Survey of Anonymity Techniques for Privacy Preserving

Neha Prajapati[1]
[1]Student of M.E
[1]Department of Computer Science and Engineering
[1]Narnarayan Shastri Institute of Technology, Jetalpur, Ahmedabad, Gujarat, India

Abstract— The advancement of information technologies has enabled various organizations (e.g., census agencies, hospitals) to collect large volumes of sensitive personal data (e.g., census data, medical records). Data in its original form, however, typically contains sensitive information about individuals, and publishing such data will violate individual privacy. Protecting data privacy is an important problem in micro data distribution. Recently, there are various methods and techniques which have been created for providing privacy to the process of data mining. Anonymity techniques typically aim to protect individual privacy, with minimal impact on the quality of the released data. In this paper we provide an overview of anonymity techniques for privacy preserving. We discuss the anonymity models, the major implementation ways and the strategies of anonymity algorithms, and analyze their advantage and disadvantage. Then we give a simple review of the work accomplished. Finally, we conclude further research directions of anonymity techniques by analyzing the existing work.

Key words: Privacy Preserving Data Mining, Anonymity techniques, Randomization, K-Anonymity, anonymity models

I. INTRODUCTION

Data mining is the process of extracting interesting patterns or knowledge from huge amount of data. In recent years, there has been a tremendous growth in the amount of personal data that can be collected and analyzed by the organizations [1]. As hardware costs go down, organizations find it easier than ever to keep any piece of information acquired from the ongoing activities of their clients. These organizations constantly seek to make better use of the data they possess, and utilize data mining tools to extract useful knowledge and patterns from the data. Also, The current trend in business collaboration shares the data and mine results to gain mutual benefit. The knowledge discovered by various data mining techniques on these data may contain private information about individual or business such as, Social security numbers , credit card numbers, income, credit ratings, type of disease, customer purchases, etc., that must be properly protected. can be re-. So one of the great challenges of data mining is finding hidden patterns without revealing sensitive information. Privacy preservation data mining (PPDM) is answer to such challenges [2]. Privacy preservation data mining (PPDM) considers problem of maintaining privacy of data and knowledge in data mining [2]. With the development of data analysis and processing technique, the privacy disclosure problem about individual or enterprise is inevitably exposed when releasing or sharing data, then give the birth to the research issue on privacy preserving. To protect privacy against re-identifying individuals by joining multiple public data sources, after a technique of privacy preserving called k-anonymity [4] was proposed by Samarati and Sweeney in 1998, the anonymity techniques became one of the most important research issue on privacy preserving. Anonymity techniques typically aim to protect individual privacy, with minimal impact on the quality of the resulting data. We provide here an overview of anonymity techniques for privacy preserving.

II. THEORY OF BACKGROUND

A. Existing Privacy Preserving Techniques:

The main objective of privacy preserving data mining is to develop data mining methods without increasing the risk of mishandling [5] of the data used to generate those methods. Most of the techniques use some form of alteration on the original data in order to attain the privacy preservation. The altered dataset is obtainable for mining and must meet privacy requirements without losing the [5] benefit of mining.

1) Randomization:

Randomization technique is an inexpensive and efficient approach for privacy preserving data mining (PPDM). In order to assure the performance [6] of data mining and to preserve individual privacy, this randomization schemes need to be implemented. The randomization approach protects the customers' data by letting them arbitrarily alter their records before sharing them, taking away some true information and introducing some noise. Some methods in randomization are numerical randomization and item set randomization Noise can be introduced either by adding or multiplying random values to numerical records (Agrawal & Srikant, 2000) or by deleting real items and adding "fake" values to the set of attributes.

2) Anonymization:

To protect individuals' identity when releasing sensitive information, data holders often encrypt or remove explicit identifiers, such as names and unique security numbers. However, unencrypted data provides no guarantee for anonymity. In order to preserve privacy, k-anonymity model has been proposed by Sweeney [5] which achieves k-anonymity using generalization and suppression [5], In K-anonymity, it is difficult for an imposter to decide the identity of the individuals in collection of data set containing personal information. Each release of data contains every combination of values of quasi-identifiers and that is indistinctly matched to at least k-1 respondents. Generalization

3) Encryption:

Encryption method mainly resolves the problems that people jointly conduct mining tasks based on the private inputs they provide. These mining tasks could occur between mutual un-trusted parties, or even between competitors, therefore, protecting privacy becomes a primary concern in distributed data mining setting. There are two different distributed privacy preserving data mining approaches such as the method on horizontally partitioned data and that on vertically partitioned data. The encryption method can

ensure that the transformed data is exact and secure, but it is much low efficient.

