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, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

Survey of Various Techniques to Provide Multilevel Trust in Privacy Preserving Data Mining

R.Mynavathi, N.Sowmiya, D.Vanitha

Dept of IT, Velalar college of Engineering and Technology, Erode, Tamilnadu, India[1, 2, 3]

ABSTRACT—The new dimension of Multilevel Trust (MLT) poses new challenges for perturbation-based PPDM. In contrast to the single-level trust scenario where only one perturbed copy is released, now multiple differently perturbed copies of the same data are available to data miners at different trusted levels.The problem of developing accurate models about aggregated data without access to precise information in individual data record is addressed. Previous solutions of this approach are limited in their assumption of single-level trust on data miners.In Single-level trust, only one perturbed copy of data is released.In the proposed system, we expand the scope of perturbation based PPDM to Multilevel Trust (MLT-PPDM). In our setting, the more trusted a data miner is, the less perturbed copy of the data it can access.A malicious data miner may have access to differently perturbed copies of the same data through various means. Prevents from jointly reconstructing the original data. Allows a data owner to generate perturbed copies of its data for arbitrary trust levels on demand. As with most existing work on perturbation-based PPDM, the existing work is limited in the sense that it considers only linear attacks (attack by single party).When two or more different parties involved in combining the perturbed copies and tries to recover the privacy, then the techniques are less suitable. More powerful adversaries may apply nonlinear techniques to derive original data and recover more information.This feature provides data owners' more flexibility.

KEYWORDS—Privacy preserving data mining, Multilevel trust, Random perturbation.

I. INTRODUCTION

In recent years, data mining has been considered as a threat to privacy since the cause of widespread proliferation of electronic data maintained by corporations. This has lead to increased concerns about the privacy of the underlying data. In recent years, a number of techniques have been proposed for modifying or transforming the data in such a way so as to preserve privacy. Privacy-preserving data mining techniques may be found in [1]. In this chapter, an overview of the state-of-the-art in privacy-preserving data mining will be studied. Privacy-preserving data mining finds numerous applications in surveillance which are naturally supposed to be "privacy-violating" applications. The key is to design methods [2] which continue to be effective, without compromising security. In [2], a number of techniques have been discussed for biosurveillance, facial de-dentification, and identity theft. More detailed discussions on some of these sssues may be found in [3, 4–6]. Most methods for privacy computations use some form of transformation on the data in order to perform the privacy preservation. Typically, such methods reduce the granularity of representation in order to reduce the privacy. This reduction in granularity results in some loss of effectiveness of data management or mining algorithms. This is the natural trade-off between information loss and privacy. Some examples of such techniques are as follows:

The randomization method: The randomization method is a technique for privacy-preserving data mining in which noise is added to the data in order to mask the attribute values of records [2, 5]. The noise added is sufficiently large so that individual record values cannot be recovered. Therefore, techniques are designed to derive aggregate distributions from the perturbed records. Subsequently, data mining techniques can be developed in order to work with these aggregate distributions. We will describe the randomization technique in greater detail in a later section.

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, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

The k-anonymity model and l-diversity: The k-anonymity model was developed because of the possibility of indirect identification of records from public databases. This is because combinations of record attributes can be used to exactly identify individual records. In the k-anonymity method, we reduce the granularity of data representation with the use of techniques such as generalization and suppression. This granularity is reduced sufficiently that any given record maps onto at least k other records in the data. The l-diversity model was designed to handle some weaknesses in the k-anonymity model since protecting identities to the level of k-individuals is not the same as protecting the corresponding sensitive values, especially when there is homogeneity of sensitive values within a group. To do so, the concept of intra-group diversity of sensitive values is promoted within the anonymization scheme.

Distributed privacy preservation: In many cases, individual entities may wish to derive aggregate results from data sets which are partitioned across these entities. Such partitioning may be horizontal (when the records are distributed across multiple entities) or vertical (when the attributes are distributed across multiple entities). While the individual entities may not desire to share their entire data sets, they may consent to limited information sharing with the use of a variety of protocols. The overall effect of such methods is to maintain privacy for each individual entity, while deriving aggregate results over the entire data.

II. DATA HIDING

2.1.DATA PERTURBATION

Perturbation has a long history in statistical disclosure control [Adam and Wortman 1989] due to its simplicity, efficiency, and ability to preserve statistical information. The general idea is to replace the original data values with some synthetic data values, so that the statistical information computed from the perturbed data does not differ significantly from the statistical information computed from the original data. The perturbed data records do not correspond to real-world record owners, so the attacker cannot perform the sensitive linkages or recover sensitive information from the published data. Following are the commonly used perturbation methods, including additive noise, data swapping, and synthetic data generation.

