ISSN 2278-3083

Volume 3, No.4, July - August 2014

International Journal of Science and Applied Information Technology

Available Online at http://warse.org/pdfs/2014/ijsait01342014.pdf

# A Study Of Privacy Preserving Data Mining Techniques

**Ms.R.Kavitha[1], Prof.D.Vanathi[2]**

[1]M.E Student(CSE), Nandha Engineering College,Erode,TamilNadu,India,

E-mail: kavithamohan926@gmail.com

[2]Associate Professor(CSE), Nandha Engineering College,Erode,TamilNadu,India,

Email: vanathidina@yahoo.com

**ABSTRACT**

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

**Keywords:** Cryptography; Distributed Privacy Preserving; k-Anonymity; Privacy-preserving;Perturbation;

## 1. INTRODUCTION

The need of privacy preserving data mining has become more **significant** in recent years because of the increasing ability to store personal data about users and the increasing sophistication of data mining algorithm to leverage this information. A number of methods such as k-anonymity, classification, association rule mining,clustering have been recommended in recent years in order to perform privacy preserving data mining. Furthermore,the problem has been discussed in multiple communities such as the database, the statistical disclosure control(SDC) and the cryptography. Data mining techniques have been developed successfully to extracts knowledge in order to support a variety of domain areas marketing, weather forecasting, medical diagnosis, and national security. But it is still a challenge to mine some kinds of data without violating the data owners 'privacy .For example, how to mine patients 'private data is an ongoing problem in health care applications . As data mining become more pervasive, privacy concerns are increasing.

Commercial concerns are also concerned with the privacy issue. Most concerns gather details about individuals for their own particular needs. More often, different departments within an organization themselves may find it necessary to share information. In those cases, each organization or unit must ascertain that the privacy of the individual is not compromised or that sensitive business information is not divulged .Consider, for example, a government, or more specifically, one of its security branches interested in developing a system for determining, from passengers whose baggage has been checked, those who must be subjected to additional security measures. The data indicative of such need for further examination stems from a lot of sources like police records; airports; banks; general government statistics; and passenger information records that generally include personal information (such as name and passport number); demographic data (such as age and gender); flight information (such as departure, destination, and duration); and expenditure data (such as transfers, purchasing and bank transactions). In most countries, this information is considered as private and to avoid intentionally or unintentionally exposing confidential information about an individual, it is illegal to make such information freely available.

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

We begin with the uses of privacy preserving in section 2 followed by different techniques of privacy

preserving in section 3, we present and relate several important notions for this task, followed by distributed privacy preserving approaches in section 4 and describe some general goals of different approaches also. In section 5, the methods are compared and contrasted and finally we sum up with conclusion and future work in later sections.

## 2. 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.

### 2.1 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.

### 2.2 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.

### 2.3. 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.

### 2.4 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.

### 2.5. 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.

## 3 METHODS USED IN PRIVACY PRESERVING DATA MINING

### 3.1 Data Perturbation

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

$$Y = X + e;$$

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

$$Y = X * e$$

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

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

Similar to SAN and MN, UMA causes bias in the variance of the confidential attribute, as well as in the relationships between attributes. Multivariate microaggregation (MMA) [3], differs from UMA in that it groups data using a clustering technique that is based on a multi-dimensional distance measure. As a result, the relationships between attributes are expected to be better preserved. However, this benefit comes with a higher computational time complexity , which could be inefficient for large data sets. Xiao-Bai Li and Sumit Sarkar proposed a method, called perturbation trees [4] that uses a recursive partitioning technique to divide a data set into subsets that contain similar data. The partitioned data are perturbed using the subset average. Since the data are partitioned based on the joint properties of multiple confidential and nonconfidential attributes, the relationships between attributes are expected to be reasonably preserved. Further, the proposed method is computationally efficient.The algorithm is based on the kd-tree technique.

