ISSN No. 0976-5697

Volume 5, No. 8, Nov-Dec 2014

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

A Summarization on Preserving Privacy Techniques in Data Mining

Sidra Anam[1], Saurabh Gupta[2]
[1]M.Tech Scholar, [2]Ass. Prof
CSE Department, Pranveer Singh Institute of Technology, Kanpur

Abstract: Data Mining is an analysis process of large quantities of data in order to discover meaningful patterns and rules. Privacy Preservation in Data Mining is designed to reduce the gap between Data Confidentiality and Data Mining. In recent years with the explosive development and vast advancements in internet, data processing and data storage technologies, privacy preserving is becoming an important issue for concern. A number of methods and techniques have been developed for privacy preserving data mining. In this paper, we provide a classification and description of the various techniques and methodologies that have been developed in the area of privacy preserving data mining.

Keywords— Data Mining, Privacy Preservation, Data Confidentiality.

I. INTRODUCTION

In today's scenario, Data mining is primarily used by the companies with a strong focus on retail, financial, marketing organization. Organizations record every single transaction. The resulting data sets can consist of terabytes of data, so efficiency and scalability is the primary consideration of most data mining algorithms. Naturally, ever-increasing data collection, along with the influx of analysis tools capable of handling huge volumes of information, has led to privacy concerns [1]. Protecting private data is an important concern for almost all the organizations. No company is ready to broadcast its sensitive information, but its importance is not limited as corporations might also need to protect their information's privacy, even though sharing it for analysis could benefit the company. Clearly, the trade-off between sharing information for analysis and keeping it secret to preserve corporate trade secrets and customer privacy is a growing challenge. Over the years a number of definitions privacy preserving has emerged. One of them defines "privacy preserving as the individual's ability to control the circulation of information relating to him/her". Privacy preserving also define as "the claim of individuals, groups, or institutions to determine for themselves when, how & what extent information about them is communicated to others". The main objective in privacy preserving data mining is to develop algorithms for modifying the original data in such a way that the private data remain private even after the mining process.

II. CLASSIFICATION OF PRIVACY PRESERVING TECHNIQUES

There are different approaches which have been adopted for privacy preserving data mining. They are as follows:

A. Data distribution:

For this approach, two models have been proposed for privacy preserving data mining, they are as follows:

a. Centralized Model:

Here all the data are owned by single data publisher. The key issues are how to modify the data and how to recover data mining result from modified data. Techniques which are basically used in centralized model are Randomization and Encryption. But major disadvantage of this model is that the load is entirely on a single site which manages all the tasks of data mining. In case of failure of this central site, all the important data can be at the risk of losing it.

b. Distributed Model:

Here multiple data publishers want to conduct a computation based on their private inputs, but no one is willing to disclose its own output to anybody else. The problem is known as Secure Multiparty Communication (CMC). The following figure shows the models of Privacy-Preserving Data Mining (PPDM) algorithms:

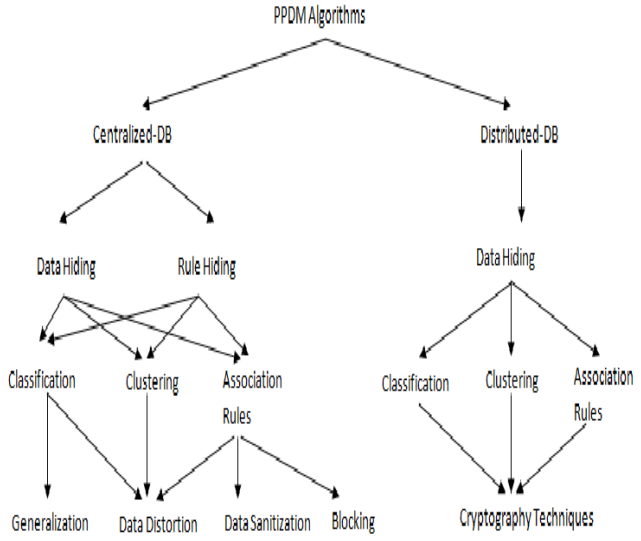

Figure 1: PPDM Algorithms

The data may be distributed in two ways across different sites:

c. Horizontal Partitioning:

Here, different sites may have different sets of records containing the same attributes i.e., the data is partitioned horizontally and stored at different sites.

d. Vertical Partitioning:

Here, different sites may have different attributes of the same sets of records i.e., the data is partitioned vertically and stored at different sites. In many cases, the data may be distributed across multiple sites, and the owners of the data across these different sites may wish to compute a common function. Then in such 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.

