IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 2, February 2016

A Review on Data Anonymization in Privacy Preserving Data Mining

Kinjal Parmar[1], Vinita Shah[2]

PG Student, Department of Information Technology, G H Patel College of Engineering and Technology, Vidhyanagar, India[1]

Assistant Professor, Department of Information Technology, G H Patel College of Engineering and Technology, Vidhyanagar, India[2]

Abstract: People today are very reluctant to share their information as they are well aware of the privacy threats of their sensitive data. Data in its original form contains sensitive information about individuals, and publishing such data without revealing sensitive information is a difficult task. The major risk is of those non-sensitive data which may deliver sensitive information indirectly. Privacy preserving data mining (PPDM) try to overcome this problem by protecting the privacy of data without sacrificing the integrity of data. A number of techniques have been proposed for privacy-preserving data mining. This paper provides a review of different approaches for privacy preserving data mining along with merits and demerits. It provides a brief explanation of anonymization approach along with its different techniques like k-anonymity, l-diversity and t-closeness. It also includes comparison between different algorithms of anonymization with their advantages and disadvantages.

Keywords: Privacy preserving, Anonymization, Randomization, Sensitive attributes, k-anonymity.

I. INTRODUCTION

Privacy Preserving Data Mining (PPDM) is a field of Data Mining which is used for the extraction of useful knowledge from large amount of data, while protecting the sensitive information simultaneously. Data Mining [1] refers to extracting or mining knowledge from large amounts of data. Privacy preserving [2] is said to be done when the attacker is not able to learn anything extra from the given data even with the presence of his background knowledge obtained from other sources.

From last decade, due to an exponential growth in the data generation and rapid increase in data storage ability, there is wide proliferation in the knowledge and information based decision making. Information about individuals is being collected on a day to day basis. According to Moor (2004), once our personal information is digitised and made available over a computer network, or via the Internet, it becomes "greased data" that can easily slip across cyberspace and personal information may no longer be controlled" by those to whom it refers and it may well be accessed by those "who have no right to do so". Many organizations publish micro data for different purposes such as business, demographic research, public health research etc. These data may contain sensitive or valuable information of any individuals, e.g., organizations such as hospitals contain medical records of the patients which they provide to the researchers or data miner for the purpose of research. Data miner analyses the medical records to gain useful global health statistics. In this process the data miner may able to obtain sensitive information and in combination with an external database may try to obtain personal attribute of an individual. So privacy becomes an important issue when data involves sensitive information.

To solve this problem, Privacy Preserving Data Mining (PPDM) has been emerged [8].

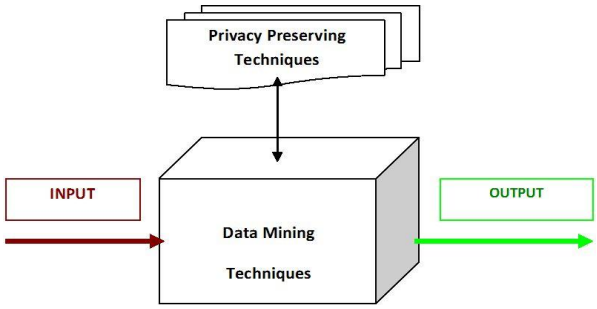

Fig 1.A Framework for Privacy Preserving Data Mining Systems: High-Level

PPDM is used for the extraction of useful knowledge from large amount of data, while protecting the sensitive information simultaneously.

As shown in the figure 1[15], PPDM modify the original dataset and release the privacy preserved dataset which protect the sensitive information of original dataset. With the help of PPDM approach the researchers can study the data without compromising privacy of any individual. Privacy preserving data mining techniques clearly depend on the definition of privacy, which captures what information is sensitive in the original data and should therefore be protected from either direct or indirect disclosure.

Paper provides a review on Anonymization approach for PPDM. It is explained as follows: Section 1 gives introduction to PPDM. Section 2 provides the classification of privacy preserving approaches. Section 3

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 2, February 2016

explains anonymization in privacy preserving data mining. Different algorithms for anonymization are discussed in section 4 along with their comparison, merits and demerits.