Li Liu *, Murat Kantarcioglu and Bhavani Thuraisingham[5] proposed an individually adaptable perturbation model, which enables the individuals to choose their own privacy levels. This method enables users to choose different privacy levels without significant data mining performance degradation. The effectiveness of the new approach is demonstrated by various experiments conducted on both synthetic and real-world data sets. Reconstruction is a very important step for the perturbation based PPDM approaches. It is found that when applied to real-world data sets reconstruction could be a problem. So some PPDM methods were proposed which skip this reconstruction step and compute the data mining results directly.

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

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

Distributed anonymous data perturbation method for privacy-preserving data mining is proposed by Feng LI†, Jin MA and Jian-hua LI[8]. Cryptography-based secure multiparty computation is a main approach for privacy preserving. However, it shows poor performance in large scale distributed systems. Meanwhile, data perturbation techniques are comparatively efficient but are mainly used in centralized privacy-preserving data mining (PPDM). In this paper, a light-weight anonymous data perturbation method is proposed for efficient privacy preserving in distributed data mining. The privacy constraints for data perturbation based PPDM are defined in a semi-honest distributed environment. Two protocols are proposed to address these constraints and protect data statistics and the randomization process against collusion attacks: the adaptive privacy-preserving summary protocol and the anonymous exchange protocol. Finally, a distributed data perturbation framework based on these protocols is proposed to realize distributed PPDM. Experiment results show that this approach achieves a high security level and is very efficient in a large scale distributed environment.

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

### 3.2. Condensation Approach

We introduce a condensation approach, [10] that constructs constrained clusters within the data set, and then generates pseudo-data from the statistics of those clusters .This technique **is also** referred to as condensation because uses condensed statistics of the clusters for generating pseudo-data. The constraints on the clusters are outlined in terms of the sizes of the clusters that are chosen **so as** to preserve k-anonymity. Since the approach works with pseudo-data instead of with modifications of original data, this helps in higher preservation of privacy than techniques that merely use modifications of the initial data. Moreover, the use of pseudo-data doesn’t require redesign of data mining algorithms, since are of same format as the original data. In contrast, once the data is constructed with the usage of generalizations or suppressions, we'd like to redesign data mining algorithms to work effectively with incomplete or partially certain data . It may also be effectively employed in situations with dynamic data updates like the data stream problem. we discuss a condensation approach

for data mining which uses a strategy that condenses the data into multiple clusters of predefined size, for every cluster, certain statistics are maintained.
Each cluster incorporates a size atleast k, that is referred to as the level of that privacy-preserving approach. The greater the level, the higher the amount of privacy. At the same time, a larger quantity of data is lost because of the condensation of a bigger number of records into one statistical group entity. We use the statistics from every cluster so as for generating the corresponding pseudo-data.

### 3.3 Method of Anonymization

When releasing micro data for analysis purposes, one must limit disclosure risks to a suitable level whereas maximizing data utility. To limit revealing risk, Samarati et al. [11]; Sweeney [12] introduced the *k*-anonymity privacy requirement, which needs every record in an anonymized table to be indistinguishable with at least *k* other records inside the dataset, with respect to a collection of quasi-identifier attributes. To achieve the *k*-anonymity requirement, they used each generalization and suppression for data anonymization. Unlike traditional privacy protection techniques like data swapping and adding noise, info in a *k*-anonymous table using generalization and suppression remains truthful. Particularly, a table is *k*- anonymous if the Ql values of every tuple are identical, to those of at least *k* other tuples. An example of 2-anonymous generalization is shown in table3. Even with the voter registration list which is shown in Table2, somebody can only infer that Ramu may be the person involved in the first 2 tuples of Table1, or equivalently, the real disease of Ramu is discovered only with probability 50%.

In general, *k* anonymity guarantees that an individual is related to his real tuple with a probability at most 1/*k*.

