A Survey on Maintaining Privacy in Data Mining

Divya Sharma
Lecturer, Information Technology, Gandhinagar Institute of Technology, divya.sharma@git.org.in

Abstract—Data Mining is the process of discovering new patterns from large datasets. The goal is to extract knowledge from dataset in human understandable structure. Now a day we all are using internet lot, data processing technologies, privacy of data is a major issue in data mining .So Privacy Preserving Data Mining has become very popular and in high demand. A number of methods and techniques have been developed for privacy preserving data mining. This paper provides a wide survey of different privacy preserving data mining algorithms. I have discussed more about one of algorithm Randomization and also discussed merits and demerits of the same.

Index Terms – **Data mining, Privacy, Privacy-preserving data mining, Randomization, Data Swapping Randomization.**

I. Introduction

The main goal of data mining is to extract knowledge and new patterns from large datasets in human understandable structure. For data mining computations we have to first collect data without much concern about privacy of data. Because of privacy concerns some people are not giving right information. Therefore Privacy preserving data mining has becoming important field of research. In order to make a publicly system secure, we must ensure that not only private sensitive data have been trimmed out, but also that certain

Inference channels should be blocked as well with respect to privacy. A number of effective methods for privacy preserving data mining have been proposed [1]. This paper provides a wide survey of different privacy preserving data mining techniques, and points out their merits and demerits.

This paper is organized as follows. Section II, will introduce the classification of privacy preserving methods. Section III, will analyze the method of randomization for privacy preserving on the original data. Section IV, will discuss the swap randomization method. Randomization to protect privacy will be discussed in section V. And section VI will discuss the applications and section VII will discuss conclusion and future work.

II Classification of Privacy Preserving Methods and Techniques

A number of effective methods for privacy preserving data mining have been proposed [2].

Most methods use some form of privacy preservation .transformed dataset is made available for mining and must meet privacy requirements without losing the benefit of mining. We classify them into the following three categories:

A. The randomization method

The Randomization method is a popular method in current privacy preserving data mining. In which noise is added to the data in order to mask the attribute values of records [3]. The noise added is sufficiently large so that the individual values of the records can no longer be recovered. In general, randomization method aims at finding an appropriate balance between privacy preservation and knowledge discovery.

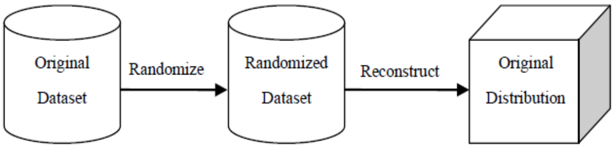


Figure 1. The Model of Randomization

In randomization method, data collection will be done in two steps. In first step, data providers randomize their data and transmit randomized data to data receiver. In second step, data receiver estimates original distribution of data using distribution reconstruction algorithm.

B. The anonymization method

Anonymization method aims at making the individual record be indistinguishable among a group records by using techniques of generalization and suppression. The representative anonymization method is k-anonymity. The motivating factor behind the k-anonymity approach is that many attributes in the data can often be considered quasi-identifiers which can be used in conjunction with public records in order to uniquely identify the records. Many advanced methods have been proposed, such as, p-sensitive k-anonymity, (a, k)-anonymity [4], l-diversity, t-closeness, M-invariance, Personalized anonymity, and so on. The anonymization method can ensure that the transformed data is true, but it also results in information loss in some extent.

C. The encryption method

Encryption method mainly resolves the problems that people jointly conduct mining tasks based on the private inputs they provide. These mining tasks could occur between mutual un-trusted parties, or even between competitors, therefore, protecting privacy becomes a primary concern in distributed data mining setting. There are two different distributed privacy preserving data mining approaches such as the method on horizontally partitioned data and that on vertically partitioned data. The encryption method can ensure that the transformed data is exact and secure, but it is much low efficient.

III The Randomization Method

In this section, I have discussed the randomization method for data privacy. The method of randomization can be described as follows:

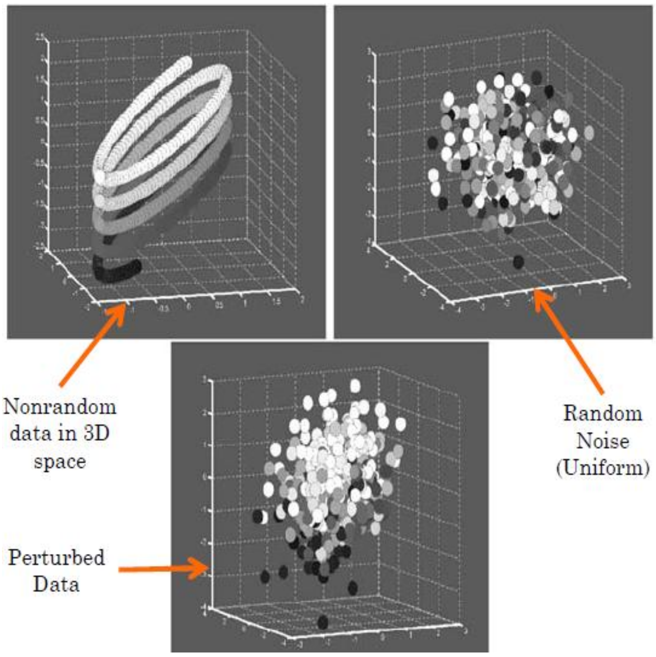
Consider a set of data records denoted by X = {$x_1 . . . x_N$}. For record $x_i \in$ X, we add a noise component which is drawn from the probability distribution Fr(R).These noise components are drawn independently, and are denoted $R_1 . . . R_N$. Thus, the new set of distorted records are denoted by $x_1 + r_1 . . . x_N + r_N$. This is denoted as Z_1…..Z_N

In general, it is assumed that the variance of the added noise is large enough, so that the original record values cannot be easily guessed from the distorted data. Thus, the original records cannot be recovered, but the distribution of the original records can be recovered. [5]

Thus, if X be the random variable denoting the data distribution for the original record, Y is the random variable describing the noise distribution, and Z is the random variable denoting the final record, we have:

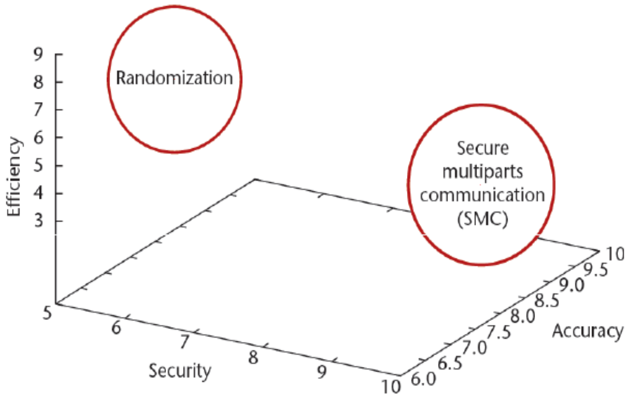
$$Z = X + R$$
$$X = Z – R$$

By subtracting R from the approximated distribution of Z, it is possible to approximate the original probability distribution X.



A Advantage

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. Noise is independent of data and does not need entire dataset for perturbation. The randomization method can be implemented at data collection time, and does not require the use of a trusted server containing all the original records in order to perform the anonymization process. The randomization approach has also been extended to other applications such as OLAP [6].
And it is much faster compared to SMC.



B Disadvantage

It treats all records equally irrespective of their local density. Therefore, outlier records are more susceptible to adversarial attacks as compared to records in more dense regions in the data.

C. Mulplicative Randomization

In this type of randomization, records are multiplied by random vectors. And then transform data so that inter-record distances are preserved approximately. These types of randomization can be applicable in Privacy-Preserving clustering and classification. Attacks can be known input-output or known sample attack. In known input-output attack, Attacker knows some linearly independent collection of records and their perturbed versions and in Known sample attack, Attacker has some independent samples from the original distribution.

D. Randomization for Association Rule Mining

This type of randomization is done through deletion and addition of items in transactions. Following steps are performed:

First we should select-a-size operator. Now assume transaction size = m and a probability distribution p[0], p[1], …, p[m], over {0, 1, …, m}.Given a transaction t of size m, generate randomized transaction t’ as: Select j at random from 0, .., m using above distribution Select j items from t (uniformly without replacement) and place in t’ For each item a not in t, place a in t’ with probability ρ, here p is the randomization level[7] .

IV DATA SWAPPING RANDOMIZATION

Noise addition or multiplication is not the only technique which can be used to perturb the data. A related method is that of data swapping, in which the values across different records are swapped in order to perform the privacy-preservation [8]. One advantage of this technique is that the lower order marginal totals of the data are completely preserved and are not perturbed at all. Therefore certain kinds of aggregate computations can be exactly performed without violating the privacy of the data. We note that this technique does not follow the general principle in randomization which allows the value of a record to be perturbed independently of the other records. Therefore, this technique can be used in combination with other frameworks such as k-anonymity, as long as the swapping process is designed to preserve the definitions of privacy for that model.