B. Data Modification:

These techniques study the different transformation methods that are associated with privacy. These techniques include methods such as randomization [2] and k-anonymity [3, 4]. Another related issue is how the perturbed data can be used in conjunction with classical data mining methods such as association rule mining [5].

a. The Randomization Method:

The randomization technique uses data distortion methods to create private representations of the records [1, 6], which can then be used for mining purpose. Certain type of noise or a pattern is added to the data so that original data cannot be read by unauthorized party. Generally, the individual records of company cannot be recovered, but only aggregate distributions can be recovered. These aggregate distributions can be efficiently used for data mining purposes. Secure multiparty communication and Data Perturbation approaches are used for protecting sensitive data during multiparty privacy preserving data mining. Here Data Perturbation means hiding the private data while still mining patters. It is important to consider the fact that randomization is done in such a way that the data can be used in synchrony with classical data mining methods such as association rule mining [6]. Two kinds of perturbation are possible with the randomization method:

b. Additive Perturbation:

Here, a randomized noise is added to the data records. The overall data distributions can be recovered from the randomized records. Data mining and management algorithms re designed to work with these data distributions.

c. Multiplicative Perturbation:

Here, random projection or random rotation techniques are used in order to perturb the records. The advantage of using randomization method lies in its simplicity as it does not need to know the knowledge of other records in the data, which is not the case of other methods such as k-anonymity. Therefore, randomization can be applied to at data collection time without the use of a trusted server containing all original records [7].

d. The k-anonymity Method:

The k –anonymity is a method used for privacy de-identification [3]. In this technique many attributes in the data can often be considered pseudo-identifiers, which can be used in conjunction with public records in order to uniquely identify the records. If the identifications from the records are removed, we can still identify the records. For example, attributes such as the birth date and zip code can be used in order to uniquely identify the identities of the underlying records. The idea is to reduce the granularity of representation of the data in such a way that a given record cannot be distinguished from at least (k - 1) other records.

This method requires the knowledge of other records in the data [7].

C. Data Encryption:

In many cases, multiple parties may wish to share aggregate private data, without leaking any sensitive information at their end [8]. For example, many stores having sensitive sales data may wish to coordinate among themselves for knowing aggregate trends without leaking the data patterns of individual stores. This requires secure and cryptographic protocols for sharing the information across the different parties. It includes encryption techniques. At sender site, either the data or mined result is first encrypted using encryption algorithms and send over the network. At receiver side, it is decrypted using decryption algorithm.

D. Data Hiding:

Sometimes the results of data mining applications such as association rule or classification rule mining can also compromise the privacy of individual data. So, results of association rule mining are modified. An example of such techniques is association rule hiding methods, in which some of the association rules are suppressed in order to preserve privacy.

E. Data Reconstruction:

It is important to consider the fact that data should be modified in such a way that original patterns can be reconstructed from the modified data. Their exists reconstruction techniques where the original distribution of the data is reconstructed from the randomized data.

III. PRIVACY PRESERVING ALGORITHMS

The algorithms and techniques based on above approaches of privacy preserving are as follows:

A. The Randomization Method:

Let each client C_i, i = 1, 2, . . . ,N, have a numerical attribute x_i. Assume that each xi is an instance of random variable X_i, where all X_i are independent and identically distributed. The cumulative distribution function (the same for every X_i) is denoted by F_X. The server wants to learn the function F_X, or its close approximation; this is the aggregate model which the server is allowed to know. The server can know anything about the clients that is derivable from the model, but we would like to limit what the server knows about the actual instances x_i [9]. The solution is as follows: Each client randomizes its x_i by adding to it a random shift y_i. The shift values y_i are independent identically distributed random variables with cumulative distribution function F_Y ; their distribution is chosen in advance and is known to the server. Thus, client C_i sends randomized value $z_i = x_i + y_i$ to the server, and the server's task is to approximate function F_X given F_Y and values $z_1, z_2, \ldots, z_N$.

More simplified method of Randomization is as follows: Consider a set of data records denoted by $X = \{x_1 \ldots x_N\}$. For record $x_i \in X$, we add a noise component which is drawn from the probability distribution f_Y (y). These noise components are drawn independently, and are denoted $y_1 \ldots y_N$. Thus, the new set of distorted records are

denoted by $x_1 + y_1 \ldots x_N + y_N$. We denote this new set of records by $z_1 \ldots z_N$. In general, it is assumed that the variance of the added noise is large enough, so that the original record values cannot be easily guessed from the distorted data. Thus, the original records cannot be recovered, but the distribution of the original records can be recovered. 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:

$Z = X + Y$

$X = Z - Y$

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 as those discussed in [10,11].