**Table-1** Microdata

| ID | Attributes | | | |
|---|---|---|---|---|
| | Age | Sex | Zip code | Disease |
| 1 | 36 | Male | 93261 | Headache |
| 2 | 34 | Male | 93234 | Headache |
| 3 | 41 | Male | 93867 | Fever |
| 4 | 49 | Female | 93849 | Cough |

**Table-2** Voter Registration List

| ID | Attributes | | | |
|---|---|---|---|---|
| | Name | Age | Sex | Zip code |
| 1 | Ramu | 36 | Male | 93261 |
| 2 | Mani | 34 | Male | 93234 |
| 3 | Ranu | 41 | Male | 93867 |
| 4 | Sonia | 49 | Female | 93849 |

**Table-3** A 2-Anonymous table

| ID | Age | Sex | Zipcode | Disease |
|---|---|---|---|---|
| | | | | |
| 1 | 3* | Male | 932** | Headache |
| 2 | 3* | Male | 932** | Headache |
| 3 | 4* | * | 938** | Fever |
| 4 | 4* | * | 938** | Cough |

**Table-4** Original patients table

| ID | Attributes | | |
|---|---|---|---|
| | Zip code | Age | Disease |
| 1 | 93261 | 36 | Headache |
| 2 | 93234 | 34 | Headache |
| 3 | 93867 | 41 | Fever |
| 4 | 93849 | 49 | Cough |

**Table-5** Anonymous versions of table

| ID | Attributes | | |
|---|---|---|---|
| | Zip code | Age | Disease |
| 1 | 932** | 3* | Headache |
| 2 | 932** | 3* | Headache |
| 3 | 938** | 4* | Fever |
| 4 | 938** | 4* | Cough |

While *k*-anonymity protects against identity disclosure, it does not give adequate protection against attribute disclosure. There are 2 attacks namely homogeneity attack and background knowledge attack. The limitations of the *k*-anonymity model stem from the 2 assumptions. First, it might be very hard for the owner of a database to determine which of the attributes are or are not available in external tables. The second limitation is that the *k*-anonymity model assumes a certain method of attack, whereas in real situations there's no reason why the attacker should not try with other methods. Example1. Table4 is the Original table, and Table5 is an anonymous version of it satisfying 2-anonymity.The Disease attribute is sensitive. Suppose Mani knows that Ranu is a 34 years old woman living in ZIP 93234 and Ranu's record is in the table. From Table5, Mani can conclude that Ranu corresponds to the first equivalence class, and thus should have fever. This is the homogeneity attack. For an example of the background knowledge attack, suppose that, by knowing Sonia's age and zip code, Mani can conclude that Sonia's corresponds to a record in the last equivalence class in Table5. More ever, suppose that Mani knows that Sonia has very low risk for cough. This background knowledge enables Mani to conclude that Sonia most likely has fever.

### 3.4 Cryptographic Technique

The another method in Privacy preserving data mining is cryptography. This branch became famous [13] 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 toolset of cryptographic algorithms .However, recent work [14] has pointed that cryptography does not defend the output of a computation. Instead, it prevents privacy leaks within the process of computation. Thus, it fails to produce a complete solution to the problem of privacy preserving data mining.

## 4. DISTRIBUTED PRIVACY PRESERVING DATA MINING

The key goal in most distributed methods for privacy-preserving data mining (PPDM) is to permit computation of useful aggregate statistics on the entire data set while not compromising the privacy of the individual data sets among the varoius participants. Thus, the participants may need to collaborate in getting aggregate results, but might not fully trust one another in terms of the distribution of their own data sets. For this reason, the data sets might either be horizontally partitioned or be vertically partitioned. In horizontally partitioned data sets, the individual records are spread out across multiple entities, each of that has the identical set of attributes. In vertical partitioning, the individual entities might have different attributes (or views) of the identical set of records. Both kinds of partitioning pose different challenges to the problem of distributed privacy-preserving data mining.