Swap randomization falls within the broad family of randomization testing methods. Given a metric of interest (e.g., the number of frequent item sets in the data), randomization testing techniques produce multiple random datasets and test the null hypothesis that the observed metric is likely to occur in the random data. If the metric of interest in the original data deviates significantly from the measurements on the random datasets, then we can reject the null hypothesis and assess the result as significant. The key characteristic of the randomization techniques is in the way that the random datasets are generated. Rather than assuming that the underlying data follows a given distribution and sampling from this distribution, randomization techniques randomly shuffle the given data to produce a random dataset. Shuffling is meant to preserve some of the structural properties of the dataset, for example, in a 0–1 matrix we may want to preserve the total number of 1’s in the dataset, or the number of 1’s in each column. In the case of swap randomization, the generated samples preserve both the column and row margins. This constraint can also be thought of.

$$\begin{array}{c|ccccc|} & & A & & B & \\ \hline & & \vdots & & \vdots & \\ u & \cdots & 1 & \cdots & 0 & \cdots \\ & & \vdots & & \vdots & \\ v & \cdots & 0 & \cdots & 1 & \cdots \\ & & \vdots & & \vdots & \\ \hline \end{array} \quad \Longleftrightarrow \quad \begin{array}{c|ccccc|} & & A & & B & \\ \hline & & \vdots & & \vdots & \\ u & \cdots & 0 & \cdots & 1 & \cdots \\ & & \vdots & & \vdots & \\ v & \cdots & 1 & \cdots & 0 & \cdots \\ & & \vdots & & \vdots & \\ \hline \end{array}$$

Assessing Data Mining Results via Swap Randomization

Swap randomization is an extension of traditional randomization methods. For instance, a chi-square test for assessing the significance of frequent item sets is a method based on studying the distribution of datasets where the column margins are fixed, but the row margins are allowed to vary. Similarly, methods that randomize the target value in prediction tasks keep the column margins fixed (e.g., Megiddo and Srikant [1998]), but impose no constraint on the row margins. These techniques are designed for assessing the significance of individual patterns or models, and are not appropriate for assessing complex results of data mining such as clustering or pattern sets. Swap randomization preserves both row and column margins, and takes into account the global structure of the dataset. A motivating example for why it is important to maintain both column and row margins is given in the next section

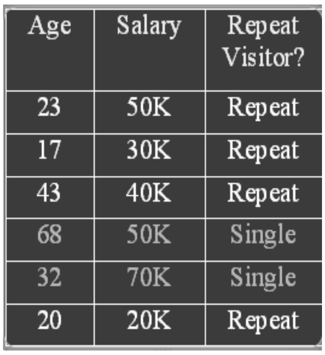
A. *Applications*

Swap randomization has been considered in various applications. An overview is presented in a survey paper by Cobb and Chen [2003]. A very useful discussion on using Markov chain models in statistical inference is Besag [2004], where the case of 0–1 data is used as an example. The problem of creating 0–1 datasets with given row and column margins is of theoretical interest in itself; see, among others Bez´akov´a et al. [2006] and Dyer [2003]. Closely related is the problem of generating contingency tables with fixed margins, which has been studied in statistics (such as Chen et al. [2005]). In general, a large body of research is devoted to randomization methods [Good 2000]

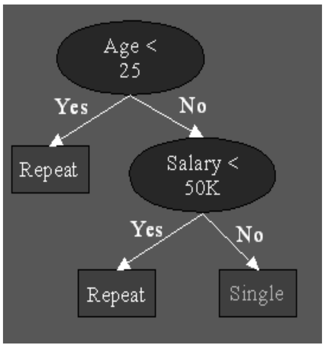
V RANDOMIZATION TO PROTECT PRIVACY

Return x+ r instead of x, where r is a random value drawn from a distribution. Uniform and Gaussian Reconstruction algorithm knows parameters of *r*'s distribution.

B. *Classification Example*



Age	Salary	Repeat Visitor?
23	50K	Repeat
17	30K	Repeat
43	40K	Repeat
68	50K	Single
32	70K	Single
20	20K	Repeat



Decision-Tree Classification:

```
Partition (Data S)
begin
        if (most points in S belong to same class)
        return;
    for each attribute A
        evaluate splits on attribute A;
    Use best split to partition S into S1 and S2;
    Partition (S1);
    Partition (S2);
End
```

C. *Training using Randomized Data*

In this we need to modify two key operations .Determining split point and partitioning data. When and how we should reconstruct distribution is primary question. First solution is to reconstruct using the whole data (globally) or reconstruct separately for each class. Second solution is to reconstruct once at the root node or at every node.

VI APPLICATIONS OF PRIVACY-PRESERVING DATA MINING

The problem of privacy-preserving data mining has numerous applications in homeland security, medical database mining, and customer transaction analysis. Some of these applications such as those involving bio-terrorism and medical database mining may intersect in scope. In this section, we will discuss a number of different applications of privacy-preserving data mining methods.

