International Journal of Engineering Innovation & Research
Volume 11, Issue 4, ISSN: 2277 – 5668

Hiding the Identity of an Individual During Microdata Publishing

C. Jency[1] and I. Jasmine Selvakumari Jeya[2*]
[1]Assistant Professor, Department of Information Technology, Hindusthan College of Engineering and Technology, Coimbatore-641032, Tamilnadu, India.
[2]Professor, Department of Information Technology, Hindusthan College of Engineering and Technology, Coimbatore-641032, Tamilnadu, India.
*Corresponding author email id: wjasminejeya@gmail.com

Date of publication (dd/mm/yyyy): 26/08/2022

Abstract **– Privacy is the most important issue in data publishing. Many of the organizations will distribute Individuals personal data for research and survey purpose. Confidence is considered as highest priority to be revealed during publishing the data which means that an adversary cannot predict sensitive information. In addition to that another important problem to be considered is that the adversary may also have access to external knowledge like public records and social networks related to the individuals. A unique multidimensional technique to measuring an adversary's external knowledge is also presented, along with a generic framework for privacy reasoning in the presence of external knowledge. In high dimensional space, the data become sparse, making it difficult to understand the idea of spatial locality. Data anonymization has received considerable attention due to the need of several organizations to release microdata without revealing the identity of individuals. Finally the impact of dimensionality on k-anonymity techniques is examined.**

Keywords **– Data Extraction, Data Integrity, Data Security, Data Publishing, Data Reconstruction, Privacy Preservation.**

I. INTRODUCTION

Preserving the privacy of publishing the microdata has been studied extensively in recent years. Microdata contain records each of which contains information about an Individual entity, such as a person, a household, or an organization. Several microdata anonymization techniques have been proposed. In all the approaches, attributes are partitioned into three categories, Identifiers, Quasi Identifiers (QI), Sensitive Attributes (SAs). Identifiers are used to uniquely identify an individual, such as Name or Social Security Number. Quasi Identifiers (QI), which the adversary may already know from the other publicly available databases in the society and which, when taken together, can potentially identify an individual, e.g., Birthdate, Sex, and Pincode. Sensitive Attributes (SAs), which are unknown to the adversary and are considered sensitive such as difficulties or disease or it may be Salary of an individual. Existing, data anonymization technique called slicing is introduced to improve the current state of the art. Slicing partitions the data set both vertically and horizontally. Vertical partitioning is done by grouping attributes into columns based on the correlations among the attributes. Each column contains a subset of attributes that are highly correlated. Horizontal partitioning is done by grouping tuples into buckets. Finally, within each bucket, values in each column are randomly permutated (or sorted) to break the linking between different columns. Anonymization techniques such as k-anonymization [2] usually use privacy metrics that measure the amount of privacy protection and transform the original data in such a way that it will be hard to infer identities of the individuals when it is released. In addition to that the released data will infer the knowledge needed to know. It has been taken into consideration that a watermarking of medical images greatly helps to provide authentication for safe storage and transmission of image databases. It presents a review on image watermarking algorithms for indexing medical images [8].

II. Related Works

A technique called slicing is introduced for privacy-preserving data publishing. First, the new technique called Slicing for privacy preserving data publishing. It has several advantages when compared with the previous methods called generalization and bucketization. But this Slicing technique preserves better data utility than generalization [14]. It preserves more attribute correlations with the SAs than bucketization. It can also handle high-dimensional data and data without a clear separation of QIs and SAs [10]. Secondly an optimal SVM for lung image classification where the parameters of SVM are feature selection and are managed by a modified grey wolf optimization algorithm combined with genetic algorithm [3] (GWO-GA). A set of experiments has been performed for investigating the results in terms of feature selection results and classifier performance. But slicing can be effectively used for preventing attribute disclosure, based on the privacy requirement. A technique called l-diverse slicing, which ensures that the adversary cannot learn the sensitive value of any individual with a probability greater than one. Third, an efficient algorithm for computing the sliced table is developed that satisfies l-diversity. The algorithm partitions attributes into columns, applies column generalization, and partitions tuples into buckets. Attributes that are highly correlated are in the same column; this preserves the correlations between such attributes [1]. The associations between uncorrelated attributes are broken this provides better privacy as the associations between such attributes are less-frequent and potentially identifying. Fourth, The intuition behind membership disclosure to explain how slicing prevents membership disclosure. A secured indexing of lung CT image (SILI), a secured way to index the lung CT images with the patient's detail. Authentication is achieved by the use of sender’s logo information and the secret key is utilized to embed the watermark into the host image. The simulation values indicated the presented method is robust to unauthorized access, noise, blurring, and intensity based attacks [6].