B. Anonymity Models:

K-anonymization techniques have been the focus of intense research in the last few years. In order to ensure anonymization of data while at the same time minimizing the information loss resulting from data modifications, several extending models are proposed, which are discussed as follows.

1) Basic Definitions [6]:

- Identifiers: These are attributes that unambiguously identify the respondent. Examples are the passport number, social security number, name, surname, etc.
- Quasi-identifiers or key attributes: These are attributes which identify the respondent with some degree of ambiguity. (Nonetheless, a combination of quasi-identifiers may provide unambiguous identification.) Examples are address, gender, age, telephone number, etc.
- Confidential outcome attributes: These are attributes which contain sensitive information on the respondent. Examples are salary, religion, political affiliation, health condition, etc.
- Non-confidential outcome attributes: Those attribute which do not fall in any of the categories above.

2) k-Anonymity [7]:

The main objective of the K-Anonymity model is to transform a table so that no one can make high-probability associations between records in the table and the corresponding entities. In order to achieve this goal, the K-Anonymity model requires that any record in a table be indistinguishable from at least ($k-1$) other records with respect to the pre-determined quasi-identifier. In the *k*-anonymous tables, a data set is *k*-anonymous ($k \geq 1$) if each record in the data set is indistinguishable from at least (*k*-1) other records within the same data set. The larger the value of *k*, the better the privacy is protected. K-anonymity can ensure that individuals cannot be uniquely identified by linking attacks.

Let *T* (i.e. TABLE I) is a relation storing private information about a set of individuals. The attributes in *T* are classified in four categories: an identifier (AI), a sensitive attribute (SA), d quasi identifier attributes (QI) and other unimportant attributes.

For example, we have a raw medical data set as in TABLE I. Attributes sex, age and post code form the quasi-identifier. Two unique patient records 1 and 2 may be re-identified easily since their combinations of sex, age and postcode are unique. The table is generalized as a 2-anonymous table as in TABLE II. This table makes the two patients less likely to be re-identified.

AI	QI			SA
Name	*Sex*	*Age*	*Postcode*	*Illness*
Bill	M	20	13000	Flu
Ken	M	24	13500	HIV
Linda	F	26	16500	Fever
Mary	F	28	16400	HIV

Table 1: Raw Medical Data Set

AI	QI			SA
Name	*Sex*	*Age*	*Postcode*	*Illness*
Bill	M	[20,24]	13*00	Flu
Ken	M	[20,24]	13*00	HIV
Linda	F	[26,28]	16*00	Fever
Mary	F	[26,28]	16*00	HIV

Table 2: A 2-Anonymos Data Set of Table 1

C. Generalization:

Generalization consists of substituting attribute values with semantically consistent but less precise values. For example, the month of birth can be replaced by the year of birth which occurs in more records so that the identification of a specific individual is more difficult. Generalization maintains the correctness of the data at the record level but results in less specific information that may affect the accuracy of machine learning algorithms applied on the k-anonymous data set. Different systems use various methods for selecting the attributes and records for generalization as well as the generalization technique [7]. Generalization can be applied at the following levels [8]

1) Attribute (AG):

Generalization is performed at the level of column; a generalization step generalizes all the values in the column.

2) Cell (CG):

Generalization is performed on single cells; as a result a generalized table may contain, for a specific column, values at different generalization levels. For instance, in the date of birth column

Another method applied in conjunction with generalization to obtain k-Anonymity is tuple suppression. The intuition behind the introduction of suppression is about the additional method which reduces the amount of generalization to satisfy the k-anonymity constraint. Suppression is also used to moderate the generalization process when there is a limited number of outlier.

D. Suppression:

Suppression refers to removing a certain attribute value and replacing occurrences of the value with a special value "?" indicating that any value can be placed instead. Suppression can drastically reduce the quality of the data if not properly used [7].Suppression can be applied at the following levels [8]

1) Tuple (TS):

Suppression is performed at the level of row; a suppression operation removes a whole tuple.

2) Attribute (AS):

Suppression is performed at the level of column; a suppression operation obscures all the values of the column.

