Vertical Partition divides the table into multiple tables that contains smaller number columns. Mainly two type of vertical partition prepared first is normalization in which redundant attributes are removed from the table and another one is row splitting in which the original table divides into table which contain fewer columns.
- **Hybrid Partitioning:**Hybrid Partition data is partition is first horizontal and then vertical partition or vertical partition and then horizontal partition.

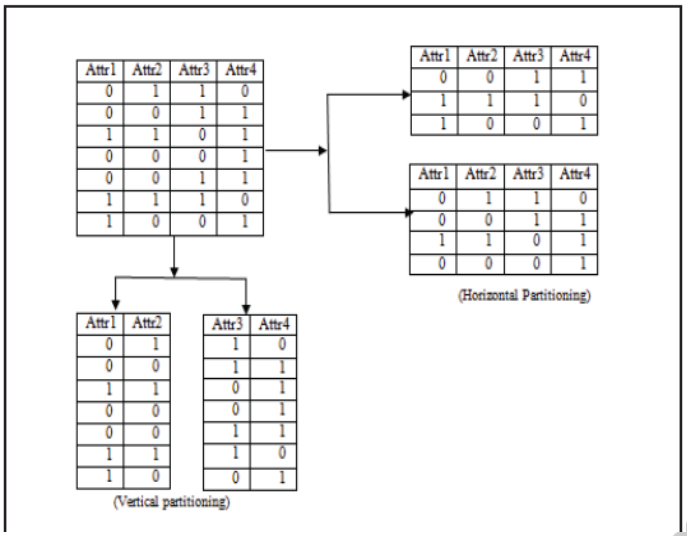

Figure 5 Partitioning of database (Tabular form) [10]

IV. METHOD ANDDATA MINING TASK WHICH USED IN RANDOMIZATION RESPONSE TECHNIQUE

In Randomization Response Technique two method are used [14]. They are: Adding Noise, Scrambling. Using these method major data mining task like classification, clustering, association rule mining and outlier detection done [13]. This data mining task is described follow:

1. **Classification:**Classification forms of data analysis that can be used to extract models describing important data classes or to predict future data trends. Such analysis can help provide us with a better understanding of the data at large. Classification predicts categorical (discrete, unordered) labels [1].

2. **Clustering:** Clustering is a process of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. The goal of clustering is to divide the data elements into groups of similar objects, where each group is referred to as a cluster, consisting of objects that are similar to one another and dissimilar to objects of other groups [14].

3. **Association Rule:** It aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories. Association rules are widely used in various areas such as telecommunication networks, market and risk management, inventory control etc. Association rule mining is to find out association rules that satisfy the predefined minimum support and confidence from a given database [15].

4. **Outlier Detection:** A database may contain data objects that do not comply with the general behavior or model of the data. These data objects are outliers. Outliers may be detected using statistical tests that assume a distribution or probabilitymodel for the data, or using distance measures where objects that are a substantialdistance from any other cluster are considered outliers. Rather than using statistical ordistance measures, deviation-based methods identify outliers by examining differencesin the main characteristics of objects in a group [1].

V. Conclusion

This paper focus on what is data mining and what is privacy for user which stores data and shares data with others. Mainly detailed about Randomization Response Technique which protecting privacy and briefly define major data mining task like classification, clustering, association rule mining and outlier detection.

REFERENCES

[1] Jiawei Han and Micheline Kamber, “Data Mining: Concepts and Techniques”, 2nd Ed.
[2] Sachin Janbandhu, Dr.S.M.Chaware, “Survey on Data Mining with Privacy Preservation”, IJCSIT, Vol. 5 (4), 2014, 5279-5283, ISSN: 0975-9646.
[3] Wenliang Du and Zhijun Zhan, “Using Randomized Response Techniques forPrivacy-Preserving Data Mining”, SIGKDD ’03, August 24-27, 2003, Washington, DC, USA.
[4] K.Saranya, K.Premalatha and S.S.Rajasekar, “A Survey on Privacy Preserving Data Mining”, 978-1-4788-7225-8/15/$31.00 ©2015 IEEE
[5] S. L. Warner. Randomized response: A survey techniquefor eliminating evasive answer bias. *The AmericanStatistical Association*, 60(309):63–69, March 1965.
[6] Pingshui WANG, “Survey on Privacy Preserving Data Mining”, International Journal of Digital Content Technology and its Applications,Volume 4, Number 9, December 2010.
[7] Charu C. Aggarwal and Philip S. Yu, “A General Survey of Privacy-Preserving Data Mining Models and Algorithms”, P.S. (Eds.) 2008, XXII, 514 p., Hardcover, ISBN: 978-0-387-70991-8.
[8] Brijal H. Patel and Ankur N. Shah, "Overview of Privacy preserving techniques and data accuracy", IJARCSMS, Volume 3, Issue 1, January 2015, ISSN: 2321-7782.
[9] S.Selva Rathna and Dr. T. Karthikeyan, “Survey on Recent Algorithms for Privacy Preserving Data mining”, IJCSIT, Vol. 6 (2) , 2015, 1835-1840, ISSN: 0975-9646.
[10] Jayanti Dansana, Raghvendra Kumar and Debadutta Dey, "PRIVACY PRESERVATION IN HORIZONTALLY PARTITIONED DATABASES USING RANDOMIZED RESPONSE TECHNIQUE", 978-1-4673-5758-6/13/$31.00 © 2013 IEEE
[11] Maiwand Khishki and Vijay Kumar, “ Research Paper on Randomization-based Privacy-Preserving Association Rule Mining”, IJARCSSE,Volume 5, Issue 6, June 2015 ISSN: 2277 128X.
[12] Agrawal, R. and Srikant, R. 2000. “Privacy-preserving data mining.” SIGMOD Rec. 29, 2 (Jun. 2000), 439-450.
[13] Jisha Jose Panackal and Dr Anitha S Pillai, "Privacy Preserving Data Mining: An Extensive Survey", © Association of Computer Electronics and Electrical Engineers, 2013, DOI: 03.AETS.2013.4.15.
[14] Apurva Juyal and Dr. O. P. Gupta, “ A Review on Clustering Techniques in Data Mining”, IJARCSSE, Volume 4, Issue 7, July 2014, ISSN: 2277 128X.
[15] Sotiris Kotsiantis, Dimitris Kanellopoulos, “Association Rules Mining: A Recent Overview”, GESTS International Transactions on Computer Science and Engineering, Vol.32 (1), 2006, pp. 71-82.