### 4.1. Distributed algorithms over horizontally partitioned data sets

In horizontally partitioned data sets, totally different sites contain different sets of records with identical (or highly overlapping) set of attributes that are used for mining purposes. Several of those techniques use specialised versions of the general strategies discussed in for various problems. The work in discusses the development of a popular decision tree induction method called ID3with the usage of approximations of the best splitting attributes. Subsequently, a range of classifiers are generalized to the problem of horizontally partitioned privacy preserving mining including the Naïve Bayes Classifier and the SVM Classifier with nonlinear kernels . An extreme solution for the horizontally partitioned case is discussed in [15], within which privacy preserving classification is performed in a fully distributed setting, where every customer has personal access to only their own record. A number of other data mining applications have been generalized to the problem of horizontally partitioned data sets [16]. These include the applications of association rule mining, clustering, and collaborative filtering.

### 4.2. Distributed algorithms over vertical partitioned data sets

For the vertically partitioned case, several primitive operations like computing the scalar product or the secure set size intersection may be useful in computing the results of data mining algorithms. As an example, the methods in [15] discuss the way to use scalar dot product computation for frequent item set counting. The method of counting can also be achieved by using the secure size of set intersection as discussed in [17]. Another technique for association rule mining uses the secure scalar product over the vertical bit representation of item set inclusion in transactions, so as to calculate the frequency of the corresponding item sets. This step is applied repeatedly within the framework of a roll up procedure of item set counting. It's been shown that this approach is quite effective in practice. The approach of vertically partitioned mining has been extended to a variety of data mining applications like decision trees, SVM Classification, Naïve Bayes Classifier, and kmeans clustering.

## 5. MERITS AND DEMERITS OF DIFFERENT TECHNIQUES OF PPDM

The following table (Table-6) lists different techniques used in PPDM along with their merits and demerits.

Table-6 Techniques in PPDM- merits and demerits

| Techniques of PPDM | Merits | Demerits |
|---|---|---|
| PERTURBATION | Independent treatment of the different attributes by the perturbation approach. It is a simple approach. | Under certain conditions , this approach is easier to breach and provides little privacy. |
| ANONYMIZATION | This method is used to protect respondents' identities while releasing truthful information. While *k*-anonymity protects against identity disclosure, it does not give sufficient protection against attribute disclosure. | There are two attacks namely homogeneity attack and background knowledge attack. Because the limitations of the *k*-anonymity model stem from the two assumptions. Firstly, it might be very hard for the owner of a database to decide which of the attributes are or are not available in external tables. The second limitation is that the *k*-anonymity model assumes a certain method of |

| | | |
|---|---|---|
| | | attack, while in real scenarios there is no reason why the attacker should not try other methods. |
| CONDENSATION | This approach works with pseudo-data rather than with modifications of original data, this helps in better preservation of privacy than techniques which simply use modifications of the original data. | The use of pseudo-data no longer necessitates the redesign of data mining algorithms, since they have the same format as the original data. |
| CRYPTOGRAPHIC | Cryptography offers a well-defined model for privacy, which includes methodologies for proving and quantifying it. There exists a huge toolset of cryptographic algorithms and constructs to implement privacy preserving data mining algorithms | This approach is especially difficult to scale when more than few parties are involved. Also, it does not address the question of whether the disclosure of the final data mining result may breach the privacy of individual records. And it is slow when the dataset involved is huge. |

## 6. CONCLUSION

With the development of data analysis and processing techniques, the privacy disclosure problem about individual or company is inevitably exposed when releasing or sharing data to mine useful decision information and knowledge, then give the birth to the research field on privacy preserving data mining. In this paper, we have dealt with different issues and present naïve privacy preserving methods to distribute ones and the methods for handling horizontally and vertically partitioned data. While all the proposed methods are only approximate to our goal of privacy preservation, we need to further perfect those approaches or develop some efficient methods. To address these issues, the following problems should be widely studied.