A. Medical Databases

The scrub system [9] was designed for de-identification of clinical notes and letters which typically occurs in the form of textual data. Clinical notes and letters are typically in the form of text which contains references to patients, family members, addresses, phone numbers or providers. Traditional techniques simply use a global search and replace procedure in order to provide privacy. However clinical notes often contain cryptic references in the form of abbreviations which may only be understood either by other providers or members of the same institution. Therefore traditional methods can identify no more than 30-60% of the identifying information in the data. The Scrub system uses numerous detection algorithms which compete in parallel to determine when a block of text corresponds to a name, address or a phone number. The Scrub System uses local knowledge sources which compete with one another based on the certainty of their findings. It has been shown in [9] that such a system is able to remove more than 99% of the identifying information from the data.

B. Bioterrorism Applications

In typical bioterrorism applications, we would like to analyze medical data for privacy-preserving data mining purposes. Often a biological agent such as anthrax produces symptoms which are similar to other common respiratory diseases such as the cough, cold and the flu. In the absence of prior knowledge of such an attack, health care’s providers may diagnose a patient affected by an anthrax attack of have symptoms from one of the more common respiratory diseases. The key is to quickly identify a true anthrax attack from a normal outbreak of a common respiratory disease, in many cases; an unusual number of such cases in a given locality may indicate a bio-terrorism attack. Therefore, in order to identify such attacks it is necessary to track incidences of these common diseases as well. Therefore, the corresponding data would need to be reported to public health agencies. However, the common respiratory diseases are not reportable diseases by law. The solution proposed in [10] is that of “selective revelation” which initially allows only limited access to the data. However, in the event of suspicious activity, it allows a “drill-down” into the underlying data. This provides more identifiable information in accordance with public health law.

C. Homeland Security Applications

A number of applications for homeland security are inherently intrusive because of the very nature of surveillance. In [11], a broad overview is provided on how privacy-preserving techniques may be used in order to deploy these applications effectively without violating user privacy. Some examples of such applications are as follows: Credential Validation Problem, Identity Theft, Web Camera Surveillance, Video-Surveillance, Watch List Problem

D. Genomic Privacy

Recent years have seen tremendous advances in the science of DNA sequencing and forensic analysis with the use of DNA. As result, the databases of collected DNA are growing very fast in the both the medical and law enforcement communities. DNA data is considered extremely sensitive, since it contains almost uniquely identifying information about an individual [12].

VI I CONCLUSION AND FUTURE WORK

In this paper, I have carried out a wide survey of the different approaches for privacy preserving data mining, and analyses the major algorithms available for randomization method and points out the existing drawback. While all the purposed methods are only approximate to our goal of privacy preservation, we need to further perfect those approaches or develop some efficient methods.

REFERENCES

[1] Han Jiawei, M. Kamber, *Data Mining: Concepts and Techniques*, Beijing: China Machine Press,pp.1-40,2006.

[2]]D. Agrawal and C. Aggarwal. *On the design and quantification of privacy preserving data mining algorithms*. In Proceedings of the 20th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, Santa Barbara, California, USA, May 21-23 2001.

[3] Agrawal R., Bayardo R., Faloutsos C., Kiernan J., Rantzau R., Srikant R.:*Auditing Compliance via a hippocratic database*. VLDB Conference, 2004..

[4] G. Loukides, J.H. Shao, *“An Efficient Clustering Algorithm for k-Anonymisation”*, International Journal of Computer Science And Technology,vol.23, no.2, pp.188-202, 2008.

[5] R. Agrawal, R. Srikant, *“Privacy-Preserving Data Mining”*, ACM SIGMOD Record, New York,vol.29, no.2, pp.439-450,2000.

[6] G Agrawal R., Srikant R., Thomas D. *Privacy-Preserving OLAP*. Proceedings of the ACM SIGMOD Conference, 2005.

[7] A. Evfimievski, R. Srikant, R. Agrawal, J. Gehrke, *“Privacy Preserving Mining of Association Rules”*, Information System, vol.29, no.4, pp.343-364,2004.

[8] Fienberg S., McIntyre J.: *Data Swapping: Variations on a Theme by Dalenius and Reiss*. Technical Report, National Institute of Statistical Sciences, 2003.

[9] Sweeney L.: *Replacing Personally Identifiable Information in Medical Records, the Scrub System*. Journal of the American Medical Informatics Association, 1996.

[10] Sweeney L.: *Privacy-Preserving Bio-terrorism Surveillance*. AAAI Spring Symposium, AI Technologies for Homeland Security, 2005.

[11] Sweeney L.: *Privacy Technologies for Homeland Security*. Testimony before the Privacy and Integrity Advisory Committee of the Deprtment of Homeland Security, Boston, MA, June 15, 2005.

[12] Malin B. *Why methods for genomic data privacy fail and what we can do to fix it*, AAAS Annual Meeting, Seattle, WA, 2004.