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. 2, Issue 11, November 2014

A Survey on Privacy Preservation Recent Approaches and Techniques

Dhivakar K [1], Mohana S [2]

P.G. Student, Department of Computer Science and Engineering, M.I.E.T Engineering College, Trichy, India[1]

Assistant Professor, Department of Computer Science and Engineering, M.I.E.T Engineering College, Trichy, India[2]

ABSTRACT: Data mining is a process of extracting useful knowledge from large data sets. The typical process of data collection and data dissemination result in a possible risk of privacy threats and attacks. Some private information about individuals, businesses and organizations has to be suppressed before it is shared or published. In recent years, privacy preserving data mining has been studied extensively, because of the wide proliferation of sensitive information on the internet. We discussed about the recent approaches involved in privacy preservation such as randomization, anonymization, perturbation and distributed privacy preservation. We discuss the computational and theoretical limits associated with privacy preservation over high dimensional data sets.

KEYWORDS: Privacy preservation, anonymization, randomization, distributed privacy preserving

I. INTRODUCTION

In recent years, data mining has been viewed as a threat to privacy because of the wide spread proliferation of electronic data maintained by corporations. This has leads to increased concerns about the privacy of the underlying data. In the last few decades a number of approaches and techniques such as classification, association rule mining have been proposed for modifying or transforming the data in such a way so as to preserve the privacy. Preservations of individuals information is an essential for the data owners to ensure his privacy. Privacy plays an important role in data publishing.

Data mining process allows a company to use large amount of data to develop correlations and relationships among the data to improve the business efficiency. Therefore privacy preserving data mining has become important field of research. The Data Mining technology can develop these analyses on its own, using commix of statistics, artificial intelligence, machine learning algorithms, and data stores.

In order to face the challenging risk, some researchers have been proposed as a remedy of this awkward situation, which target at accomplishing the balance of data utility and information privacy when publishing dataset. The ongoing research is called Privacy Preserving Data Publishing. Balancing the privacy of the data as per the legitimate need of the user is the major problem.

The original data is modified by the sanitization process to conceal sensitive knowledge before release so the problem can be addressed. Privacy preservation of sensitive knowledge is addressed by several researchers in the form of association rules by suppressing the frequent item sets. As the data mining deals with generation of association rules, the change in support and confidence of the association rule for hiding sensitive rules is done. A new concept named 'not altering the support' is proposed to hide an association rule.

Confidentiality issues in data mining. A key problem that arises in any en masse collection of data is that of confidentiality. The need for privacy is sometimes due to law (e.g., for medical databases) or can be motivated by business interests. The irony is that data mining results rarely violate privacy. The objective of data mining is to generalize across populations, rather than reveal information about individuals.

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. 2, Issue 11, November 2014

The hitch is that data mining works by evaluating individual data that is subject to privacy concerns. The paper is organized as follows. Section 2 is about Related works in privacy preservation Section 3 introduces the techniques for privacy preservation Section 3 summarize the concepts about privacy preservation in distributed data base. Section 4 discuss about the modern privacy attacks. Section 5 gives explanation about the comparative study about the various approaches of privacy preservation.

II. RELATED WORKS

Data privacy has been an active research topic in the statistics, database, and security communities for the last three decades[5] Interactive versus non-interactive. In an interactive framework, a data miner can pose queries through a private mechanism, and a database owner answers these queries in response. In a non interactive framework, a database owner first anonymizes the raw data and then releases the anonymized version for data analysis. Once the data are published, the data owner has no further control over the published data. This approach is also known as privacy preserving data publishing (PPDP)[5], In the distributed (multiparty) scenario, data owners want to achieve the same tasks as single parties on their integrated data without sharing their data with others Our proposed algorithm addresses the distributed and non-interactive scenario. Below, we briefly review the most relevant research works. Iyengar[9] has presented the anonymity problem for classification and proposed a genetic algorithmic solution. Fung et al. have proposed another anonymization technique for classification using multidimensional recoding. Research proposals[10] address the problem of non-interactive data release only consider the single-party scenario. Therefore, these techniques do not satisfy the privacy requirement of our data integration application for the financial industry. Jurczyk and Xiong[12] have proposed an algorithm to securely integrate horizontally partitioned data from multiple data owners without disclosing data from one party to another.

