Privacy Preserving Data Mining: A Parametric Analysis

Darshana Patel and Radhika Kotecha

Abstract With technological revolution, a huge amount of data is being collected and as a consequence the need of mining knowledge from this data is triggered. But, data in its raw form comprises of sensitive information and advances in data mining techniques have increased the privacy breach. However, due to socio-technical transformations, most countries have levied the guidelines and policies for publishing certain data. As a result, a new area known as Privacy Preserving Data Mining (PPDM) has emerged. The goal of PPDM is to extract valuable information from data while retaining privacy of this data. The paper focuses on exploring PPDM in different aspects, such as types of privacy, PPDM scenarios and applications, methods of evaluating PPDM algorithms etc. Also, the paper shows parametric analysis and comparison of different PPDM techniques. The goal of this study is to facilitate better understanding of these PPDM techniques and boost fruitful research in this direction.

Keywords Data mining · Privacy · Privacy preserving data mining techniques

1 Introduction

Internet has brought a drastic change in today's world. The Internet today is a widespread information infrastructure presently allowing many people all over the globe to communicate and better understand each other. Such information is stored in huge databases. Knowledge discovery from such databases is the aim of data mining and is attracting researchers vastly. This knowledge discovery has resulted

D. Patel (✉) · R. Kotecha
V.V.P. Engineering College, Rajkot, Gujarat, India
e-mail: darshana.h.patel@gmail.com

R. Kotecha
e-mail: kotecha.radhika7@gmail.com

S.C. Satapathy et al. (eds.), *Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications*, Advances in Intelligent Systems and Computing 516, DOI 10.1007/978-981-10-3156-4_14

in a remarkable increase in the disclosure of private information about individuals. As a result, preserving privacy has become an active research area. Privacy preservation implies that an access to the published data should not enable the adversary to learn anything extra about any target victim as compared to having no access to the database, even with the presence of any adversary's background knowledge obtained from other sources [1]. Thus, Privacy preserving data mining (PPDM) deals with hiding sensitive information of individuals like names, age, zip code etc. without compromising the usability of data [2, 3]. An efficient privacy preserving data mining technique must ensure that any information disclosed should not: (1) be traced to a specific individual, and, (2) form an intrusion.

1.1 Types of Privacy

Privacy is a valuable characteristic of a person. It is the fundamental right of every human being and needs to be preserved. Following are the types of privacy [4]:

(a) Information Privacy: Entails the establishment of rules governing the gathering and managing of private data such as credit information and medical records.
(b) Bodily privacy: Concerns the fortification of people's physical identity in opposition to invasive procedures such as drug testing and cavity searches.
(c) Privacy of communications: Protections of messages, telephones, email and other forms of communication.
(d) Territorial privacy: Involves the setting of confinement on infringing of the domestic and other environments such as the place of work or civic space.

1.2 Different Scenarios in PPDM

With the advance of the information age, data gathering and data investigation have exploded out both in magnitude and complications. Consequently, there arises a need of data sharing amongst stakeholders. For sharing data in privacy preserving, considering in broader aspect, following two different scenarios exist specifically in PPDM [2]:

(a) Central Server Scenario: Firstly, data owner conceals the micro-data by applying various PPDM techniques before publishing it to the data miner which then performs different data mining tasks on such concealed data. In this scenario, data owners/data miners are independent of managing privacy issues. It is also referred as Data Publishing scenario.

(b) Distributed Scenario: The data owners can also be the miners and get collective outcomes on the amalgamation of their records. This is a situation where the privacy is ensured on results of data mining. Distributed scenario can further be classified into three following different models [3, 5]:

(1) Trust Third Party Model: In such kind of model, each and every party gives the data to a trusted third party keeping blind faith on it. On the other hand, this trusted third party carries out the computation and conveys only the outcomes. However, any such reliable third party does not exist, so this is a superlative model.
(2) Semi-honest Model: In the semi-honest model, every party pursue the policy of the etiquette using its truthful put in, but may try to interpret facts from the data interchange process.
(3) Malicious Model: In this sort of model, no restrictions are placed on any of the contributors. As a consequence, any party is completely free to treat in whatever way it wants. Thus, it becomes quite complicated to conduct data mining under the malicious model.