III. Architecture Diagram

This paper aims to prevent privacy information of an individual to the public and also it prevents that information using a technique called masking into some other form of related information. Meanwhile providing the necessary information needed to be published. The data to be released must follow the procedure for masking the privacy related information before publishing. The data being published will be collected from the organization and it is extracted and selected from the dataset. Then the masking process takes place. The data which reveals the identity of an individual will be grouped into a broader category based on the generalization process. The data will be identified based on their group called as bucketization. Then Anonymity process has been carried out. This ensures the clear concept of hiding the identity of an individual through the given data or information shown in Figure 1.

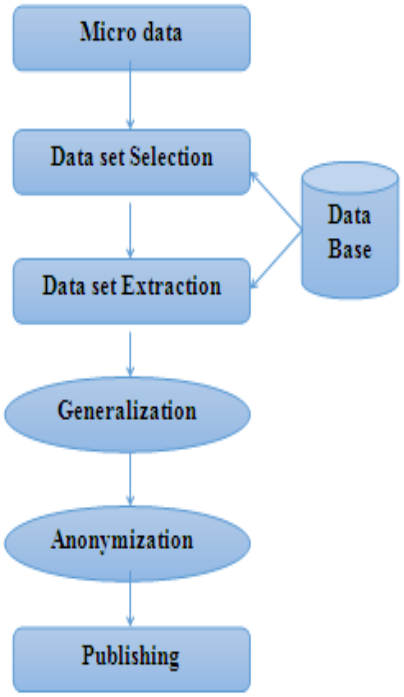

Fig. 1. Data Flow of Publishing Information.

IV. PROPOSED WORK

Slicing technique partitions the given dataset in vertically and horizontally. Vertical partitioning is done by grouping attributes into columns based on the correlations among the attributes. Each column contains a subset of attributes that are related to each other. Horizontal partitioning is done by grouping tuples into groups called buckets. Finally, within each bucket, values in each column are randomly permutated (or sorted) to break the linking between different columns. The basic idea of slicing is to break the association cross columns, but to preserve the association within each column. This reduces the dimensionality of the data and preserves better utility than generalization and bucketization. Slicing preserves utility because it groups highly-correlated attributes together, and preserves the correlations between those attributes. Slicing also protects the privacy of an individual because it breaks the associations between uncorrelated attributes. The original table is shown in Table A. The three QI attributes are {Age, Sex, Pincode}, and the sensitive attribute SA is Disease.

Table 1. Original Table.

Age	Sex	Pincode	Disease
21	M	4790687	Fever
22	F	4790687	Dyspepsia
34	F	4790567	Flu
50	F	4790567	Bronchitis
54	M	4730234	Flu
60	M	4730234	Dyspepsia
60	M	4730456	Fever
64	F	4730456	Gastritis

The Sliced table for the original table depicted in Table 1 is shown in Table 2. It contains two buckets, each containing four tuples.

Table 2. Sliced Table.

(Age, Sex)	(Pincode, Disease)
(21, M)	(4790567, fev)
(22, F)	(4790687, dysp.)
(34, F)	(4790567, flu)
(50, F)	(4790687, bron)
(54, M)	(4730456, flu)
(60, M)	(4730234, dysp)
(60, M)	(4730234,fev)
(64, F)	(4730456, gast.)

The Table 2 contains the sliced data, within each bucket; values in each column are randomly permutated to break the linking between different columns. For example, in the first bucket of the sliced table shown in Table 2, the values are randomly permutated so that the linking between the two columns within one bucket is hidden. Thus preserves the privacy of an individual.