III. TECHNIQUES FOR PRIVACY PRESERVATION

In this paper, we will provide a broad overview of the different techniques for privacy preserving data mining. We will provide a broad view of the major algorithms available for each method, and the variations on the different techniques. We will also discuss a combination of different concepts.

3.1 THE RANDOMIZATION METHOD

In this section, we will discuss the randomization method for the privacy preserving data mining. The randomization method has been traditionally used in the context of distorting data by probability distribution for methods such as surveys which have an evasive answer bias because of privacy concerns

The method of randomization can be described as follows. Consider a set of data records denoted by X ={x1,….xn}. For record xi ε X, we add a noise component which is drawn from the probability distribution $f_y(y)$. These noise components are drawn independently, and are denoted $y_1....y_{n..}$ Thus , new set of distorted records are denoted by $x_1+y_1…..x_n+y_n$. we denote this new set of records by $z_1…….z_{n.}$

The randomization method has been extended to a variety of data mining problems. A number of other techniques have also been proposed which seem to work well over a variety of different classifiers. Techniques have also been proposed for privacy preserving methods of improving the effectiveness of classifiers. The problem of association rules is especially challenging because of the discrete nature of the attributes corresponding to presence or absence of items.

3.1.1 PRIVACY QUANTIFICATION

The quantity used to measure privacy should indicate how closely the original value of the attribute can be estimated. If the original value can be estimated with c% confidence to lie in the interval, then the interval width defines the amount of privacy at c% confidence level.

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. 2, Issue 11, November 2014

3.1.2 ADVERSARIAL ATTACKS ON RANDOMIZATION

In general , a systematic approach can be used to do this in multi-dimensional data sets with the use of spectral filtering or PCA based techniques. The broad idea in techniques such as PCA is that correlation structure in the original data can be estimated fairly accurately even after the noise addition. Once the broad correlation structure in the data has been determined, one then try to remove the noise in the data in such a way that it fits the aggregate correlation structure of the data.

The another kind of the adversarial attack is with the use of public information. Consider a record X $=\{x_1 \ldots\ldots x_d\}$, which is perturbed to Z $=\{z_1 \ldots.z_d\}$. Then, since the distribution of the perturbations is known, we can try to use a maximum likelihood fit of the potential perturbation of Z to a public record. The higher the log-likelihood fit, the greater the probability that the record W corresponds to X. If it is known that the public data set always includes X, then the maximum likelihood fit can provide a high degree of certainty in identifying the correct record, especially in the case where d is large.

3.1.3 RANDOMIZATION METHODS FOR DATA STREAMS

The randomization approach is particularly well suited to privacy-preserving data mining of streams, since the noise added to a given record is independent of the rest of the data. However, streams provide a particularly vulnerable target for adversarial attacks with the use of PCA techniques because of the large volume of the data available for analysis. In an interesting techniques for randomization has been proposed which uses the auto correlation in different time series while deciding the noise to be added to any particular value.

3.1.4 MULTIPLICATIVE PERTURBATIONS

The most common method of the randomization is that of additive perturbations. However, multiplicative perturbations can also be used to good effect for privacy preserving data mining. Multiplicative perturbations can also be used for distributed privacy preserving data mining.

3.1.5 DATA SWAPPING

We note the noise addition or multiplication is not only technique which can be used to perturb the data. A related method is that of data swapping, in which values across different records are swapped in order to perform the privacy preservation. One advantage of this techniques is that the lower order marginal totals of the data are completely preserved and are not perturbed at all. We note that this technique does not follow the general principle in randomization which allows the value of the record to be perturbed independent; y of 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.

