INTERNATIONAL JOURNAL OF MANAGEMENT, IT & ENGINEERING

Vol. 9 Issue 8, August 2019,
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gage as well as in Cabell's Directories of Publishing Opportunities, U.S.A

A STUDY ON MICRO DATA PRESERVATION USING INTERSECT CARVING

Sugitha Arumugam*

Abstract- In terms of privacy for sensitive information anonymization techniques are used. Microdata clutches information about an individual along with sensitive information. Existing anonymization techniques are Bucketization, Generalization, Slicing and Overlap slicing. Generalization does not support high dimensional data and loses information. Bucketization does not suitable to preserve identity disclosure. The Slicing method delivers more secrecy and data efficacy than other existing methods but still, it has few limitations. To overcome the limitations a new technique called intersect carving is proposed. Intersect carving provides a high level of privacy and better data utility than slicing technique. In this method, attributes are duplicated in more than one row to obscure opponents and preserve identity disclosure.

Keywords- Privacy-Preserving, Anonymization, Sensitive Information, Data Utility.

I. Introduction

Government officialdoms uphold a medical record for each individual those who are affected by diseases. Research candidates may seek medical record for their research drive. So, an organization should not make available of the original database to researchers because the original database may encroach upon the privacy of an individual. There are many anonymization techniques available to safeguard secret information of an individual. The most common anonymization techniques are Generalization and suppression for the k-anonymity method [3] and Bucketization for the l - diversity method [4]. These methods divide attributes into three categories: 1) Explicit identifiers that used to recognize an individual exactly e.g.: Name 2) Quasi-Identifier a freely available data which others may know easily e.g.: D.O.B and 3) Sensitive attribute clandestine information about an individual e.g.: Disease, Salary.

Generalization first partition tuple values into buckets after eradicating identifiers. Then it swaps the QI-values in all buckets with "*less specific but semantically consistent values*" [2]. Bucketization method also partitions tuples into the buckets but it permutes the secret information column and does not perform any protection process across QI-values [5]. A

* is the author of this work. She has completed Master of Technology specialized in Computer Software and Engineering. She worked in Dr.Udhayamurthy school of excellence (CBSE) for 3 years and currently she is working as an ICT Lecturer in International Training Institute, Papua New Guinea. She is striving to publish more articles in her profession.

system called Slicing is presented to overcome the restrictions of the existing methods [1]. It divides the database values horizontally and vertically. Horizontal partitioning is achieved by combining the tuples into buckets. Vertical partition is done by grouping the attribute values in the column. Attributes in the column are highly interrelated [6]. The fundamental idea of slicing is to preserve the relationship within each column but split the association across columns. Slicing disruptions relationship between unrelated attributes. It provides enhanced privacy than the existing methods but it limits data utility [7]. So, '*Intersect Carving*' is introduced to overcome the drawback of slicing.

II. Scope Of The Study

As there are various contributions in the area of Privacy-Preserving Data Mining, the scope of the work is to minimize the attack of the hackers and enhance the performance of the experts using data mining preserving functionalities.

The author has puzzled out on the tabular data wherever the information sets are bestowed with a set of attributes; the sensitivity rate of the attribute is determined by taking the classification procedures of the information mining. Within the classification, the importance of the attributes is outlined as a category and additional necessary attributes or cluster of attributes as sub-classes. The ways of finding the importance of the attributes to make membership of a class pose the importance as sensitivity and therefore the gain magnitude relation are calculated because of the sensitivity rate. The sensitivity rate is directly proportional to the quantum of noise to be added to the information. According to the sensitivity the noise of data in increased to conceal the originality or to preserve the privacy of the information. Further, it's terribly tough for an attacker to predict the distribution of noise since the sensitivity rate is calculated from the average variety of queries expose to the information set.

Thus, during this work, the author tended to propose a life to quantify the individual privacy disclosure and additional propose in some way to live how shut the inter-quantile vary obtained by attackers or snoopers is to individual's privacy interval for a few specific sensitive variables. Author tend to conjointly extend such life to variable case supported the confidential region.

III. Literature Review

K. LeFevre, (K.LeFevre 2008)R. Ramakrishnan and D. DeWitt [10] propose Mondrian k-anonymity for multidimensional which supports for tuple partitioning. Mondrian algorithm also checks for diversity check which considers l- diversity. Tuple partitioning is partitioning row values into buckets. Tiancheng Li, Jian Zhang, Ian Molloy and Ninghui Li [1] defines a Modified Mondrian algorithm with few modifications of the Mondrian algorithm. The slicing method is introduced here to overcome the existing drawbacks of generalization and bucketization.