A. *Slicing Process*

Let T represent the published microdata table. T has d attributes, with A = A1, A2,...,Ad and their attribute domains being D[A1], D[A2],...,D[Ad]. A tuple t[A1], t[A2],..., t[Ad]) may be written as t, where t[Ai] (1 I d) is the Ai value of t. Each attribute in an attribute partition only belongs to one of multiple subsets of the entire set A. A column is a subset of an attribute. Assume there is just one sensitive property, S, for the sake of simplicity in the discussion. One can either examine each sensitive characteristic separately or simultaneously if the data contains many sensitive attributes [13]. Each tuple is a member of exactly one subset of the many subsets of T that make up a tuple partition. Each group.

B. *Generalization Process*

There are many generalisation recoding types. Local recoding is the type of recoding that retains the most information [7]. In local recoding, tuples are first divided into buckets, and for each bucket, all of one attribute's values are then replaced with a generic value. Since the same property value may be generalised differently depending on which bucket it appears in, such recoding is local. If the same tuple partition is employed, slicing maintains more data than such a local recoding method. One utilises the multiset of precise values in each bucket rather than replacing more particular attribute values with a generalist value [19]. In comparison to the generalised interval, the multiset of precise values offers additional details about the distribution of values in each characteristic.

C. *Bucketization Process*

In Bucketization, there are precisely two columns: one column includes only the SA, while the other column contains all the QIs. This is a specific example of slicing. As seen below, slicing has several advantages over

bucketization [17]. First, membership disclosure may be avoided by using slicing to divide characteristics into more than two columns. The empirical analysis on a real dataset demonstrates that membership disclosure is not prevented by bucketization. Second, slicing may be utilised without a clear division between the sensitive attribute and the QI attributes, in contrast to bucketization, which demands it. Due to the lack of a single external public database that one may use to distinguish between QIs and SAs for datasets like the census data, it is sometimes impossible to properly separate them.

D. *Issues on Privacy*

Three different categories of concerns to privacy exposure exist when posting microdata [9].

(1) Membership disclosure: Since the dataset to be published is chosen from a wide population and the selection criteria are sensitive (only diabetes patients, for example), one wants to prevent enemies from discovering if their record is included in the published dataset.

(2) Identity disclosure takes place when a person is associated with a specific entry in the leaked table. When the opponent is unsure about membership, one may wish to defend against identity revelation [18].

(3) Attribute disclosure: This occurs when new information about certain persons is made public, allowing for a more precise inference of a person's qualities than would have been feasible prior to the publication of the data [20]. Also we have concentrated on the tuning of weights and bias of Radial Basis Function Neural Network (RBFNN) classifier by the use of presented Real Coded Genetic Algorithm (RCGA). The operators present d in RCGA allows the method to determine the weights and bias value so that minimum Mean Square Error (MSE) is attained. It is tested against Lung Image Database Consortium (LIDC) database and Real time database [11]. The attained results exhibited superior results over the compared methods.

V. ALGORITHMS AND PROCEDURES

The Slicing algorithm consists of three phases: attribute partitioning, column generalization, and tuple partitioning.

A. *Partitioning Based on Attributes*

Highly connected qualities are placed in the same column by the algorithm when it divides up the attributes. This benefits both privacy and utility [16]. Grouping highly correlated features retains the relationships [4] between those qualities in terms of data usefulness. Because the correlation of uncorrelated attribute values is far less common and hence more traceable, it poses greater identifying concerns in terms of privacy than the linkage of highly correlated characteristics. The mean-square contingency coefficient [5] and Pearson correlation coefficient [5] are two commonly used measures of relationship. Mean square contingency coefficient is a chi-square measure of correlation between two categorical qualities, whereas Pearson correlation coefficient is used to quantify correlations between two continuous attributes.

B. *Generalization of Column*

Tuples are generalised in the second step to meet a minimum frequency requirement. Although column generalisation is not a necessary step, it may be beneficial in a number of ways. To begin with, column

generalisation could be necessary to prevent against identity and membership disclosure [12]. Second, bucket sizes can be decreased when column generalisation is used to achieve the same level of privacy against attribute disclosure. Smaller bucket sizes provide improved data usefulness but column generalisation may cause information loss.