3.2 ANONYMIZATION TECHNIQUES

The randomization method is a simple technique which can be easily implemented at data collection time , because the noise added to a given record is independent of the behavior of other records. This is also the weakness because outlier records can often be the difficult to mask. Another key weakness of the randomization frame work is that it does not consider the possibility that publicly available record can be used to identify the identity of the owners of those records. Therefore, a broad approach to many privacy transformations is to construct groups of anonymous records which are transformed to group specific way

3.2.1 K-ANONYMITY

The data records are made available by simply removing key identifiers such as the name and social-security numbers from personal records. K-anonymity provide protection by ensuring that released information map to no, k or

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. 2, Issue 11, November 2014

incorrect entities, respectively. To determine how many individuals each released tuple actually matches requires combining the released data with externally available data and analyzing other possible attacks.

DEFINITION k-anonymity

Let $RT(A_{1,\ldots}A_n)$be a table and QI_{RT} be the quasi identifier associated with it. RT is said to satisfy k-anonymity if and only if each sequence of values in $RT[QI_{RT}]$ appears with at least k occurrences in $RT[QI_{RT}]$.

The approaches assumes on ordering among the quasi-identifiers attributes. The values of the attributes are discretized in to intervals or grouped into different sets of values. An index is created using these attribute-interval pair and a set enumeration tree is a systematic enumeration of all possible generalizations with use of these groupings. A branch and bound techniques can be used to successively improve the quality of the solution during traversal process. Eventually it is possible to terminate the algorithm at a maximum computational time, and use the current solution at that point. The reduction in anonymity is always controlled, so that k-anonymity is never violated. Since the problem of k-anonymization is essentially a search over a space of possible multi-dimensional solutions, standard heuristics search techniques such as genetic algorithms or simulated annealing can be effectively used. This kind of technique can be useful in situations where it is desirable to determine whether or not anonymization should be used as the technique of choice for a particular situation.

3.2.1.1 **ATTACK AGAINST K ANONYMITY**

a) Unsorted matching attack against k-anonymity

This attack is based on the order in which tuples appear in the released table. These can be corrected by randomly sorting the tuples of the solution table. Otherwise , the release of a related table can leak sensitive information

b) Complementary release attack against k-anonymity

In this attack, that the attributes that constitute the quasi identifier are themselves a subset of the attributes released. As a result, when a table T, which adheres to k-anonymity, is released, it should be considered as joining other external information. Therefore, subsequent releases of the same privately held information must considered all of the released attributes of T a quasi identifier to prohibit linking on T, unless of course, subsequent releases are based on T.

c) Temporal attack against k-anonymity

Data collections are dynamic. Tuples are added, changed, and removed constantly. As a result, releases of generalized data over time can be subject to a temporal attack. Let table T_0 be the original privately held table at time t=0. Assume a k-anonymity solution based on T_0, which call as table RT_0, is released. Let RT_t be a k-anonymity solution based on T_t that is released at time t. Because there is no requirement that RT_t respect RT_o linking the tables RT_t and RT_o may reveal sensitive information and thereby compromise k-anonymity protection.

K-anonymity algorithm fails to protect privacy or overly reduce the utility of the data. The anonymity was achieved by the use of secure multiparty computation. Privacy requirement for anonymizations were personalized based on the individuals preferences on sensitive attributes.

Bottom-up search strategy is proposed for finding optimal anonymizations. This strategy works particularly well when the value of k is small. More sophisticated generalization schemes allow more valid generalizations and produce a dataset with better data quality.

3.3 PERTURBATION TECHNIQUES

Privacy concerns over the ever-increasing gathering of personal information by various institutions led to the development of privacy preserving data. The approach protects the privacy of the data by perturbing the data through a

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. 2, Issue 11, November 2014

method. The major challenge of data perturbation is to achieve the desired result between the level of data privacy and the level of data utility.

Data privacy and data utility are commonly considered as a pair of conflicting requirements in privacy-preserving of data for applications and mining systems. Multiplicative perturbation algorithms aim at improving data privacy while maintaining the desired level of data utility by selectively preserving the mining task and model specific information during the data perturbation process.

The multiplicative perturbation algorithm may find multiple data transformations that preserve the required data utility. Thus the next major challenge is to find a good transformation that provides a satisfactory level of privacy data.

Evaluation of data perturbation technique

The data perturbation technique has the benefits of efficiency, and does not require knowledge of the distribution of other records in the data. This is not true of other methods such as k-anonymity which require the knowledge of other records in the data. This technique does not require the use of a trusted server containing all the original records in order to perform the anonymization process.

