IJCSN International Journal of Computer Science and Network, Volume 6, Issue 5, October 2017
ISSN (Online) : 2277-5420
www.IJCSN.org
Impact Factor: 1.5

An Extensive Survey of Privacy Preserving Data Mining Techniques

[1] Bhargav Sundararajan; [2] Deepthi Peri; [3] Nita Radhakrishnan; [4] Mehul Awasthi

[1] SRM University, Chennai, Tamil Nadu 603203, India

[2] SRM University, Chennai, Tamil Nadu 603203, India

[3] SRM University, Chennai, Tamil Nadu 603203, India

[4] SRM University, Chennai, Tamil Nadu 603203, India

Abstract - Data mining techniques are rising trends to aid organizations to analyze, find un-obvious patterns and details to benefit from the customer or user data. But this is classified as proprietary information disclosure and mining misuse. To avoid this, we introduce the concept of privacy preserving data mining (PPDM). The fundamental notions of the existing privacy preserving data mining methods, their merits, and shortcomings are presented. We discus five techniques namely Anonymization based PPDM, Perturbation based PPDM, Randomized response based PPDM, Condensation based PPDM and Cryptography based PPDM.

Keywords – *privacy, data mining, anonymization, perturbation.*

1. Introduction

Data mining is the task of creating aggregate models from the available data and hence it requires a vast amount of precise data for the same. The biggest challenge in these mining tasks is to ensure the privacy of the acquired data [1]. Varied internet services such as electronic commerce, online-banking, research, and online trade exploiting both human and software vulnerabilities, are highly sensitive as they involve personal and financial data. The corporations have been widely proliferating this data and hence data mining has been viewed as a threat to privacy. Therefore, enhanced privacy preserving data mining methods are in high demand for secured and reliable information exchange over the internet. This led to the advent of highly complex data mining algorithms, enabling secure information sharing. Privacy preserving data mining is one of the most efficient algorithms that were crafted to this cause. Truly, the privacy must protect all the three mining aspects including association rules, classification, and clustering. The problems faced in data mining are widely deliberated in many communities such as the database, the statistical disclosure control and the cryptography community. Currently, several privacy preservation methods for data mining are available. These include K-anonymity, classification, clustering, association rule, distributed privacy preservation, L-diverse, randomization, taxonomy tree, condensation, and cryptographic [7] . The PPDM methods protect the data by changing them to mask or erase the original sensitive one to be concealed. Typically, they are based on the concepts of privacy failure, the capacity to determine the original user data from the modified one, loss of information and estimation of the data accuracy loss. There are several challenges also in PPDM. The major challenges of PPDM method for association rule hiding are high information loss, expensive, difficult to recover original data after hiding and should be efficient enough for very large datasets.

In this paper, we provide a general overview of the privacy preserving data mining algorithms available and briefly discuss each one in detail.

2. Privacy in Data Mining

Perception of privacy in data mining requires comprehension of how privacy can be violated and the possible means for preventing privacy violation. In general, one major factor contributes to privacy breach in data mining is the misuse of data. User's privacy can be violated in different ways and with different intentions. Although data mining can be extremely valuable in many

applications (e.g., business, medical analysis, etc.), it can also, in the absence of adequate safeguards, violate informational privacy. Privacy can be violated if personal data is used for other purposes subsequent to the original transaction between an individual and an organization when the information was collected [2]. There are several known methods of carrying out this data accumulation. For example, Data magnets are techniques and tools used to collect personal data. Examples of data magnets include explicitly collecting information through on-line registration, identifying users through IP addresses, software downloads that require registration, and indirectly collecting information for secondary usage. In many cases, users may or may not be aware that information is being collected or do not know how that information is collected.

Privacy preservation can be classified into two domains namely individual privacy preservation and collective privacy preservation. Individual privacy protection refers to the safeguarding of personally identifiable information. In general, information is considered personally identifiable if it can be linked, directly or indirectly, to an individual person. Thus, when personal data are subjected to mining, the attribute values associated with individuals are private and must be protected from disclosure. On the other hand, collective privacy protection refers to the protection of sensitive knowledge representing the activities of group.