2 Applications of Privacy Preserving Data Mining

Privacy-preserving data mining studies techniques for assembling the potentially contradictory objectives with regard to human privileges and allowing applicable organization to bring together and excavate it for massive data sets. This technique has been consumed in plentiful application areas, some are listed below [6, 7]:

(a) Clinical Records: Registration of patients in clinics leads to collection of enormous amount of clinical records. Such kinds of data are shared with the scientific and government organizations for research and development. These clinical records can be anonymized before publishing the records with other firms so that privacy of an individual is preserved.
(b) Cyber Terrorism/Social Networking: Internet is very famous at the present time for pleasing people with a variety of services related to different fields. But, people are oblivious of the privacy issues. The published data in social network can be anonymized such that it does not infer to any one individual.
(c) Homeland security applications: Some examples of such applications are as follows: identity theft, credential validation problem, web camera surveillance, video-surveillance, and watch list problem.
(d) Government and Financial companies: The government and financial companies are required to publish the data to civilians. If such data is sanitized then no one could reveal any identity and lead to self humiliation.

(e) Business: Enormous data is generated by piling up data from daily shopping of customers. Also, the retailers may share data with other retailers for mutual benefit. In such cases, various anonymization methods can be applied on the data to ensure the privacy of each individual.
(f) Surveillance: Raw samples gathered from surveillance may end up in the hands of adversary. These adversaries could forecast the description of biometric qualities of an individual stored in the database and takes benefit of the existing feature. Privacy preserving algorithms can be adapted and deployed on the databases storing raw data of each individual.

3 Evaluation of Privacy Preserving Data Mining Algorithms

Privacy preserving data mining, the recognition of appropriate evaluation principles and the development of related standards is an important aspect in the expansion and evaluation of algorithms. It is thus vital to provide abusers with a set of metrics which will facilitate them to select the most suitable privacy preserving technique for the data at furnish; with respect to some specific consideration they are interested in optimizing. The chief list of evaluation parameters that are used for evaluation of privacy preserving data mining algorithms are given below [3, 6]:

(a) Performance: Performance of PPDM algorithms is evaluated in terms of time taken to preserve the private attributes.
(b) Data Utility: Data utility is basically a measure of information loss or loss in the functionality of data in supplying the results, which could be created without PPDM algorithms.
(c) Uncertainty level: It measures uncertainty with which the sensitive information that has been concealed can still be forecasted.
(d) Resistance: It depicts the measure of acceptance shown by PPDM algorithm against various data mining algorithms.

4 Related Work

4.1 Different Privacy Preserving Techniques

Varied forethoughts from an attacker could lead to an information disclosure of a particular user. Therefore, one needs to limit this disclosure risks to an acceptable level while maximizing data efficiency before releasing original data for research

and analysis purposes. To limit such risks, modification of data is done by applying variety of operations to the original stuff [8, 9]. They can broadly be classified into three categories mainly: Anonymization, Randomization and Cryptography.

(a) Anonymization: Anonymization aims to make individual record impossible to differentiate among a group of records by using following techniques:
    (1) Generalization and Suppression: Replaces definite values from the original data [10, 11].
    (2) Anatomization and Permutation: De-associates the correlation between attributes [1].
    (3) Perturbation: Distorts the data by adding noise [12].
    (4) Bucketization: Separates the SAs (Sensitive attributes) from the QIs (Quasi-Identifiers) by arbitrarily permuting the SA values in each bucket [13].
    (5) Microaggregation: Unified approach consisting of partition and aggregation [14, 15].

(b) Randomization: Applied generally to provide estimates for data distributions rather than single point estimates. It distorts values of each attribute in a sample independently and doesn't use information about other samples [16].

(c) Cryptography: Cryptographic techniques are ideally meant for multiparty scenarios where most frequently used protocol is SMC (Secure Multiparty Computation) which mainly constitutes secure sum, secure union, secure intersection, etc. operations [17, 18].