3) Cell (CS):

Suppression is performed at the level of single cells; as a result a k-anonymized table may wipe out only certain cells of a given tuple/attribute.

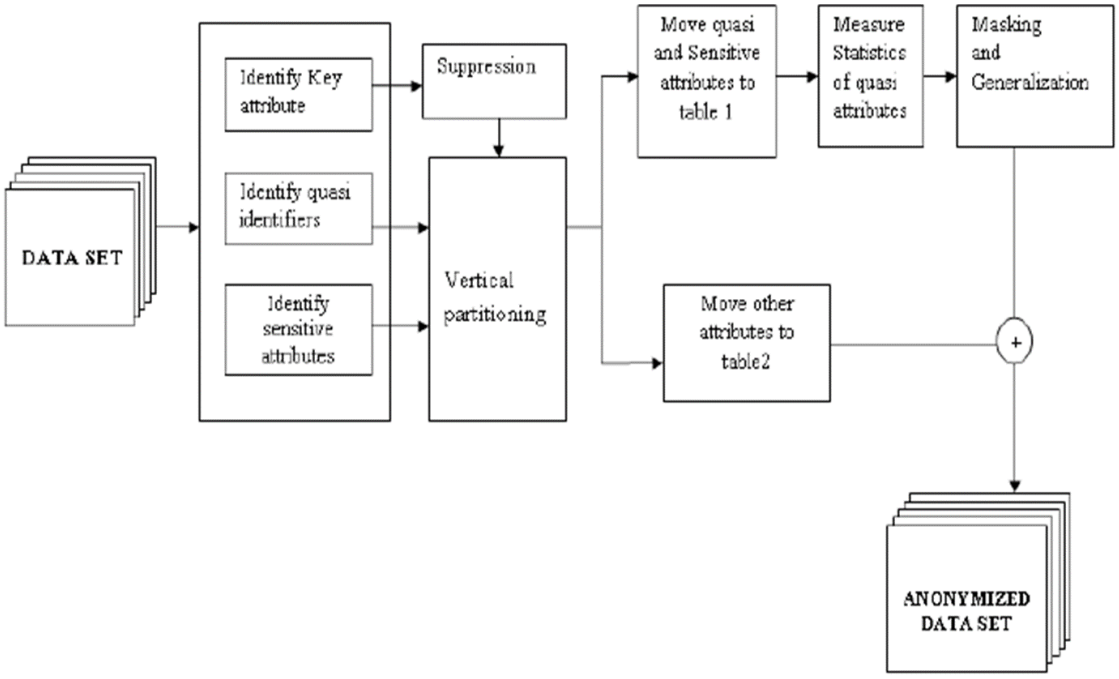

Fig. 1: Process of Anonymization of dataset [9]

4) l-diversity:

While k-anonymity is effective in preventing identification of a record, it may not always be effective in preventing inference of the sensitive values of the attributes of that record. Therefore, the technique of l-diversity was proposed which not only maintains the minimum group size of k, but also focuses on maintaining the diversity of the sensitive attributes.

Therefore, the l-diversity model for privacy is defined as follows [3, 10]:

"Let a q∗-block be a set of tuples such that its non-sensitive values generalize to q∗. A q∗-block is l-diverse if it contains l "well represented" values for the sensitive attribute S. A table is l-diverse, if every q∗- block in it is l-diverse."

A number of different instantiations for the l-diversity definition is available. When there are multiple sensitive attributes, then the l-diversity problem becomes especially challenging because of the curse of dimensionality, methods have been proposed in for constructing l-diverse tables from the data set, though the technique remains susceptible to the curse of dimensionality. Other methods for creating l-diverse tables are discussed in, in which a simple and efficient method for constructing the l-diverse representation is proposed.

5) t-closeness[4,10]:

The t-closeness model is a further enhancement on the concept of l-diversity. One characteristic of the l-diversity model is that it treats all values of a given attribute in a similar way irrespective of its distribution in the data. This is rarely the case for real data sets, since the attribute values may be much skewed. This may make it more difficult to create feasible l-diverse representations. Often, an adversary may use background knowledge of the global distribution in order to make inferences about sensitive values in the data. Furthermore, not all values of an attribute are equally sensitive. For example, an attribute corresponding to a disease may be more sensitive when the value is positive, rather than when it is negative. A t-closeness model was proposed which uses the property that the distance between the distributions of the sensitive attribute within an anonymized [14] group should not be different from the global distribution by more than a threshold t.