II. CLASSIFICATION OF PPDM APPROACHES

Privacy-preservation methods can generally be executed at different steps of the data mining process:

A. Privacy Preserving during Data collection
The only known method for privacy protection at data collection is the randomization method.

i) The Randomization Method
In this method, noise is fused in the data at data collection time. It creates private representations of the records using different data distortion methods. The randomization method is easily implemented at data collection time, because the added noise is independent of the behaviour of other data records. There are two steps to be carried out [12]: During the first step, the data providers randomize their data and send the randomized data to the data receiver. In second step, the original distribution of the data is reconstructed. The randomization response model is shown in figure 2.

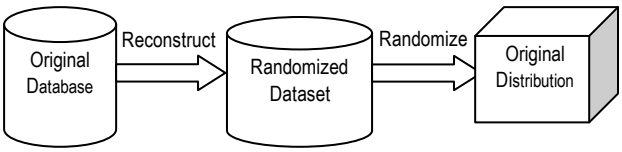

Fig. 2 Model of Randomization [6]

There are some advantages and drawbacks of this technique.

Advantages:
This method is easily implemented during the data collection phase. It is useful for hiding individual sensitive data.

Disadvantage: One of the main disadvantages of this method is that it adds noise to the original data which reduces data utility.

B. Privacy-Preserving Data Publishing
Privacy-preserving data publishing assumed that all the records are already available to a trusted party, who might be the current owner of the data. This party then wants to release (or publish) this data for analysis. Privacy-preserving data publishing is typically performed using k-anonymity, l-diversity and t-closeness which will be explained later in this paper. The eventual goal of all these methods is to prevent the release of sensitive information about individuals.

C. Output privacy of data mining algorithms:
The output of data mining algorithms may contain much sensitive information and can be used by an adversary to reveal the private data. Therefore, in many cases, the output needs to be restricted to prevent the release of sensitive information.

i) Association Rule Hiding
It is also known as frequent pattern hiding approach. In this technique the modification is applied on the output of the data mining algorithm, rather than the base data.

The main purpose of this technique is to hide the rules themselves, instead of changing the entries [18].A set of sensitive rules are specified by the system administrator. The task is to mine all association rules, such that none of the sensitive rules are discovered, but all non-sensitive rules are discovered. Association rule hiding methods are either heuristic methods, border-based methods, or exact methods [19].

ii) Downgrading Classifier Effectiveness
In this approach the data is modified in such a way that the accuracy of the classification process is reduced , while retaining the utility of the data for other kinds of applications.

iii) Query Auditing and Inference Control
Many sensitive databases are not available for public access, but may have a public interface through which aggregate querying is allowed. A smart adversary may pose a sequence of queries through which he may deduce sensitive facts about the data. In query auditing, to prevent the disclosure, one or more queries are denied from a sequence of queries or the responses to some of the queries are audited[21].

D. Distributed Privacy Preservation
Along with the centralized data scenario, the distributed data also exist and preserving the privacy in this scenario is very difficult as data is distributed at different places. Distributed privacy preservation can is classified into horizontal data distribution and vertical data distribution. It uses cryptographic approach to preserve privacy. The approach is based on the special encryption protocol named as secure multiparty computation (SMC) technology. The aim of secure multiparty computation is to enable parties to carry out distributed computing tasks in a secure manner.

III. ANONYMIZATION IN PRIVACY PRESERVING DATA MINING

Anonymization method aims at making the individual record be indistinguishable among a group records by utilizing techniques of generalization and suppression [3]. Different attributes in a data set may play different roles in either facilitating identification or facilitating sensitive information release. There are three main types of attributes [14]:

i) Key Attribute/Explicit identifiers: The attribute that can identify an individual directly is known as the key attribute. It is always removed during the release of data. e.g. Name, Social Security Number (SSN).
ii) Quasi-Identifier or Pseudo-identifier: The attributes that do not provide a unique identification, but which in combination might yield a unique identification by means of linking attacks are known as Quasi Identifiers. e.g Date of Birth, ZIP Code.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 2, February 2016

iii) Sensitive Attribute: The attribute containing the sensitive information about an individual is the sensitive attribute. E.g. Salary, Health Problem.