1. Though data perturbation approach provides good privacy, under certain conditions it is relatively easy to breach the privacy protection offered by the random perturbation based techniques.
2. Random data distortion preserves very little privacy.
3. In distributed privacy preserving data mining areas, efficiency is an essential issue. Efforts should be made to develop more efficient algorithms and achieve a balance between disclosure cost, computation cost and communication cost.
4. Privacy and accuracy don't go together; improving either usually incurs a cost in the other. How to apply various optimizations to achieve a trade-off should be deeply researched.
5. Side-effects are inevitable in data sanitization process. The question of reducing their negative impact on privacy preserving needs to be considered carefully. We additionally need to outline some metrics for measuring the side-effects resulting from data processing.

## REFERENCES


1. J.F. Traub, Y. Yemini, and H. Wozniakowski, "**The Statistical Security of a Statistical Database**," ACM Trans. Database system vol. 9, no. 4, pp. 672-679, 1984.
2. N.R. Adam and J.C. Wortmann, "**Security-Control Methods for Statistical Databases: A Comparative Study**," ACM Computing Surveys, vol. 21, no. 4, pp. 515-556, 1989.
3. J. Domingo-Ferrer and J.M. Mateo-Sanz, "**Practical Data-Oriented Microaggregation for Statistical Disclosure Control**," IEEE Trans.Knowledge and Data Eng., vol. 14, no. 1, pp. 189-201, 2002
4. Xiao-Bai Li and Sumit Sarkar, "**A Tree based Data Perturbation Approach for Privacy-Preserving Data Mining**",IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 18, NO. 9, SEPTEMBER 2006
5. Murat Kantarcioglu, Bhavani Thuraisingham ,"**The applicability of the perturbation based privacy preserving data mining for real-world data**", IEEE, Li Liu, 2007
6. Hillol Kargupta , Souptik Datta, Qi Wang and Krishnamoorthy Sivakumar , " **On the Privacy Preserving Properties of Random Data Perturbation Techniques**",ICDM2003.Third IEEE International conference on 19-22,Nov 2003

7. Yingpeng Sang, Hong Shen, and Hui Tian, "**Effective reconstruction of data perturbed by random projections**", IEEETRANSACTIONS ON COMPUTERS, VOL. 61, NO. 1, JANUARY 2012
8. Feng LI†, Jin MA, Jian-hua LI , "**Distributed anonymous data perturbation method for privacy preserving data mining**", Journal of Zhejiang University SCIENCE A
9. Hillol KarGupta and Souptik Datta,Qi Wang and Krishnamoorthy Sivakumar,"**Random Data Perturbation Techniques and Privacy Preserving data Mining**",IEEE International Conference on data Mining 2003
10. C.C. Aggarwal and P.S. Yu, "**A Condensation Approach to Privacy Preserving Data Mining,**" Proc. Ninth Int'l Conf. Extending Database Technology, pp. 183-199, 2004.
11. P.Samarati, "**Protecting respondent's privacy in micro data release",** IEEE Transaction on knowledge and Data Engineering,pp.010-027,2001.
12. L. Sweeney, "**k-anonymity: a model for protecting privacy** ", International Journal on Uncertainty, Fuzziness and Knowledge based Systems, pp. 557-570,2002.
13. Laur, H. Lipmaa, and T. Mieli' ainen,"**Cryptographically private support vector machines**". In Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 618-624,2006
14. Ke Wang, Benjamin C. M. Fung1 and Philip S. Yu, "**Template based privacy preservation in classification problems**", In ICDM, pp. 466-473,2005
15. Yang Z., Zhong S.Wright R." **Privacy-Preserving Classification of Customer Data without Loss of Accuracy**" SDM Conference, 603-610,2006
16. M. Kantarcioglu and C. Clifton, "**Privacy-preserving distributed mining of association rules on horizontally partitioned data**",2002
17. Clifton C. Kantarcioglou M., Lin X., Zhu M,"**Tools for privacy-preserving distributed data mining. ACM SIGKDD Explorations**" 4(2),2002