Additive noise. Additive noise is a widely used privacy protection method in statistical disclosure control. It is often used for hiding sensitive numerical data (e.g., salary). The general idea is to replace the original sensitive value s with s+, r where r is a random value drawn from some distribution. Privacy was measured by how closely the original values of a modified attribute can be estimated [Agrawal and Aggarwal 2001]. Fuller [1993] and Kim and Winkler [1995] showed that some simple statistical information, like means and correlations, can be preserved by adding random noise. Experiments in Agrawal and Srikant [2000], Du and Zhan [2003], and Evfimievski et al. [2002] further suggested that some data mining information can be preserved in the randomized data. However, Kargupta et al. [2003] pointed out that some reasonably close sensitive values can be recovered from the randomized data when the correlation among attributes is high but the noise is not. Huang et al. [2005] presented an improved randomization method to limit this type of privacy breach. Some representative statistical disclosure control methods that employ additive noise.

Synthetic data generation. Many statistical disclosure control methods use synthetic data generation to preserve record owners' privacy and retain useful statistical information [Rubin]. The general idea is to build a statistical model from the data and then to sample points from the model. These sampled points form the synthetic data for data publication instead of the original data. An alternative synthetic data generation approach is condensation [Aggarwal and Yu 2008a, 2008b]. The idea is to first condense the records into multiple groups. For each group, extract some statistical information, such as sum and covariance, that suffices to preserve the mean and correlations across the different attributes. Then, based on the statistical information, for publication generate points for each group following the statistical characteristics of the group.

1) Additive perturbation

The additive perturbation is a technique for privacy-preserving data mining in which noise is added to the data in order to mask the attribute values of records (Agrawal & Srikant, 2000). The noise added is sufficiently large so that individual

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, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

record values cannot be recovered. Therefore, techniques are designed to derive aggregate distributions from the perturbed records. Subsequently, data mining techniques can be developed in order to work with these aggregate distributions.

2)Matrix multiplicative perturbation

The most common method of data perturbation is that of additive perturbations. However, matrix multiplicative perturbations can also be used to good effect for privacy-preserving data mining. The product of a discrete cosine transformation matrix and a truncated perturbation matrix, then the perturbation approximately preserves Euclidean distances. New Fundamental Technologies in Data Mining

3) 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. While this is a strength of the data perturbation technique, it also leads to some weaknesses, since 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. In order to guard against this, one may need to be needlessly more aggressive in adding noise to all the records in the data. This reduces the utility of the data for mining purposes.

4)Data swapping. The general idea of data swapping is to anonymize a data table by exchanging values of sensitive attributes among individual records, while the swaps maintain the low-order frequency counts or marginals for statistical analysis. It can be used to protect numerical attributes [Reiss et al. 1982] and categorical attributes [Reiss 1984]. An alternative swapping method is rank swapping: First rank the values of an attribute A in ascending order. Then for each value v $\in$ A, swap v with another value u $\in$ A, where u is randomly chosen within a restricted range p% of v. Rank swapping can better preserve statistical information than the ordinary data swapping [Domingo-Ferrer and Torra 2002].

Perturbation, which is accomplished by the alteration of an attribute value by a new value (i.e., changing a 1-value to a 0-value, or adding noise),

• Blocking, which is the replacement of an existing attribute value,

• Aggregation or merging which is the combination of several values into a coarser category,

• Swapping that refers to interchanging values of individual records, and

• Sampling, which refers to releasing data for only a sample of a population?

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 independent;y 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.

2.2.SECURE MULTIPARTY COMPUTATION

Substantial work has been done on secure multiparty computation. The key result is that a wide class of computations can be computed securely under reasonable assumptions. We give a brief overview of this work, concentrating on material that is used later in the paper. The definitions given here are from Goldreich [10]. For simplicity, we concentrate on the two-party case. Extending the definitions to the multiparty case is straightforward.

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, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

Security in Semihonest Model

A semihonest party follows the rules of the protocol using its correct input, but is free to later use what it sees during execution of the protocol to compromise security. This is somewhat realistic in the real world because parties who want to mine data for their mutual benefit will follow the protocol to get correct results. Also, a protocol that is buried in large, complex software cannot be easily altered. A formal definition of private two-party computation in the semihonest model is given below. Computing a function privately is equivalent to computing it securely. The formal proof of this can be found in Goldreich [10]. The above definition says that a computation is secure ifthe view of each party during the execution of the protocol can be effectively simulated by the input and the output of the party. This is not quite the same as saying that private information is protected. For example, if two parties use a secure protocol to mine distributed association rules, a secure protocol still reveals that if a particular rule is not supported by a particular site and that rule appears in the globally supported rule set, then it must be supported by the other site. A site can deduce this information by solely looking at its locally supported rules and the globally supported rules. On the other hand, there is no way to deduce the exact support count of some itemset by looking at the globally supported rules. With three or more parties, knowing a rule holds globally reveals that at least one site supports it, but no site knows which site (other than, obviously, itself). In summary, a secure multiparty protocol will not reveal more information to a particular party than the information that can be induced by looking at that party's input and the output.