C. *Tuple Partitioning*

Tuples are divided into buckets during the tuple partitioning process [15]. The programme keeps track of two data structures: a collection of sliced buckets SB and a queue of buckets Q. SB is empty at first, while Q only has one bucket that holds all tuples. The algorithm divides each bucket into two buckets and eliminates a bucket from Q on each iteration. Checking if a sliced table fulfils l-diversity is the primary function of the tuple-partition algorithm. To record the frequency of each column in each bucket B, the algorithm first does one scan of each bucket B. The method then does a single scan of each tuple to identify all tuples that match B, record their matching probability, and record the distribution of potential sensitive tuples that match B.

VI. CONCLUSIONS

Slicing gets over the drawbacks of generalisation and bucketization and protects against privacy risks while maintaining an individual's privacy. The general technique suggested is that one may examine the data features before anonymizing the data and use these qualities in data anonymization. The idea is that by understanding the data more thoroughly, one may create better data anonymization solutions. Numerous research paths are encouraged by this study. First, in this paper, Slicing gets around the drawbacks of generalisation and bucketization while still preserving an individual's privacy and thwarting threats to it. This work's overall technique suggests that one can examine the data's attributes before anonymizing it and then utilize those attributes for data anonymization. It is argued that when we are more familiar with the data, we may create better data anonymization solutions. Many directions for further investigation are encouraged by this work. When each characteristic occupies precisely one column, the operation is known as slicing. The idea of "overlapping slicing" is an extension that repeats an attribute over many columns. This update includes more attribute correlations. However, the privacy considerations must be well researched and understood in order to deliver the greatest data value. Observing is fascinating It is interesting to study tradeoff between privacy and utility. Second is to study the additional information about membership disclosure protection. The results of the trials indicate that random grouping is ineffective. Third, processing high dimensional data using slicing is a promising strategy. By breaking the link between uncorrelated qualities, which would compromise privacy, and maintaining the relationship between highly correlated attributes, which would maintain data utility, attributes are divided into columns. Finally, even though several anonymization methods have been developed, the question of how to use the anonymized data is still unanswered.

REFERENCES

[1] C.Aggarwal, “On k-Anonymity and the Curse of Dimensionality,” Proc. of the Int’l Conf. on Very Large Data Bases (VLDB), pp. 901909, 2005.
[2] B.C. Chen, K. Le Fevre, and R. Ramakrishnan, “Privacy Skyline: Privacy with multi dimensional adversarial knowledge,” Proc. of the Int’l Conf. on Very Large Data Bases (VLDB), pp. 770–781, 2007.
[3] I Jasmine Selvakumari Jeya and Dinesh Valluru, “IoT with Cloud Based Lung Cancer Diagnosis Model using Optimal Support Vector Machine”, Health care management science, Springer Publication, pp. 1-10, 2019.
[4] I. Dinur and K. Nissim, “Revealing Information while Preserving Privacy”, Proc. of the ACM Symp.on Principles of Database Systems (PODS), pages 202-210, 2003.
[5] R.C.W. Wong, A.W.C. Fu, K. Wang, and J. Pei, “Minimality attack in Privacy Preserving Data Publishing,” Proc.of the Int’l Conf. on

