ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 7, October 2015

A Survey on Privacy Preserving Data Mining

Yashaswini

Dept. of Computer Application, AIMIT, St. Aloysius College (Autonomous), Mangalore, Karnataka, India

ABSTRACT: In recent years, advances in hardware technology have lead to an increase in the capability to store and record personal data about consumers and individuals. This has lead to concerns that the personal data may be misused for a variety of purposes. In order to alleviate these concerns, Privacy preserving data mining has become increasingly popular because it allows sharing of sensitive data for analysis purposes. Several techniques of privacy preserving data mining have been proposed in literature. In this paper, I have studied all these state of art techniques. A tabular comparison of work done by different authors is presented and merits and demerits of several techniques are pointed out.

KEYWORDS: privacy preserving, sensitive data, data mining.

I. INTRODUCTION

The tremendous growth in Information and Communications technology increases the need for electronic data to be stored and shared securely. The huge amount of data, if publicly available can be utilized for many research purposes. Data Mining can be one of the technologies used to extract knowledge from massive collection of data. On the other hand, being published, the sensitive information about individuals may be disclosed which create ethical or privacy issues. Due to privacy issues many individuals are reluctant to share their data to the public which leads to data unavailability. Thus, privacy should be an important concern in the field of Data Mining. Privacy Preserving Data Mining (PPDM) is becoming a popular research area to address various privacy issues. This paper provides an extensive study of various literatures and gives some conclusions based on certain parameters.

II. TECHNIQUES OF PRIVACY PRESERVING

A. Method of anonymization

TABLE-IV
ORIGINAL PATIENTS TABLE

Attributes			
ID	Zip Code	Age	Disease
1	93461	36	Headache
2	93434	34	Headache
3	93867	41	Fever
4	93849	49	Cough

TABLE-V
ANONYMOUS VERSIONS OF TABLE-I

Attributes			
ID	Zip Code	Age	Disease
1	93*	3*	Headache
2	93*	3*	Headache
3	93*	4*	Fever
4	93*	4*	Cough

While k-anonymity protects against identity disclosure, it does not provide sufficient protection against attribute disclosure. There are two attacks: the homogeneity attack and the background knowledge attack. Because the limitations of the k-anonymity model stem from the two assumptions. First, it may 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 mode assumes a certain method of attack, while in real scenarios there is no reason why the attacker should not try other methods. Example1. Table-IV is the Original data table, and Table-V is an anonymous version of it satisfying 2- anonymity. The Disease attribute is sensitive. Suppose Manu knows that Ranu is a 34 years old woman living in ZIP 93434 and Ranu's record is in the table. From Table-V, Manu can conclude that Ranu corresponds to the first equivalence class, and thus must have fever. This is the homogeneity attack. For an example of the background knowledge attack, suppose that, by knowing Sonu's age and zip code, Manu can conclude that Sonu's corresponds to a record in the last equivalence class in Table-V. Furthermore, suppose that Manu knows that Sonu has very low risk for cough. This background knowledge enables Manu to conclude that Sonu most likely has fever.

B. Perturbation approach

The perturbation approach works under the need that the data service is not allowed to learn or recover precise records. This restriction naturally leads to some challenges. Since the method does not reconstruct the original data values but only distributions, new algorithms need to be developed which use these reconstructed distributions in order to perform mining of the underlying data. This means that for each individual data problem such as classification, clustering, or association rule mining, a new distribution based data mining algorithm needs to be developed. For example, Agrawal [3] develops a new distribution-based data mining algorithm for the classification problem, whereas the techniques in Vaidya and Clifton and Rizvi and Haritsa[4] develop methods for privacy-preserving association rule mining. While some clever approaches have been developed for distribution-based mining of data for particular problems such as association rules and classification, it is clear that using distributions instead of original records restricts the range of algorithmic techniques that can be used on the data [5].

In the perturbation approach, the distribution of each data dimension reconstructed independently. This means that any distribution based data mining algorithm works under an implicit assumption to treat each dimension independently. In many cases, a lot of relevant information for data mining algorithms such as classification is hidden in inter-attribute correlations. For example, the classification technique uses a distribution-based analogue of single-attribute split algorithm. However, other techniques such as multivariate decision tree algorithms cannot be accordingly modified to work with the perturbation approach. This is because of the independent treatment of the different attributes by the perturbation approach. This means that distribution based data mining algorithms have an inherent disadvantage of loss of implicit information available in multidimensional records. Another branch of privacy preserving data mining which using cryptographic techniques was developed. This branch became hugely popular for two main reasons: Firstly, cryptography offers a well-defined model for privacy, which includes methodologies for proving and quantifying it. Secondly, there exists a vast tool set of cryptographic algorithms and constructs to implement privacy-preserving data mining algorithms. However, recent work has pointed that cryptography does not protect the output of a computation. Instead, it prevents privacy leaks in the process of computation. Thus, it falls short of providing a complete answer to the problem of privacy preserving data mining.