Anonymization [9] is a PPDM approach that hides the identity and sensitive data of record owners, assuming that sensitive data must be retained for data analysis. Explicit identifiers are removed first. Even with all explicit identifiers being removed, [8] showed a real-life privacy threat in which an individual was identified uniquely using his name in a public voter list linked with his record in a published medical database through the combination of zip code, date of birth, and sex, as shown in Figure 2.

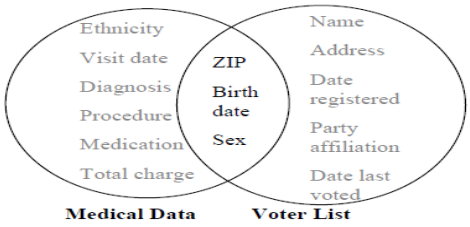

Fig.2 Linking to re-identify record owner [13]

Each of these attributes does not uniquely identify a record owner, but their combination, called the quasi-identifier identifies a unique or a small number of record owners. In the above example, the owner of a record is re-identified by linking his quasi-identifier. To perform such linking attacks, the attacker needs two pieces of prior knowledge: the victim's record in the released data and the quasi-identifier of the victim. The anonymization problem is to produce an anonymous table T` that satisfies a given privacy requirement with no or less information loss.

A. k-anonymity

k-anonymity states that – "There should be at least k tuples having the same quasi-identifier values to guarantee an individual's privacy. Every tuple in a table should be similar to at least (k-1) tuples then only the table will achieve k-anonymity".[8] K-anonymity is achieved by using generalization and suppression.

Generalization: Transformation of any value to a more general form is the process of generalization. E.g. "Male" and "Female" can be generalized to "Person". Generalization can be applied at the following levels:

i) Attribute (AG): Generalization is performed at the level of column.
ii) Cell (CG): Generalization is performed on single cells.

Suppression: Removing any value completely from a data table is the process of suppression. Suppression can be applied at the following levels:

i) Tuple (TS): Suppression is performed at the level of row; removes a whole tuple.
ii) Attribute (AS): Suppression is performed at the level of column.
iii) Cell (CS): Suppression is performed at the level of single cells.

Models of k-anonymity

The possible combinations of different types of generalizations and suppressions result in different models of k-anonymity. The following are the different models:-

i) AG_TS: Generalization is applied at the level of attribute (column) and suppression at the level of tuple (row).
ii) AG_AS: Both generalization and suppression are applied at the level of column.
iii) AG_CS: Generalization is applied at the level of column, while suppression at the level of cell.
iv) AG: Generalization is applied at the level of column, suppression is not considered.
v) CG_CS: Both generalization and suppression are applied at the cell level. Then, for a given attribute we can have values at different levels of generalization.
vi) CG: Generalization is applied at the level of cell, suppression is not considered.
vii) TS: Suppression is applied at the tuple level, generalization is not allowed.
viii) AS: Suppression is applied at the attribute level, generalization is not allowed.
ix) CS: Suppression is applied at the cell level, generalization is not allowed.

Advantage: k-anonymity prevents record linkage by generating large equivalence class.

Drawback: If most records in an equivalence class have similar values on a sensitive attribute, the attacker can still relate the sensitive value of an individual without identifying his record.

B. l-diversity

l-diversity is based on the concept of intra-group diversity of sensitive values.

l-Diversity [11]: "A data set is said to satisfy l-diversity if, for each group of records sharing a combination of key attributes, there are at least l "well represented" values for each confidential attribute."

Advantage: l-Diversity prevents from homogeneity attack and background knowledge attack.

Drawback: l-Diversity may be difficult and unnecessary to achieve and it is insufficient to prevent attribute disclosure.

C. t-closeness

t-closeness: "A data set is said to satisfy t-closeness if, for each group of records sharing a combination of key attributes, the distance between the distribution of the confidential attribute in the group and the distribution of the attribute in the whole data set is no more than a threshold t"[13].

Advantage: t-Closeness solves the attribute disclosure vulnerabilities inherent to l-diversity: skewness and similarity attack.

Drawback: t-closeness limits the amount of useful information that is released. It destroys the correlations between key attributes and confidential attributes.