4.2 Types of Attack Models

Privacy preserving data mining is a significant asset that any mining system must satisfy. User's data is considered to be effectively protected when an opponent could not fruitfully recognize a particular user's data through linkages between an record owner to sensitive attribute in the published data. Thus, these linkage attacks can be classified broadly into three types of attack models namely Record Linkage, Attribute Linkage and Table Linkage [1, 19]. In all these three types of attacks or linkage, it is assumed that opponent knows the QIs (Quasi-identifiers) of the victim.

(a) Record Linkage: If an opponent is capable to link a record holder to a record in a published data table then such kind of privacy risk is known as record linkage. It is assumed that the adversary can make out the victim's record is in the released table and tries to recognize the victim's record from the table. To prevent record linkage privacy models such as k-Anonymity, MultiR

k-Anonymity, l-Diversity, (α, k)-Anonymity, (X, Y)-privacy, (c, t)-Isolation, etc. can be used [1].

(b) Attribute Linkage: If an opponent is capable to link a record holder to a sensitive attribute in a published data table then such kind of privacy risk is known as attribute linkage. It is assumed that the adversary can make out the victim's record is in the released table and tries to recognize the victim's sensitive information from the table. To prevent attribute linkage privacy models such as l-Diversity, Confidence Bounding, (α, k)-Anonymity, (X, Y)-privacy, (k, e)-Anonymity, (ε, m)-Anonymity, Personalized Privacy, t-closeness, etc. can be used [1].

(c) Table Linkage: If an opponent is capable to link a record holder to the published data table then such kind of privacy risk is known as table linkage. In this scenario, the attacker seeks to determine the occurrence or nonappearance of the victim's record in the released table. To prevent table linkage privacy models such as δ-presence, ε-Differential privacy, (d, γ)-privacy and distributional privacy can be used.

4.3 Data Mining Tasks

Privacy preserving [20] has been extensively studied by the Data Mining community in recent years. Currently, the PPDM algorithms are mainly used on the functionality of data mining such as Classification, Association and Clustering [21–23].

(a) Classification: Classification is the process of finding a set of model that differentiates data classes such that the model can be used to predict the class of objects whose class label is unknown. The common algorithms used for Classification are Decision Tree, Bayesian classification, etc.

(b) Association Rule: It is a technique in data mining that recognizes the regularities found in bulky amount of data. Such a technique may identify and reveal hidden information that is private for an individual or organization. The common approaches used for Association are a priori, FP-Growth etc.

(c) Clustering: Preserving the privacy of individuals when data are shared for clustering is a complex problem. An important task is to protect the primary data values without interpretation of the similarity between objects under analysis in clustering functionality. The common algorithms used for Clustering are Partitioning methods, Hierarchical methods, Grid methods etc.