Very Large Data Bases (VLDB), pp. 543-554, 2007.
[6] I. Jasmine Selvakumari Jeya, and Suganthi, J., 2015. RONI based secured and authenticated indexing of lung CT images. Computational and Mathematical Methods in Medicine, 2015.
[7] C. Dwork, "Differential Privacy", Proc. of the International Colloquium on Automata, Languages and Programming (ICALP), pages 1-12, 2006.
[8] I. Jasmine Selvakumari Jeya and J Suganthi, "Using Visible and invisible watermarking algorithms for Indexing Medical Images", The International Arab Journal of Information Technology, vol. 15, pp. 748- 755, 2018.
[9] B. C.M. Fung, K. Wang, and P.S. Yu, "Top-down specialization for information and privacy preservation," Proc. Int'l Conf. Data Engineering (ICDE), pp. 205216, 2005.
[10] G. Ghinita, Y. Tao, and P. Kalnis, "On the Anonymization of Sparse High-Dimensional Data", Proc. Int'l Conf. Data Engineering (ICDE), pages 715-724, 2008.
[11] . Jasmine Selvakumari Jeya, and S.N, Deepa, 2016. Lung cancer classification employing proposed real coded genetic algorithm based radial basis function neural network classifier. Computational and mathematical methods in medicine, 2016..
[12] L. Kaufman and P. Rousueeuw, "Finding Groups in Data: an Introduction to Cluster Analysis", John Wiley & Sons, 1990.
[13] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "l-Diversity: PrivacyBeyond k-nonymity," Proc. Int'l Conf. Data Engineering (ICDE), pp. 24, 2006.
[14] T. Li and N. Li, "On the Tradeoff Between Privacy and Utility in Data Publishing," Proc. of the ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (KDD), pp. 517-526, 2009.
[15] T. Li and N. Li, "Injector: mining Background knowledge for Data Anonymization," In Proc. Int'l Conf. Data Engineering (ICDE), pp. 446- 455, 2008.
[16] D. J. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J.Y. Halpern, "Worst case background knowledge for privacy preserving Data Publishing,"Proc.Int'lConf.Data Engineering (ICDE), pp. 126135, 2007.
[17] M. E. Nergiz, M. Atzori, C. Clifton, "Hiding the Presence of Individuals from Shared Databases," Proc. of the ACM SIGMOD Int'l Conf. on Management of Data (SIGMOD), pp. 665-676, 2007.
[18] R.C.W. Wong, A.W.C. Fu, K. Wang, and J. Pei, "Minimality Attack in Privacy Preserving Data Publishing," Proc. of the Int'l Conf. on Very Large Data Bases (VLDB), pp. 543-554, 2007.
[19] F. Abdul Rahim, A. A. Bakar, S. Yussof, R. Ramli, R. Ismail and B. M. Yusof, "Privacy preserving technique for smart metering Data : A Preliminary Result", *Adv. Sci. Lett*, vol. 24, pp. 1839-1842, 2018.
[20] A. Kumar, M. Gyanchandani and P. Jain, "A comparative review of privacy preservation techniques in data publishing", *2018 2nd International Conference on Inventive Systems and Control (ICISC)*, pp. 1027-1032, 2018.

AUTHOR'S PROFILE

First Author

Ms.C. Jency, working as an Assistant Professor in the Department of Information Technology, Hindusthan College of Engineering and Technology, Coimbatore. She had received her B.E., Degree in Computer Science and Engineering from Institute of Road and Transport Technology in the year 2005. M.E., Degree in Computer Science and Engineering from Sasurie College of Engineering and honoured as 29th Rank Holder of Anna University, Chennai in the year 2013. Her Area of Research Specializations are Data Mining, Medical Image Processing, Optimization Techniques. She has actively participated in 5 National/ International Conferences, published 2 patents and attended 15 FDP and Seminars. She has re--ceived Elite & Gold for Introduction to Programming in C course in NPTEL.

Second Author

Dr. I. Jasmine Selvakumari Jeya, currently working as an Professor & Head, Department of Information Technology, Hindusthan College of Engineering and Technology, Coimbatore. She has more than 19 years of teaching and research experience. She has completed bachelor degree in Computer Science and Engineering from Manonmaniam Sundaranar University in 2001, M.E degree in Computer Science and Engineering from Karunya University in 2007 and her PhD degree under faculty of Information and Communication Engineering in the year 2016 at Anna University. She has published more than 25 research papers in International Journals with good impact factor, 28 patent, 40 International and National Conferences and three books. She has received Active Participation Award for Woman Member Inspiring Innovative Faculty Award, Best Teacher Award(Senior), Special Appreciation Award for NPTEL Country Topper and Longest Continuous Student Branch Coordinator Award from the Computer Society of India (CSI). The area of interest includes Medical Image Database Security, Information Security, Image Processing, Data Mining, Cloud Computing, Big Data etc. She is a life member of Institution of Engineers, Computer Society of India and International Association of Engineers. She is a recognized supervisor in Anna University and guiding 7PhD scholars under her supervision.