C. Randomized response techniques

The method of randomization can be described as follows. Consider a set of data records denoted by X = {x1 . .xN}. For record xi א X, we add a noise component which is drawn from the probability distribution fY(y). These noise components are drawn independently, and are denoted y1 . . .yN Thus, the new set of distorted records are denoted by x1+y1 . . .xN +yN. We denote this new set of records by z1 . . . zN. 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. Thus, if X be the random variable denoting the data distribution for the original record, Y be the random variable describing the noise distribution, and Z be the random variable denoting the final record, we have:

$Z = X + Y$
$X = Z – Y$

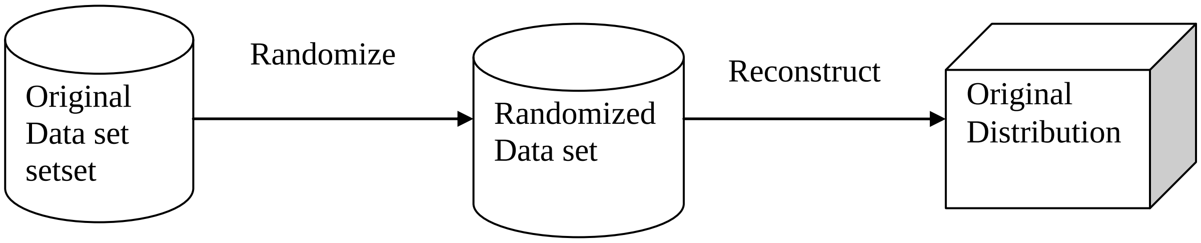

Fig. 1 The Model   Randomization

Now, we note that N instantiations of the probability distribution Z are known, whereas the distribution Y is known publicly. For a large enough number of values of N, the distribution Z can be approximated closely by using a variety of methods such as kernel density estimation. By subtracting Y from the approximated distribution of Z, it is possible to approximate the original probability distribution X. In practice, one can combine the process of approximation of Z with subtraction of the distribution Y from Z by using a variety of iterative methods. Such iterative methods typically have a higher accuracy than the sequential solution of first approximating Z and then subtracting Y from it. The basic idea of randomized response is to scramble the data in such a way that the central place cannot tell with probabilities better than a pre-defined threshold whether the data from a customer contain truthful information or false information. Although information from each individual user is scrambled, if the number of users is significantly large, the aggregate information of these users can be estimated with decent accuracy. Such property is useful for decision-tree classification since decision-tree classification is based on aggregate values of a data set, rather than individual data items. Randomized Response technique was first introduced by Warner 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. In the Related-Question Model, instead of asking each respondent whether he/she has attribute A the interviewer asks each respondent two related questions, the answers to which are opposite to each other .When the randomization method is carried out, the data collection process consists of two steps .The first step is for the data providers to randomize their data and transmit the randomized data to the data receiver. In the second step, the data receiver estimates the original distribution of the data by employing a distribution reconstruction algorithm. The model of randomization is shown in Fig 1. One key advantage of the randomization method is that it is relatively simple, and does not require knowledge of the distribution of other records in the data. Therefore, 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.

D. Condensation approach

We introduce a condensation approach, [8] which constructs constrained clusters in the data set, and then generates pseudo-data from the statistics of these clusters .We refer to the technique as condensation because of its approach of using condensed statistics of the clusters in order to generate pseudo-data. The constraints on the clusters are defined in terms of the sizes of the clusters which are chosen in a way so as to preserve k-anonymity. This method has a number of advantages over the perturbation model in terms of preserving privacy in an effective way. In addition, since the 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. Furthermore, the use of pseudo-data no longer necessitates the redesign of data mining algorithms, since they have the same format as the original data. In contrast, when the data is constructed with the use of generalizations or suppressions, we need to redesign data mining algorithms to work effectively with incomplete or partially certain data. It can also be effectively used in situations with dynamic data updates such as the data stream problem. We discuss a condensation approach for data mining. This approach uses a methodology which condenses the data into multiple groups of predefined size, for each group, certain statistics are maintained. Each group has a size atleast k, which is referred to as the level of that privacy preserving approach. The greater the level, the greater the amount of privacy. At the same time, a greater amount of information is lost because of the condensation of a larger number of records into a single statistical group entity. We use the statistics from each group in order to generate the corresponding pseudo-data.

E. Cryptographic technique

Another branch of privacy preserving data mining which using cryptographic techniques was developed. This branch became hugely popular [6] for two main reasons: Firstly, cryptography offers a well-defined model for privacy, which includes methodologies for proving and quantifying it. Secondly, there exists a vast toolset of cryptographic algorithms and constructs to implement privacy-preserving data mining algorithms. However, recent work [7] has pointed that cryptography does not protect the output of a computation. Instead, it prevents privacy leaks in the process of computation. Thus, it falls short of providing a complete answer to the problem of privacy preserving data mining.