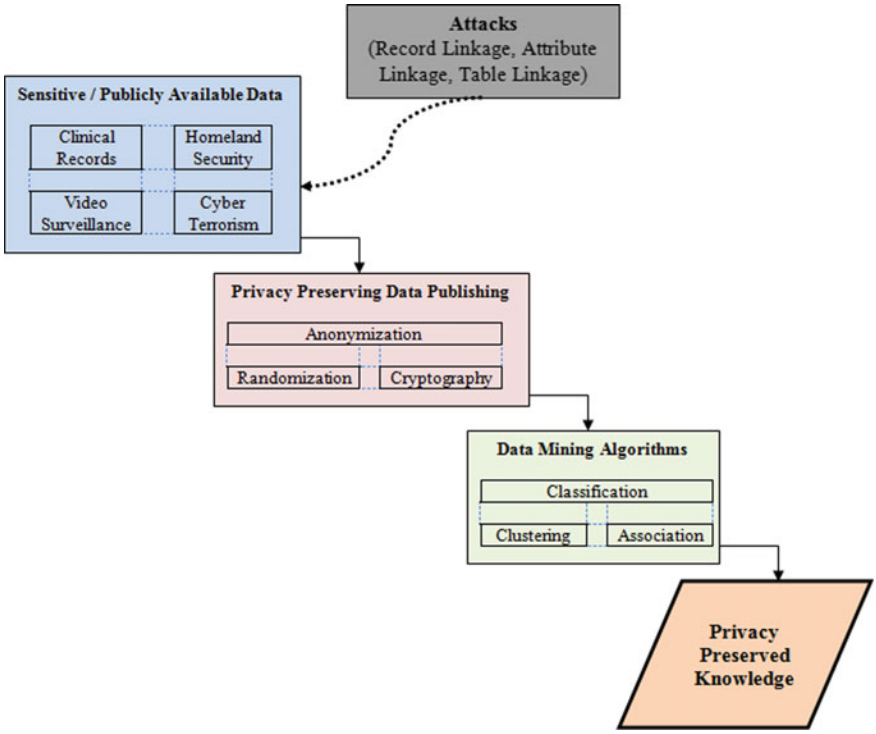

Fig. 1 Generic view of privacy preserving data mining

As a summary of the above details, Fig. 1 shows the generic view of privacy preserving data mining.

Table 1 shows the comparative and parametric analysis of various privacy preserving data mining techniques considering advantages and disadvantage of each privacy model, possible attacks on the data and different data mining tasks.

In Table 1, the correct mark in attack models represent that the technique prevents data from the particular attack and the correct mark in data mining task represents that the functionality can be applied to perform the said data mining method.

Table 1 Parametric analysis of privacy preserving data mining techniques

Privacy Models		Advantages	Disadvantages	Attack models[a]			Data mining tasks[a]		
				RL	**AL**	**TL**	**1**	**2**	**3**
Anonymization	Generalization and suppression	• Simple Technique • Protects Identity Disclosure • More flexible	• Suffers from homogeneity attack and background knowledge attack • Significant loss of granularity • Not applicable for continuous data. • Suppression complicates analysis	√			√	√	√
	Anatomization and Permutation	• More Accurate than Generalization • Certain aggregate computations can exactly be performed without violating the privacy of the data	• Linking Attack • Cannot be applied to High dimensional data without complete loss of utility • Lacks a formal framework for providing how much privacy is guaranteed	√			√	√	√
	Perturbation	• Attributes are preserved independently • Direct protection for privacy of data is possible due to statistical nature of data mining	• Does not reconstruct the original data values but only data distributions. • Loss of information • Need to develop distribution based algorithm every time	√			√	√	√
	Bucketization	• Used for high dimensional data • Better data utility than Generalization	• Does not prevent membership disclosure • Requires a clear separation between Quasi-identifiers and sensitive attributes	√	√		√	√	√
	Micro-aggregation	• Unified approach unlike Suppression and Generalization • Reduces the impact on outliers • Reduces data distortion	• Finding an optimal partition in multidimensional micro-aggregation is NP-Hard problem	√			√	√	√

(continued)

Table 1 (continued)

Privacy Models	Advantages	Disadvantages	Attack models[a]			Data mining tasks[a]		
			RL	**AL**	**TL**	**1**	**2**	**3**
Randomization	• Simple and easily implemented at data collection phase • Efficient as compared to Cryptography • Doesn't require knowledge of distributions of other records of data • Doesn't require trusted server	• High Information Loss • Cannot be used for multiple attribute databases • Treats all the record equally and reduces the utility of the data	√			√		
Cryptography	• Offers a well-defined model for privacy • Better privacy as compared to Randomization • Vast toolset of cryptographic algorithms for implementing PPDM	• Complexity increases when more parties are involved • Does not address the question of whether the disclosure of the final data mining result may break the privacy of distinct records • Long Process	√	√	√	√	√	√

[a] *RL* Record Linkage, *AL* Attribute Linkage, *TL* Table Linkage
[b] *1* Classification, *2* Clustering, *3* Association

5 Conclusion and Future Work

With proliferation in data mining techniques, the privacy of the individuals or organizations is being disclosed which has fuelled the field of privacy preserving data mining. In this paper, a broad study on privacy preserving data mining has been conducted with respect to various parameters. The field of PPDM mainly considers information privacy and the paper highlights various scenarios and applications that emphasize the importance of PPDM. Further, a tabular analysis depicting the relation between privacy models, attack models and data mining tasks is presented. Also, the merits and demerits of some of the popular PPDM techniques have been described that would help beginners to carry out research considering different dimensions of PPDM. From the study, it has been concluded that a single algorithm, in its naive form, is not efficient enough for effectively protecting data privacy since the process of preserving privacy leads to decrease in accuracy of the final data mining result. Thus, the future work can be to attempt applying variants of privacy models and their enhancements to improve the accuracy of data mining tasks while preventing the data from various attack models.