IV. PRIVACY PRESERVATION IN THE DISTRIBUTED DATABASE.

The key goal in most distributed methods for privacy preserving data mining is to allow computation of useful aggregate statistics over the entire dataset without compromising the privacy of the individual data sets with in the different participants. Thus , 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 entities, each of which have the same set of attributes. In vertically partitioning, the individual entities may have different attributes of the same set of the records. Both kinds of partitioning pose different challenges to the problem of distributed privacy-preserving data mining.

The problem of distributed privacy preserving data mining overlaps closely with a field in cryptography for determining secure multiparty computations. A broad overview of the intersection between the fields of cryptography.

4.1 DISTRIBUTED ALGORITHM OVER HORIZONTALLY PARTITIONED DATA SETS.

In horizontally partitioned data sets, different sites contain different sets of records with the same sets of attributes which are used for the mining purpose. Subsequently, a variety of classifiers have been generalized to the problem of horizontally partitioned case in which privacy preservation classification is performed in a fully distributed setting, where each customer has the private access to only their own record.

4.2 DISTRIBUTED ALGORITHM OVER VERTICALLY PARTITIONED DATA.

For the vertically partitioned case, many primitives operation such as computing the scalar product or the secure set size intersection can be useful in computing the results of the data mining algorithms. Another method for association rule mining uses the secure scalar product over the vertical bit representations of the item set inclusions in transactions, in order to compute the frequency of the corresponding item sets.

4.3 DISTRIBUTED ALGORITHMS FOR K-ANONYMITY

In many cases, it is important to maintain k-anonymity across different distributed parties. In a k anonymous protocol for data which is vertically partitioned across two parties. The issue of k-anonymity is also important in the context of distributed location based services. In this case k-anonymity of the user identity is maintained even when the location information is released.

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. 2, Issue 11, November 2014

Similar issues arise in the context of communication protocols in which the anonymity of senders may need to protect. A message is said to sender k-anonymous. If it is guaranteed that an attacker can at most narrow down the identity of the sender to k-individuals. Similarly a message is said to receiver k-anonymous, if it is guaranteed that an attacker can at most narrow down the identity of the receiver to k individuals.

V. MODERN PRIVACY ATTACKS

5.1 **Background Knowledge Attack**

Anatomy as an alternative anonymization technique to generalization. Anatomy releases all the quasi-identifier and sensitive data directly into two separate tables proposed an anatomizing algorithm to compute the anatomized tables. The algorithm first hashes the records into buckets based on the sensitive attribute, i.e., records with the same sensitive values are in the same bucket. Then the algorithm iteratively obtains the buckets that currently have the largest number of records and selects one record from each of the buckets to form a group. Each re- maining record is then assigned to an existing group.

5.2 **Unsorted Matching Attack**

This attack is based on the order in which tuples appear in the released table. While we have maintained the use of a relational model, and so the order of tuples cannot be assumed, in real-world use this is often a problem. It can be corrected of course, by randomly sorting the tuples of the solution. Otherwise, the release of a related table can leak sensitive information.

Solution: Random shuffling of rows.

5.3 **Complementary Release Attack**

It is more common that the attributes that constitute the quasi-identifier are themselves a subset of the attributes released. As a result, when a k-minimal solution, which we will call table T is released, it should be considered as joining other external information. Therefore, subsequent releases of generalizations of the same privately held information must consider all of the released attributes of T a quasi-identifier to prohibit linking on T, unless of course, subsequent releases are themselves generalizations of T.

Solution:
1) Consider all of the released tables before release the new one, and try to avoid linking.
2) Other data holders may release some data that can be used in this kind of attack. Generally, this kind of at-tack is hard to be prohibited complete

5.4 **Temporal Attack**

Data collections are dynamic. Tuples are added, changed, and removed constantly. As a result, releases of generalized data over time can be subject to a temporal inference attack.

Solution: Subsequent releases must use the already re- leased table.