DOI 10.17148/IJARCCE.2016.5216

IV. k- ANONYMITY ALGORITHMS

A. Samarati's Algorithm

This algorithm searches for the possible k-anonymous solutions by jumping at different levels in Domain Generalization Hierarchy (DGH). It uses the binary search to obtain the solution in less time. This algorithm implements the AG_TS model. Therefore, suppression can be used to achieve k-anonymity. Samarati makes the assumption that the best solutions are the ones that result in a table having minimal generalizations [15].

Advantage: Samarati's output always has a chance to be an optimal solution, and that is why it had good results when compared to Datafly algorithm.

B. Sweeney's Algorithm (Datafly Algorithm)

Sweeney considers that the best solutions are the ones that are attained after generalizing the variables with the most distinct values (unique items).This approach only goes through a very small number of nodes in the lattice to find its solution. Thus, from a time perspective, this approach is very efficient (hence the name Datafly)[9].

In other words, at each node in the lattice, it check for the which attribute having the most unique items and generalize that attribute one level up according to the corresponding hierarchy. Keep doing this until there are fewer than k rows not complying to k-anonymity, then suppress these remaining rows.

Advantage: The algorithm checks very few nodes for k-anonymity due to which it is able to give results very fast.

Disadvantage: The algorithm skips many nodes, therefore, the resulting data is very generalized and sometimes this released data may not be suitable for research purpose as it provides very little information.

C. Incognito Algorithm

This algorithm produces all the possible k-anonymous full-domain generalizations of a relation(say T), with an optional tuple suppression threshold. It begins by checking single- attribute subsets of the quasi-identifier, and then iterates, checking k-anonymity with respect to larger subsets of quasi-identifiers[30].

Advantage: The algorithm always locates the optimal solution.

Disadvantage: The algorithm uses breadth first search method which takes a lot of time to traverse the solution space.

D. Comparison of algorithms

TABLE I COMPARISON OF ANONYMITY ALGORITHMS

Samarati	Sweeney	Incognito
Evaluates all the nodes at a generalization level	Skips a lot of nodes when moving between levels	Generates the set of all possible solutions.
Provides a solution with minimal generalization and Suppression.	The solution doesn't contain minimum generalization.	Provides a solution with minimal generalization and Suppression.
More execution time.	Execute very quickly on large data sets.	More execution time.
Do not necessarily locate the optimal solution	Do not necessarily locate the optimal solution	Always locates the optimal solution
Min Information loss	Max Information loss	Min Information loss

V. CONCLUSION

Privacy Preserving Data Mining is a vast area of research and there are different approaches to classify it. This paper explains different Privacy Preserving Data Mining approaches. It discusses the anonymization approach in brief along with comparison of its different algorithms. Data anonymization is an efficient approach of PPDM which modifies the dataset to prevent the sensitive information. There are many future direction in data anonymization which includes anonymization on multiple sensitive attributes, anonymizing sequentially released data and non homogeneous data anonymization.

REFERENCES

[1] J. Han and M. Kamber, "Data Mining Concepts and Techniques", Elsevier, 2008.
[2] Stan Matwin, "Privacy Preserving Data Mining Techniques: Survey and Challenges", Discrimination and Privacy in the information Society, Springer, pp.209-222, 2013.
[3] A.S.Shanthi, , Dr. M. Karthikeyan" A Review on Privacy Preserving Data Mining "IEEE International Conference on Computational Intelligence and Computing, 2012 .
[4] Charu C. Aggarwal, Philip S. Yu ,"A General Survey of Privacy-Preserving Data Mining Models and Algorithms" ,Springer, 2008.
[5] C. C. Aggarwal, Data Mining: The Textbook, Springer International Publishing Switzerland, 2015.
[6] K.Saranya , K.Premalatha , S.S.Rajasekar ,"Survey On Privacy Preserving Data Mining", IEEE Sponsored 2nd International Conference On Electronics And Communication System (ICECS 2015)
[7] KeWang and Benjamin C. M. Fung. Anonymizing sequential releases. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '06, pages 414–423, New York, NY, USA, 2006. ACM.
[8] Benjamin C. M. Fung, Ke Wang, Rui Chen, and Philip S. Yu. Privacy-preserving data publishing: A survey of recent developments. ACM Comput. Surv. 42(4):14:1–14:53, June 2010.
[9] Sweeney L, "Achieving k-Anonymity privacy protection uses generalization and suppression" International journal of Uncertainty, Fuzziness and Knowledge based systems, 10(5), 571-588, 2002.
[10] KeWang and Benjamin C. M. Fung. Anonymizing sequential releases.12th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '06, pages 414–423, New York, NY, USA, 2006,ACM.
[11] M.E. Nergiz, C. Clifton, and A.E. Nergiz. Multirelational k-anonymity. Knowledge and Data Engineering, IEEE Transactions on, 21(8):1104 –1117, aug. 2009.
[12] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. L-diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data (TKDD), 1(1):146, 2007.