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. This is not true of other methods such as k-anonymity which require the knowledge 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.

B. Privacy Quantification:

The quantity used to measure privacy should indicate how closely the original value of an attribute can be estimated. The work in [10] uses a measure that defines privacy as follows: If the original value can be estimated with c% confidence to lie in the interval $[\alpha_1, \alpha_2]$, then the interval width $(\alpha_2 - \alpha_1)$ defines the amount of privacy at c% confidence level. For example, if the perturbing additive is uniformly distributed in an interval of width 2α, then α is the amount of privacy at confidence level 50% and 2α is the amount of privacy at confidence level 100%. However, this simple method of determining privacy can be subtly incomplete in some situations.

C. Group Based Anonymization Method:

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 data records. This is also a weakness because outlier records can often be difficult to mask. Clearly, in cases in which the privacy-preservation does not need to be performed at data-collection time, it is desirable to have a technique in which the level of inaccuracy depends upon the behavior of the locality of that given record. So, a framework named k-Anonymity Framework was introduced. In many applications, the data records are made available by simply removing key identifiers such as the name and social-security numbers from personal records. However, other kinds of attributes (known as pseudo-identifiers) can be used in order to accurately identify the records. For example, attributes such as age, zip-code and sex are available in public records such as census rolls. When these attributes are also available in a given data set, they can be used to infer the identity of the corresponding individual. A combination of these attributes can be very powerful, since they can be used to narrow down the possibilities to a small number of individuals. In k-anonymity techniques, we reduce the granularity of representation of these pseudo-identifiers with the use of techniques such as generalization and suppression. In the method of generalization, the attribute values are generalized to a range in order to reduce the granularity of representation. For example, the date of birth could be generalized to a range such as year of birth, so as to reduce the risk of identification. In the method of suppression, the value of the attribute is removed completely. It is clear that such methods reduce the risk of identification with the use of public records, while reducing the accuracy of applications on the transformed data.

D. Heuristic-Based Techniques:

a. Centralized Data Perturbation-Based Association Rule Confusion:

A formal proof that the optimal sanitization is an NPHard problem for the hiding of sensitive large itemsets in the context of association rules discovery, have been given in [12]. The specific problem which was addressed in this work is the following one. Let D be the source database, R be a set of significant association rules that can be mined from D, and let R_h be a set of rules in R. How can we transform database D into a database D, the released database, so that all rules in R can still be mined from D, except for the rules in R_h. The heuristic proposed for the modification of the data was based on data perturbation, and in particular the procedure was to change a selected set of 1-values to 0-values, so that the support of sensitive rules is lowered in such a way that the utility of the released database is kept to some maximum value. The utility in this work is measured as the number of non-sensitive rules that were hidden based on the side-effects of the data modification process.

b. Centralized Data Blocking-Based Classification Rule Confusion:

The work in [13] provides a new framework combining classification rule analysis and parsimonious downgrading. Here in the classification rule framework, the data administrator, has as a goal to block values for the class label. By doing this, the receiver of the information is unable to build informative models for the data that is not downgraded. Parsimonious downgrading is a framework for formalizing the phenomenon of trimming out information from a data set for downgrading information from a secure environment to a public one. Classification rules, and in particular decision trees are used in the parsimonious downgrading context in analyzing the potential inference channels in the data that needs to be downgraded.

E. Cryptography-Based Techniques:

Consider a scenario in which two or more parties owning confidential databases wish to run a data mining algorithm on the union of their databases without revealing any unnecessary information. For example, consider separate medical institutions that wish to conduct a joint

research while preserving the privacy of their patients. In this scenario it is required to protect privileged information, but it is also required to enable its use for research or for other purposes. In particular, although the parties realize that combining their data has some mutual benefit, none of them is willing to reveal its database to any other party. Note that we consider here a distributed computing scenario, rather than a scenario where all data is gathered in a central server, which then runs the algorithm against all data. In the multi-party scenario, there are protocols that enable the parties to compute any joint function of their inputs without revealing any other information about the inputs. That is, compute the function while attaining the same privacy as in the ideal model. This was shown to be possible in principle by Goldreich, Micali and Wigderson [14], Ben-Or, Goldwasser and Wigderson [15], and by Chaum, Crepau and Damgard [16], for different scenarios. These constructions, too, are based on representing the computed function as a circuit and evaluating it. The constructions do have, however, some additional drawbacks, compared to the two-party case:

a. The computation and communication overhead of the protocol is linear in the size of the circuit, and the number of communication rounds depends on the depth of the circuit1, unlike the two-party case where the number of rounds is constant. Furthermore, the protocol that is run for every gate of the circuit is more complex than the computation of a gate in the two-party case, especially in the malicious party scenario, and requires public-key operations (although the overhead is still polynomial).
b. The multi-party protocols require each pair of parties to exchange messages (in order to compute each gate of the circuit). The required communication graph is, therefore, a complete graph, whereas a sparse communication graph could have been sufficient if no security was required. In many applications, for example applications run between a web server and many clients, it is impossible to require all pairs of parties to communicate.
c. The security of the multi-party protocols is assured as long as there is no corrupt coalition of more than one half or one third of the parties (depending on the scenario). In many situations, however, it is impossible to ensure that the number of corrupt parties is smaller than such a threshold (for example, consider a web application in which anyone can register and participate, and which, therefore, enables an adversary to register any number of corrupt participants). In such cases the security of the protocol is not guaranteed.

These drawbacks prevent most applications from using the generic solutions for secure distributed computation.

IV. CONCLUSION AND FUTURE SCOPE

This paper summarizes the different dimensions that can exist for the approaches used for preserving privacy in data mining. Also the algorithms for the same are summarized. Conclusions that we have reached from reviewing this area of privacy preservation in data mining shows that privacy issues can be effectively considered only within the limits of certain data mining algorithms. The inability to generalize the results for classes of categories of data mining algorithms might be a tentative threat for disclosing information. The future work can be based upon the evaluation of these privacy preserving algorithms based on the data utility, uncertainty level and resistance accomplished by some privacy algorithms to different data mining techniques.

V. REFERENCES

[1] Jaideep Vaidya, Chris Clifton, "Privacy-Preserving Data Mining: Why, How, and When", IEEE Security and Privacy, vol. 2, no. 6, pp. 19-27, November-December, 2004.

[2] Agrawal R., Srikant R., "Privacy-Preserving Data Mining", ACM SIGMOD Conference, 2000.

[3] Samarati P., Sweeney L., "Protecting Privacy when Disclosing Information: k-Anonymity and its Enforcement Through Generalization and Suppression", IEEE Symp. on Security and Privacy, 1998.

[4] Bayardo R. J., Agrawal R., "Data Privacy through optimal kanonymization", ICDE Conference, 2005.

[5] Rizvi S., Haritsa J., "Maintaining Data Privacy in Association Rule Mining", VLDB Conference, 2002.

[6] Agrawal D. Aggarwal C.C., "On the Design and Quantification of Privacy Preserving Data Mining Algorithms", ACM PODS Conference, 2002.

[7] V.Thavavel, S.Sivakumar, "A generalized Framework of Privacy Preservation in Distributed Data mining for Unstructured Data Environment, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 2, January 2012 ISSN (Online): 1694-0814.

[8] PinkasB., "Cryptographic Techniques for Privacy-Preserving Data Mining", ACM SIGKDD Explorations, 4(2), 2002.

[9] Alexandre Evfimievski, "Randomization in Privacy Preserving Data Mining", Cornell University Ithaca, NY 14853, USA.

[10] Agrawal R., Srikant R. Privacy-Preserving Data Mining. Proceedings of the ACM SIGMOD Conference, 2000.

[11] Agrawal D. Aggarwal C. C. On the Design and Quantification of Privacy-Preserving Data Mining Algorithms. ACM PODS Conference, 2002.

[12] Mike J. Atallah, Elisa Bertino, Ahmed K. Elmagarmid, Mohamed Ibrahim, and Vassilios S. Verykios, "Disclosure Limitation of Sensitive Rules", In Proceedings of the IEEE Knolwedge and Data Engineering Workshop (1999), 45–52.

[13] LiWu Chang and Ira S. Moskowitz, "Parsimonious downgrading and decision trees applied to the inference problem", In Proceedings of the 1998 New Security Paradigms Workshop (1998), 82–89.

[14] O. Goldreich, S. Micali and A. Wigderson, "How to Play any Mental Game - A Completeness Theorem for Protocols with Honest Majority", Proceedings of the 19th Annual Symposium on the Theory of Computing (STOC), ACM, 1987, pp. 218–229.

[15] M. Ben-Or, S. Goldwasser and A. Wigderson, "Completeness theorems for non cryptographic fault tolerant distributed computation", Proceedings of the 20th

Annual Symposium on the Theory of Computing (STOC), ACM, 1988, pp. 1–9.

[16] D. Chaum, C. Crepeau and I. Damgard, "Multiparty unconditionally secure protocols", Proceedings of the 20th Annual Symposium on the Theory of Computing (STOC), ACM, 1988, pp. 11–19.