The Earth Mover distance metric is used in order to quantify the distance between the two distributions. Furthermore, the t-closeness[15] approach tends to be more effective than many other privacy-preserving data mining methods for the case of numeric attributes.

III. Conclusion

In above three Privacy preservation methods Anonymization is better method because in Randomization method add some noise in original data so its not preserve privacy for big dataset. Encryption especially difficult to scale when more than a few parties are involved and also it does not hold good for large databases.so anonymization preserve privacy of individual data through k-anonymity.in k-anonymity mainly preserve individual privacy and 2-attacks are possible in k-anonymity. That solution is l-diversity but in l-diversity information loss is more, and in t-closeness computing time is more. So from three methods k-anonymity, l-diversity, and t-closeness we use k-anonymity for preserve privacy in data mining.

References

[1] Han, Jiawei, Micheline Kamber, and Jian Pei. Data mining: concepts and techniques. Morgan kaufmann, 2006.

[2] Oliveira, Stanley R.M. Privacy-Preserving Data Mining. Encyclopedia of Data Warehousing and Mining, Second Edition. IGI Global, pp 1582-1588(2009).

[3] V. Kavitha and M. Poornima, "Disclosure Prevention in Privacy- Preserving Data Publishing", International Journal of Modern

Engineering Research (IJMER) Vol. 3, Issue. 3, May.-June. 2013

[4] Sachin Janbandhu1 and Dr. S. M. Chaware, "Survey on Data Mining with Privacy Preservation," (IJCSIT) International Journal of Computer Science an Information Technologies, Vol. 5 (4), 2014, 5279-5283.

[5] R. Mahesh and Dr. T. Meyyappan, "A New Method for Preserving Privacy in Data Publishing Against Attribute and Identity Disclosure Risk", International Journal on Cryptography and Information Security (IJCIS), Vol.3, No. 2, June 2013

[6] Ashwin Machanavajjhala Johannes Gehrke Daniel Kifer, "ℓ-Diversity: Privacy Beyond k-Anonymity", Muthuramakrishnan Venkitasubramaniam Department of Computer Science, Cornell University

[7] V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati, "k-Anonymity", Universitμa degli Studi di Milano, 26013 Crema, Italia, Springer US, Advances in Information Security (2007)

[8] latana sweeney, "k-ANONYMITY: A MODEL FOR PROTECTIN PRIVACY1", School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA,MAY 2012

[9] Arvind Batham,Mr. srikant lade,Mr. Deepak patel, "A Robust data preserving techniques by K-anonymity and hiding Association rules" ,International journal of Advanced Research in computer Science and software engineering volume4,issue1,January 2014.

[10] Nagendra kumar.S , and Aparna.R, "Sensitive Attributes based Privacy Preserving in Data Mining using k-anonymity", International Journal of Computer Applications (0975 – 8887) Volume 84 – No 13, December 2013

[11] Shweta Taneja, "A Review on Privacy Preserving Data Mining: Techniques and Research Challenges", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 2310-2315

[12] Manish Sharma, Atul chaudhari, Manish mathuria and shalini chaudhri, "A Review study on Privacy Preserving Data Mining: Techniques and Approches," international Journal of Computer Science and Telecommunications, [Vol. 4,issue 9,September 2013]

[13] M. Sampoorna*and V. Dineshkumar, "Evaluating Clustering Performance of K-Anonymity Methods and Techniques in PPDM" Volume 3, Issue 10, October 2013 International Journal of Advanced Research in Computer Science and Software Engineering

[14] Nagercha Kumar S and Aparna R, "Sensitive Attributes based Privacy Preserving in Data mining using k-anonymity," International Journal of Computer Applications (0975 - 8887) volume 84 – No 13, December 2013.

[15] Tiancheng Li,Ninghui Li, Senior Member ,IEEE , Jian Zhang IEEE, and lan Molloy "A new Approach for privacy preserving data publishing", IEEE Transactions on Knowledge And Data Engineering , Vol. 24. March 2012.

[16] Neha Jamdar and Vanita Babane "Survey on Privacy-Preservation in Data Mining Using Slicing Strategional", International journal of Science and research, Volume2 Issue 11, November 2013.