References

1. P. Tan, M. Steinbach, V. Kumar, *Introduction to Data Mining*, 1st edn. (Addison Wesley Longman Publishing, Co. Inc., 2005)
2. J. Panackal, A. Pillai, Privacy preserving data mining: an extensive survey, in *ACEEE, Proceedings of International Conference on Multimedia Processing, Communication and Information Technology* (2013)
3. M. Dhanalakshmi, S. Sankari, Privacy preserving data mining techniques-survey, in *Proceedings of information communication & embedded systems* (2014)
4. K. Babu, Utility-Based Privacy Preserving Data Publishing. Ph.D. thesis, National Institute of Technology Rourkela (2013)
5. X. Ge, J. Zhu, *New Fundamental Technologies in Data Mining*, ch. Privacy preserving data mining (In Tech Publishing, 2011)
6. S. Gokila, P. Venkateswari, A survey on privacy preserving data publishing. Int. J. Cybern. Inf. **3**(1) (2014)
7. C. Aggarwal, P. Yu, *Advances in Database Systems*, ch. A general survey of privacy-preserving data mining models and algorithms (Springer, 2008)
8. G. Nayak, S. Devi, A survey on privacy preserving data mining: approaches and techniques. Int. J. Eng. Sci. Technol. (2011)
9. V. A-Rivera, P. McDonagh, T. Cerqueus, L. Murphy, A systematic comparison and evaluation of k-anonymization algorithms for practitioners. Trans. Data Privacy, 337–370 (2014)
10. P. Samarati, Protecting Respondents' identities in Microdata Release. IEEE Trans. Knowl. Data Eng. (2001)
11. L. Sweeney, Achieving K-anonymity privacy protection using generalization and suppression. Int. J. Uncertain. Fuzziness Knowl.-Based Syst. **10**(5), 571–588 (2002)
12. R. Brand, Microdata protection through noise addition, in *Inference Control in Statistical Databases*. Lecture Notes in Computer Science, vol. 2316, pp. 97–116 (2002)

13. X. Xiao, Y. Tao Anatomy: simple and effective privacy preservation, in *Proceedings of the 32nd International Conference on Very Large Data Bases*, pp. 139–150 (2006)
14. J. D-Ferrer, V. Torra Ordinal, continuous and heterogeneous k-anonymity through Microaggregation, in *Data Mining Knowledge Discovery*, vol. 11, no. 2 (2005)
15. P. Samarati, L. Sweeney, Generalizing data to provide anonymity when disclosing information, in *Proceedings of ACM Symposium on Principles of Database Systems* (1998)
16. C. Aggarwal, P. Yu, *Advances in Database Systems*, ch. A survey of randomization methods for privacy-preserving data mining (Springer, 2008)
17. Y. Lindell, B. Pinkas, Privacy preserving data mining. J. Cryptol. **15**(3), 177–206 (2002)
18. L. Vasudevan, D. Sukanya, N. Aarthi, Privacy preserving data mining using cryptographic role based access control approach, in *Proceedings of the International Multi-Conference of Engineers and Computer Scientists*, Hong Kong, pp. 19–21 (2008)
19. B. Fung, K. Wang, R. Chen, P. Yu, Privacy-preserving data publishing: a survey of recent developments. ACM Comput. Surv. **42**(4) (2010)
20. R. Kotecha, S. Garg, Data streams and privacy: two emerging issues in data classification, in *Nirma University International Conference on Engineering* (IEEE, 2015)
21. K. Saranya, K. Premalatha, S. Rajasekar, A survey on privacy preserving data mining, in *International Conference on Electronics & Communication System* (IEEE, 2015)
22. J. Han, M. Kamber, *Data Mining: Concepts and Techniques* (Morgan Kaufmann Publishers Inc., 2005)
23. R. Kotecha, V. Ukani, S. Garg, An empirical analysis of multiclass classification techniques in data mining, in *Nirma University International Conference on Engineering* (IEEE, 2011)