DOI 10.17148/IJARCCE.2016.5216

[13] M. B. Malik, M. A. Ghazi and R. Ali, "Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects", in proceedings of Third International Conference on Computer and Communication Technology, IEEE 2012.
[14] Li N., Li T., Venkatasubramanian: t-Closeness: Privacy beyond k-anonymity and l-diversity. ICDE Conference, 2007.
[15] P. Samarati. Protecting respondents' identities in microdata release. IEEE Transactions on Knowledge and Data Engineering, 13(6):1010–1027, 2001.
[16] J. Indumathi, "A Generic Scaffold Housing the Innovative Modus Operandi for Selection of the Superlative Anonymisation Technique for Optimized Privacy Preserving Data Mining", "Data Mining Applications in Engineering and Medicine",pp.133-156,2012.
[17] Jaideep Vaidya, Chris Clifton, "Privacy-Preserving Data Mining: Why, How, and When", IEEE Security & Privacy, 2004.
[18] Dhanalakshmi.M, Siva Sankari.E, "Privacy Preserving Data Mining Techniques-Survey", IEEE, Information Communication and Embedded Systems (ICICES), 2014.
[19] A. Evfimievski, R. Srikant, R. Agrawal and J. Gehrke, "Privacy Preserving Mining of Association Rules", SIGKDD 2002. pp. 217-228.
[20] Vijayarani, Tamilarasi, "Privacy preserving data mining based on association rule- a survey",Communication and Computational Intelligence (INCOCCI),IEEE,2010.
[21] Moskowitz I., Chang L.: A decision theoretic system for information downgrading. Joint Conference on Information Sciences, 2000.
[22] Nabar S., Marthi B., Kenthapadi K., Mishra N., Motwani R.: Towards Robustness in Query Auditing. VLDB Conference, 2006.
[23] M. Kantarcioglu. A survey of privacy-preserving methods across horizontally partitioned data. Privacy-Preserving Data Mining: Models and Algorithms, Springer, pp. 313–335, 2008.
[24] G. Nageswara Rao, M. Sweta Harini, Ch. Ravi Kishore, "A Cryptographic Privacy Preserving Approachover Classification", Springer, 2014.
[25] J. Vaidya. A survey of privacy-preserving methods across vertically partitioned data. Privacy-Preserving Data Mining: Models and Algorithms, Springer, pp. 337–358, 2008.
[26] Pingshui Wang1, and Jiandong Wang , "L-Diversity Algorithm for Incremental Data Release" , IEEE ,2013.
[27] Yousra Abdul Alsahib ,S. Aldeen, Mazleena Salleh and Mohammad Abdur Razzaque, "A comprehensive review on privacy preserving data mining",Springer,2015.
[28] Grigorios Loukides, Aris Gkoulalas-Divanis, and Jianhua Shao. Assessing disclosure risk and data utility trade-off in transaction data anonymization. Int. J. Software and Informatics, 6(3):399–417, 2012.
[29] Hongwei Tian and Weining Zhang , "Privacy-Preserving Data Publishing Based On Utility Specification" ,IEEE, 2013.
[30] Kristen LeFevre David J. DeWitt Raghu Ramakrishnan " Incognito: Efficient FullDomain KAnonymity". SIGMOD '05 Proceedings of the 2005 ACM SIGMOD international conference on Management of data ,Pages 49-60 .

DOI 10.17148/IJARCCE.2016.5216