III. RULE HIDING

3.1.ASSOCIATION RULE MINING STRATEGY

Association rule mining algorithms scan the database of transactions and calculate the support and confidence of the rules and retrieve only those rules having support and confidence higher than the user specified minimum support and confidence threshold. Association rule hiding algorithms prevents the sensitive rules from being disclosed. The problem can be stated as follows: "Given a transactional database D, minimum confidence, minimum support and a set R of rules mined from database D. A subset RH of R is denoted as set of sensitive association rules which are to be hidden. The objective is to transform D into a database D‟ in such a way that no association rule in RH will be mined and all non sensitive rules in R could still be mined from D‟.

ASSOCIATION RULE HIDING APPROACHES

A. Heuristic approach

This approach involves efficient, fast and scalable algorithms that selectively sanitize a set of transactions from the original database to hide the sensitive association rules [11]. Various heuristic algorithms are based on mainly two techniques: (1) Data distortion technique (2) Blocking technique. Data distortion is done by the alteration of an attribute value by a new value. It changes 1's to 0's or vice versa in selected transactions. There are two basic approaches for rule hiding in data distortion based technique: Reduce the confidence of rules and reduce the support of rules.

IV. DISTRIBUTED PRIVACY-PRESERVING DATA MINING

The key goal in most distributed methods for privacy-preserving data mining is to allow computation of useful aggregate statistics over the entire data set without compromising the privacy of the individual data sets within 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 records are spread out across multiple entities, each of which have the same set of attributes. In vertical partitioning, the individual entities may have different attributes (or views) of the same set of records. Both kinds of partitioning pose different challenges to the problem of distributed privacypreserving data mining. The problem of distributed privacy-preserving data mining overlaps closely with a field in cryptography for determining secure multi-party computations. A broad overview of the intersection between the fields of

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, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

cryptography and privacy-preserving data mining. The broad approach to cryptographic methods tends to compute functions over inputs provided by multiple recipients without actually sharing the inputs with one another. For example, in a 2-party setting, Alice and Bob may have two inputs x and y respectively, and may wish to both compute the function f(x, y) without revealing x or y to each other. This problem can also be generalized across k parties by designing the k argument function $h(x_1 \ldots x_k)$. Many data mining algorithms may be viewed in the context of repetitive computations of many such primitive functions such as the scalar dot product, secure sum etc. In order to compute the function f(x, y) or $h(x_1 \ldots, x_k)$, a protocol will have to designed for exchanging information in such a way that the function is computed without compromising privacy. We note that the robustness of the protocol depends upon the level of trust one is willing to place on the two participants Alice and Bob. This is because the protocol may be subjected to various kinds of adversarial behavior:

Semi-honest Adversaries: In this case, the participants Alice and Bob are curious and attempt to learn from the information received by them during the protocol, but do not deviate from the protocol themselves. In many situations, this may be considered a realistic model of adversarial behavior.

Malicious Adversaries: In this case, Alice and Bob may vary from the protocol, and may send sophisticated inputs to one another to learn from the information received from each other. A key building-block for many kinds of secure function evaluations is the 1 out of 2 oblivious-transfer protocol. This protocol involves two parties: a sender, and a receiver. The sender's input is a pair (x0, x1), and the receiver's input is a bit value $\sigma \in \{0, 1\}$. At the end of the process, the receiver learns $x\sigma$ only, and the sender learns nothing. A number of simple solutions can be designed for this task. In one solution, the receiver generates two random public keys, K0 and K1, but the receiver knows only the decryption key for $K\sigma$. The receiver sends these keys to the sender, who encrypts x0 with K0, x1 with K1, and sends the encrypted data back to the receiver. At this point, the receiver can only decrypt $x\sigma$, since this is the only input for which they have the decryption key. We note that this is a semihonest solution, since the intermediate steps require an assumption of trust.

4.1.DISTRIBUTED ALGORITHMS OVER HORIZONTALLY PARTITIONED DATA SETS

In horizontally partitioned data sets, different sites contain different sets of records with the same (or highly overlapping) set of attributes which are used for mining purposes. Many of these techniques use specialized versions of the general methods discussed in [18, 19] for various problems. The work discusses the construction of a popular decision tree induction method called ID3 with the use of approximations of the best splitting attributes. Subsequently, a variety of classifiers have been 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 [19], in which privacy-preserving classification is performed in a fully distributed setting, where each customer has private access to only their own record. A host of other data mining applications have been generalized to the problem of horizontally partitioned data sets. These include the applications of association rule mining, clustering and collaborative filtering. Methods for cooperative statistical analysis using secure multi-party computation methods. A related problem is that of information retrieval and document indexing in a network of content providers. This problem arises in the context of multiple providers which may need to cooperate with one another in sharing their content, but may essentially be business competitors. In [17], it has been discussed how an adversary may use the output of search engines and content providers in order to reconstruct the documents. Therefore, the level of trust required grows with the number of content providers. A solution to this problem[17] constructs a centralized privacy-preserving index in conjunction with a distributed access control mechanism. The privacy-preserving index maintains strong privacy guarantees even in the face of colluding adversaries, and even if the entire index is made public.