3. Privacy Preserving Data Mining

Privacy Preserving Dara Mining (PPDM) is a method used to ensure that the privacy of a particular individual is not compromised during the process of data mining [3]. PPDM is not a one step process. It needs to be applied during the whole process of data mining. Data mining can be broadly classified into three steps which are, data acquisition, application of data mining algorithm and interpretation of information. Hence, PPDM must be applied in each of these steps in order to ensure maximum preservation of an individual's privacy. The framework of PPDM is illustrated in fig. 1. As we can see from the framework, there are three stages in PPDM which are applied for the three steps in the data mining process respectively [6]. The first stage of PPDM is applied during the data acquisition step. In stage 1, it is mandatory to make sure that private information such as mobile numbers, addresses etc. are not acquired without the consent of that particular individual. Stage 2 of PPDM is applied when the collected data is being mined by an algorithm. Here, various processes are applied to make the data sanitized so that it can be revealed to even untrustworthy data miners. At level 3, the information or knowledge so revealed by the data mining algorithms is checked for its sensitiveness towards disclosure risks.

Fig. 1 PPDM Framework

4. PPDM Techniques

There are mainly two methodologies in data mining that in turn aided researchers to come up with various techniques for PPDM. Which are,

- Methodologies that protect the sensitive data itself in the mining process.
- Methodologies that protect the sensitive data mining results (i.e. extracted knowledge) that were produced by the application of the data mining.

The first methodology refers to application of techniques such as perturbation, sampling, generalization, etc. in order to sanitize the collected data so that it can be shared to untrustworthy parties for data mining. Whereas, the second methodology involved techniques that prohibits the information gained from the data mining process to fall into the hands of malicious people.

Based on these methodologies, PPDM techniques can be broadly divided into these 5 categories [4].

- Anonymization based PPDM
- Perturbation based PPDM
- Randomized response based PPDM
- Condensation based PPDM
- Cryptography based PPDM.

4.1 Anonymization based PPDM

Anonymization refers to an approach where identity or/and sensitive data about record owners are to be hidden. Explicit identifiers such as name and social security numbers are usually removed from the collected data. However, this alone would not be enough as quasi identifiers can also be linked with publicly available data to obtain sensitive information. These type of attacks are called linking attacks. For example attributes such as DoB, Sex, Race, Zip are available in public records such as voter list. Such records are available in medical records also, when linked, can be used to infer the identity of the corresponding individual with high probability. For

tackling this, a *k*-anonymity model using generalization and suppression to achieve *k*-anonymity can be used i.e. any individual is distinguishable from at least *k*-1 other ones with respect to quasi-identifier attribute in the anonymized dataset. In other words, we can define a table as *k*-anonymous if the Q1 values of each tuple are identical to those of at least *k*- 1 other tuples. Replacing a value with less specific but semantically consistent value is called as generalization and suppression involves blocking the values. Releasing such data for mining reduces the risk of identification when combined with publically available data.

4.2 Perturbation based PPDM

Perturbation is the swapping of the original values with synthesized values at the same time preserving the statistical information of the original values. This means that, the computed statistical information of the synthesized values should we similar if not exact to the one computed using original values. The perturbed data are sometimes meaningless and do not correspond to real-world record owners. Hence, an attacker cannot use linking attacks or perform sensitive linkages to the published data. Perturbation can be done by using additive noise or data swapping or synthetic data generation. Since the perturbation method does not reconstruct the original values but only the distributions, new algorithms are to be developed for mining of the data. This means that a new distribution based mining algorithms need to be developed for each individual data problems like classification, clustering or association rule mining.

4.3 Randomized response based PPDM

Randomization is considered as one of the frequently used approaches in PPDM research. Tis method involves adding noise on to the authentic data for creating values of each record [5]. Perturbation mixed with authentic data are sufficiently huge for maintaining the privacy and hence one cannot recover the actual data. Randomized Response scheme and random-noise-based perturbation help Randomization techniques to achieve both knowledge discovery and privacy preservation. Although involving huge loss of information, this technique is comparatively a better and efficient process. Randomization proves to have the ability of preserving some semantics and anonymizing the entire dataset [8]. Among the currently used privacy preserving data mining methods. Randomization is treated as the crucial method. Harmony between utility and privacy as well as knowledge discovery are provided by this. After being balanced, the randomized data gets transmitted to the concerned recipient. Using distribution reconstruction algorithm, the recipient will receive the data. Tis method overs effective and simple way for ensuring the person's privacy and also preserve is used of data to some extent.