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. 2, Issue 11, November 2014

VI. COMPARISON STUDY

Techniques	Merits	Demerits
anonymization	This method is used to protect respondents identities while releasing truthful information	There are two attacks: the homogeneity attack and the background knowledge attack. Because the limitation of the k-anonymity model stem from the two assumptions. 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
ℓ-Diversity	Sensitive attribute would have at most same frequency	Homogeneity and background knowledge attack has lacked
t-closeness	Measure the distance between two probabilistic distribution that were indistinguishable from one another	Information gain was unclear
perturbation	Independent treatment of the different attributes by perturbation approach	The method does not reconstruct the original data values, but only distributions to carry out mining of the data available
Randomized response	The randomized method is a simple technique which can easily implemented at data collection time.	Randomized response technique is not for multiple attribute databases.
Distributed K -Anonymity framework (DKA)	Global Anonymization to ensure privacy	Utility and potential were misused
Slicing	Randomization on sensitive attribute	Utility and risk measures not matched
condensation	This approach works with pseudo data rather than with modifications of original data , this help 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 the data mining algorithms, since they have the same format as the original data.

VII. CONCLUSIONS

This paper, discussed about various approaches and techniques used in privacy preservation of data mining. Due to the large collection of information, it is important to maintain the Privacy of sensitive information. Each technique has its own advantages and dis-advantages. Most privacy attacks can be effectively destroyed by the advanced techniques and approaches. In distributed privacy preserving data mining areas, efficiency is an essential issue. Privacy and accuracy is

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. 2, Issue 11, November 2014

the pair of contradiction; improving one usually incurs a cost in other. All methods are approximate to our goal of privacy preservation, we need to further perfect those approaches or develop some efficient methods.

REFERENCES

[1] N.Mohammed, D.Alhadidi, B.C.M. Fung "Secure Two-Party Differentially Private Data Release for Vertically Partitioned Data", Ieee transactions on dependable and secure computing, 2014.
[2] N. Mohammed, B. C. M. Fung, P. C. K. Hung, and C. Lee, "Centralized and distributed anonymization for high- dimensional healthcare data," ACM Transactions on Knowledge Discovery from Data (TKDD), vol. 4, no. 4, pp.18:1–18:33, October 2010.
[3] Ravindra S, Wanjari Prof .Devi,(2013), "Improving the implementation of new approach for Data Privacy Preserving in Data Mining using slicing". International Journal of Modern Engineering Research (IJMER), Vol. 3, Issue. 3
[4] Mohnish Patel, Prashant Richariya, Anurag Shrivastava, (2013),''A review paper on Privacy-Preserving Data Mining", Review article on Scholars Journal of Engineering and Technology (SJET) , pp.359-361
[5] B. C. M. Fung, K. Wang, R. Chen, and P.S.Yu,"Privacy-preserving data publishing: A survey of recent developments," ACM Computer. Survey., vol. 42, pp. 14:1–14:53, June 2010.
[6] Charu C.Aggarwal, "A General survey of privacy preserving Data Mining Models and Algorithms", IBM,T. J. Watson Research Centre.
[7] P. Jurczyk and L. Xiong, "Distributed Anonymization: Achieving Privacy for Both Data Subjects and Data Providers," Proc. Ann. . IFIP WG 11.3 Working Conf. Data and Applications Security (DBSec '09), 2009
[8] Charu C. Aggarwal and Philip S. Yu. Privacy-Preserving Data Mining: Models and Algorithms. Springer Publishing Company, Incorporated, July 2008.
[9] Transforming Data to Satisfy Privacy Constraints,"Proc. ACM Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '02), 2002
[10] M. Hay, V. Rastogi, G. Miklau, and D. Suciu, "Boosting the Accuracy of Differentially Private Histograms through Consistency," Proc. Int'l Conf. Very Large Data Bases (VLDB '10), 2010
[11] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, "Our Data Ourselves: Privacy via Distributed Noise Generation," Proc. 25th Ann. Int'l Conf. Theory and Applications of Cryptographic Techniques (EUROCRYPT '06), 2006.
[12] P. Jurczyk and L. Xiong, "Distributed Anonymization: Achieving Privacy for Both Data Subjects and Data Providers," Proc. Ann. IFIP WG 11.3 Working Conf. Data and Applications Security (DBSec '09), 2009.