4.2.DISTRIBUTED ALGORITHMS OVER VERTICALLY PARTITIONED DATASETS

For the vertically partitioned case, many primitive operations such as computing the scalar product or the secure set size intersection can be useful in computing the results of data mining algorithms. For example, the methods in discuss how to use to scalar dot product computation for frequent itemset counting. The process of counting can also be achieved by using the secure size of set intersection. Another method for association rule mining discussed in uses the secure scalar

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, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

product over the vertical bit representation of itemset inclusion in transactions, in order to compute the frequency of the corresponding itemsets. This key step is applied repeatedly within the framework of a roll up procedure of itemset counting. It has been shown in that this approach is quite effective in practice. The approach of vertically partitioned mining has been extended to a variety of data mining applications such as decision trees,SVM Classification, Naive Bayes Classifier, and k-means clustering.A number of theoretical results on the ability to learn different kinds of functions in vertically partitioned databases with the use of cryptographic approaches are discussed.

REFERENCES

[1] Verykios V. S., Bertino E., Fovino I. N., Provenza L. P., Saygin Y., Theodoridis Y.: State-of-the-art in privacy preserving data mining. ACM SIGMOD Record, v.33 n.1, 2004.

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

[3] Newton E., Sweeney L.,Malin B.: Preserving Privacy by De-identifying Facial Images. IEEE Transactions on Knowledge and Data Engineering, IEEE TKDE, February 2005.

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

[5] Sweeney L.: AI Technologies to Defeat Identity Theft Vulnerabilities. AAAI Spring Symposium, AI Technologies for Homeland Security, 2005.

[6] Sweeney L., Gross R.: Mining Images in Publicly-Available Cameras for Homeland Security. AAAI Spring Symposium, AI Technologies for Homeland Security, 2005.

[7] Verykios V. S., Elmagarmid A., Bertino E., Saygin Y.,, Dasseni E.: Association Rule Hiding. IEEE Transactions on Knowledge and Data Engineering, 16(4), 2004.

[8] Moskowitz I., Chang L.: A decision theoretic system for information downgrading. Joint Conference on Information Sciences, 2000.

[9] Adam N., Wortmann J. C.: Security-Control Methods for Statistical Databases: A Comparison Study. ACMComputing Surveys, 21(4), 1989.

[10] O. Goldreich, "Secure Multiparty Computation," (working draft), Sept. 1998, available: http://www.wisdom.weizmann.ac.il/ ~oded/pp.html.

[11] Aris Gkoulalas–Divanis;Vassilios S. Verykios "Association Rule Hiding For Data Mining" Springer, DOI 10.1007/978-1-4419-6569-1, Springer Science + Business Media, LLC 2010

[12] M. Atallah, E. Bertino, A. Elmagarmid, M. Ibrahim, and V. S. Verykios "Disclosure limitation of sensitive rules,".In Proc. of the 1999 IEEE Knowledge and Data Engineering Exchange Workshop (KDEX'99), pp. 45–52, 1999.

[13] Vassilios S. Verykios, A.K. Elmagarmid, E. Bertino, Y. Saygin, and E. Dasseni, "Association Rule Hiding," IEEE Transactions on Knowledge and Data Engineering, vol. 16, no. 4, pp. 434-447, 2004.

[14] Shyue-LiangWang ;Dipen Patel ;Ayat Jafari ;Tzung-Pei Hong, "Hiding collaborative recommendation association rules", Published online: 30 January 2007, Springer Science+Business Media, LLC 2007

[15] Naor M., Pinkas B.: Efficient Oblivious Transfer Protocols, SODA Conference, 2001.

[16] Yao A. C.: How to Generate and Exchange Secrets. FOCS Conference, 1986.

[17] Chaum D., Crepeau C., Damgard I.: Multiparty unconditionally secure protocols. ACM STOC Conference, 1988.

[18] Clifton C., Kantarcioglou M., Lin X., Zhu M.: Tools for privacy preserving distributed data mining. ACM SIGKDD Explorations, 4(2), 2002.

[19] Du W., Atallah M.: Secure Multi-party Computation: A Review and Open Problems.CERIAS Tech. Report 2001-51, Purdue University, 2001.