III. MERITS AND DEMERITS OF DIFFERENT TECHNIQUES OF PPDM

After reviewing different techniques of privacy preserving the pros and cons are tabulated.

Techniques of PPDM	Merits	Demerits
ANONYMIZATION	This method is used to protect respondent's identities while releasing truthful information. While k-anonymity protects against identity disclosure, it does not provide sufficient protection against attribute disclosure.	There are two attacks: the homogeneity attack and the background knowledge attack. Because the limitations of the k-anonymity model stem from the two assumptions. First, it may be very hard for the owner of a database to determine which of the attributes 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.
PERTURBATION	Independent treatment of the different attributes by the perturbation approach.	The 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	The randomization method is a simple technique which can be easily implemented at data collection time. It has been shown to be a useful technique for hiding individual data in privacy Preserving data mining. The Randomization method is more efficient. However, it results in high information loss.	Randomized Response technique is not for multiple attribute databases.
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 vast toolset of cryptographic algorithms and constructs to implement privacy preserving data mining algorithms.	This approach is especially difficult to scale when more than a 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.

IV. COMPARISON BETWEEN DIFFERENT TECHNIQUES

There are many different techniques proposed in the field of Privacy Preserving Data Mining but one outperforms over other or vice versa on different criteria. Algorithms are classified on the basis of performance, utility, cost, complexity, tolerance against data mining algorithms etc. I have shown a tabular comparison of the work done by different authors in a chronological order (from past to present). I have taken the parameters like technique used for PPDM, its approach, results and accuracy

TABULAR COMPARISON OF DIFFERENT TECHNIQUES

Technique Used for PPDM	Approach	Result and Accuracy
Cryptographic Technique	A technique through which sensitive data can be encrypted. There is also a proper toolset for algorithms of cryptography.	This approach is especially difficult to scale when more than a few parties are involved. Also it does not hold good for large databases.
Data Perturbation	They tried to preserve data privacy by adding random noise, while making sure that the random noise still preserves the “signal” from the data so that the patterns can still be accurately estimated.	Randomization-based Techniques are used to generate random matrices.
Condensation Approach	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.
Anonymization	Anonymization is a technique for hiding individual’s sensitive data from owner’s record. K-anonymity is used for generalization and suppression for data hiding.	Background Knowledge and Homogeneity attacks of K-Anonymity Algorithm do not preserve sensitivity of an individual.

V. CONCLUSION

The main intension of this paper is accomplished through reiterating various PPDM techniques in literatures for handling privacy issues in data mining. Most of the studies show that there exist tradeoffs between privacy, information loss and computational overhead. Maximizing the data utility by preserving information is the critical challenge while protecting privacy. To provide accurate results in data mining, many PPDM techniques are task based. Since, no such technique exists which overcomes all privacy issues, research in this direction can make significant contributions. The study can be carried out using any one of the existing techniques or using a combination of these as illustrated in [9] or by developing entirely a new technique. This survey will definitely help the researchers to set their own privacy goals according to specific demands.

REFERENCES

[1] P.Samarati,(2001), *Protecting respondent's privacy in micro data release,*In IEEE Transaction on knowledge and Data Engineering, pp.010-027.

[2] L. Sweeney, (2002)."*k-anonymity: a model for protecting privacy",*International Journal on Uncertainty, Fuzziness and Knowledge based Systems, pp. 557-570.

[3] Agrawal, R. and Srikant, R. (2000),"*Privacy-preserving data mining*".In Proc. SIGMOD00, pp. 439-450.

[4] Evfimievski, A.Srikant, R.Agrawal, and Gehrke J(2002),"*Privacy preserving mining of association rules*". In Proc.KDD02, pp. 217-228.

[5] Hong, J.I. and J.A. Landay,(2004), “*Architecture for Privacy Sensitive Ubiquitous Computing*", In Mobisys04, pp. 177- 189.

[6] Laur, H. Lipmaa, and T. Mieli' ainen,(2006)."*Cryptographically private support vector machines*". In Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 618-624.

[7] Ke Wang, Benjamin C. M. Fung1 and Philip S. Yu, (2005) "*Template based privacy preservation in classification problems*", In ICDM, pp. 466- 473.

[8] Charu C. Aggarwal and Philip S. Yu,(2004) “*A condensation approachto privacy preserving data mining*”, In EDBT, pp. 183–199.

[9] Syed Zahid Hassan and Brijesh Verma, “A Hybrid Data Mining Approach for Knowledge Extraction and Classification in Medical Databases”, Seventh International Conference on Intelligent Systems Design and Applications.