4.4 Condensation based PPDM

Condensation approach constructs constrained clusters in dataset and then generates pseudo data from the statistics of these clusters. It is called as condensation because of its approach of using condensed statistics of the clusters to generate pseudo data [9]. It constructs groups of non-homogeneous size from the data, such that it is guaranteed that each record lies in a group whose size is at least equal to its anonymity level. Subsequently, pseudo data is generated from each group so as to create a synthetic data set with the same aggregate distribution as the original data. The use of pseudo-data provides an additional layer of protection, as it becomes difficult to perform adversarial attacks on synthetic data. This approach helps in better privacy preservation as compared to other techniques as it uses pseudo data rather than modified data. Moreover, it works even without redesigning data mining algorithms since the pseudo data has the same format as that of the original data. At the same time, data mining results get affected as large amount of information is lost because of the condensation of a larger number of records into a single statistical group entity.

4.5 Cryptography based PPDM

Cryptography is one method used to preserve sensitive data. Cryptographic technique is very much favored as it o□ers safety and security of sensitive attributes, and was suggested by authors in. The privacy of a person's record may be broken by final data mining. Consider for example a situation in which several medical institutions look for conducting a joint study on the datasets for certain mutual benefits, while not disclosing unwanted information [10]. It is possible that sometimes when a data mining algorithm is passed to a dataset formed by combining two data sets there is some possibility that the results may disclose private information about the individual. But, this kind of leakage is inevitable. Oblivious transfer is method used in cryptography based PPDM to ensure that if two are more parties are collectively working on a data mining algorithm to mine their own respective data sets, a function is generated such that no other party with get their hands on sensitive information of another.

5. Conclusion

The primary objective of PPDM is promoting algorithm to conceal sensitive data or over privacy. These sensitive data do not get revealed to unapproved parties or invader. In data mining there exists a trade of between utility and privacy of data. When we accomplish one it inevitably leads to the detrimental impact on the other. Many PPDM techniques in existence are reviewed in the paper. Ultimately, it is concluded with the fact that there is no single PPDM technique in existence that outshines every other technique with relation to each possible criteria such as use of data, performance, difficulty, compatibility with procedures for data mining, and so on. A particular algorithm may function better when compared to another, on a specific criterion. Various algorithms may be found to function better than one another on given criterion. Researchers are doing extensive research in ensuring that the sensitive data of a person is not revealed as well as not compromising the utility of data so that the data can be useful for many purposes.

References

[1] Ann Cavoukian, Information and Privacy Commissioner, Ontario, "*Data Mining Staking a Claim on Your Privacy*", 1997.

[2] The Economist. "*The End of Privacy*", May 1st, 1999. pp: 15.

[3] R. Agrawal and R. Srikant. "*Privacy Preserving Data Mining*", ACM SIGMOD Conference on Management of Data, pp: 439-450, 2000.

[4] D. Agrawal and C. Aggarwal, "*On the Design and Quantification of Privacy Preserving Data Mining Algorithms*", PODS 2001. pp: 247-255.

[5] W. Du and Z. Zhan, "*Using Randomized Response Techniques for Privacy Preserving Data Mining",* SIGKDD 2003. pp. 505-510.

[6] Elisa, B., N.F. Igor and P.P. Loredana. "*A Framework for Evaluating Privacy Preserving Data Mining Algorithms*", Published by Data Mining Knowledge Discovery, 2005, pp.121- 154.

[7] Sweeney L, *"Achieving k-Anonymity privacy protection using generalization and suppression"* International journal of Uncertainty, Fuzziness and Knowledge based systems, 10(5), 571- 588, 2002.

[8] Evfimievski A., "Randomization in Privacy-Preserving Data Mining", ACM SIGKDD Explorations, 4, 2003.

[9] Aggarwal C, Philip S Yu, *"A condensation approach to privacy preserving data mining"*, EDBT, 183-199, 2004.

[10] Benny Pinkas,*"Cryptographic Techniques for Privacy preserving data mining"*, SIGKDD Explorations, Vol. 4, Issue 2, 12-19, 2002.