N.Koudas et al [6]., D.J.Martin et al [7]., X.Xiao and K.Wang [8] proposed bucketization method which contains only two columns. All QI- values are presented in the first column and the second column holds only the sensitive information. In the bucketization method, only the secret information is permuted and all other QI- values represent its original form. It does not have an appropriate separation between QI-values and sensitive values.

J.Xu, W.Wang, J.Pei, X.Wang, B.Shi, and A.W.-C.Fu [3] designed a method for local recoding which preserves most of the information in the anonymity method. The local recording method does not have limitations like other global and regional recoding methods. It allows discrete occurrences of the akin value to be generalized but it did not support a large data set.

P. Samarati [4] proposed the concept of protecting the respondent's privacy in data publishing. The main feature of this concept is to provide "less specific but semantically consistent value" for generalization, where the values are generalized in the same column. Values in the same column contain generalized value to obscure opponents. L.Sweeny [5] proposed the same concept in privacy protection of k-anonymity using generalization

Y.He et al [2]., M.Terrovistis et al [9]., Y. Xu et al [10]., proposed slicing method to overcome the drawbacks of generalization and bucketization. Slicing supports for high dimensional data and does not require any separation between QI-values and sensitive information. T.Li et al [1] proposed slicing as an approach to privacy-preserving for publishing data. It details about generalization, bucketization, comparison of existing methods with slicing method and results as slicing to provide better data privacy but data utility is not guaranteed.

C. Dwork [8] describes survey results about the Theory and Applications of Models of Computation (TAMC) which includes randomized function k, Achieving Differential Privacy in Statistical Data Inference, Statistical Databases, Contingency Table Release, Learning (Nearby) Halfspaces. C. Aggarwal [9] defines the Curse of Dimensionality which is the drawback of the generalization method. Curse of dimensionality does not support for large database access. When it works for the large database it will loss much data and privacy also will get affected.

IV. Research Methodology

INTERSECT CARVING

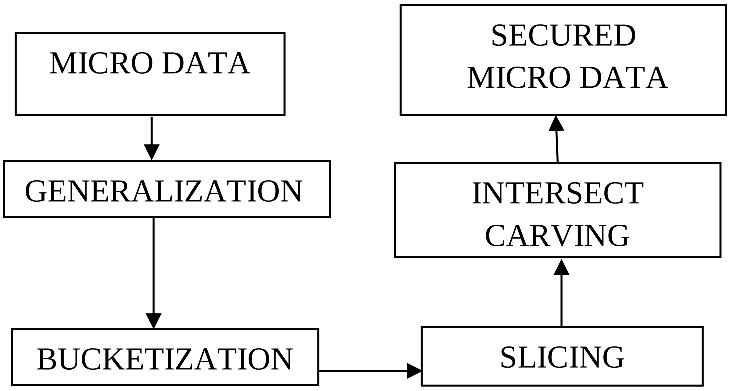

Fig.1 System Architecture

Step 1: Repossess the records from large databases.

Step 2: Anonymity method splits the records into two.

Step 3: Interchange sensitive values.

Step 4: Combine the attributes.
Step 5: Overlap the attribute combination.
Step 6: Demonstrate secured data.

Intersect Carving can be achieved in three steps. Column Generalization, Attribute Segregating, and Tuple Segregating.

A. ATTRIBUTE SEGREGATING

Attribute partitioning is allocating the record values into columns and association among the partitioned attributes should be dignified to improve the privacy of an attribute. Attribute correlation can be measured in the following ways.

1. Pearson Correlation Coefficient
2. Mean Square Contingency Coefficient.
3. Chi-Square Coefficient.

Pearson correlation coefficient is used to measure correlation among incessant value, but here only the categorical value is used so, Mean square contingency coefficient and Chi-square coefficients are used for categorical values. The chi-square coefficient method performs better than the mean square method so in intersect carving Chi-square method is used to measure the correlation among attributes.

Chi - square coefficient correlation measure for categorical value is

$$\chi 2 = \sum_{i=1}^{r}\sum_{j=1}^{c}(Aij - Bij)^2 / Bij$$

Aij= Observed value, Bij= Expected value ((row*column) /n)

B. COLUMN GENERALIZATION

Column generalization will split the attributes into the column and simplify the attribute values in the same column. This method preserves attributes from adversary without finding the original value.

C. TUPLE SEGREGATING

Tuples partitioning algorithm is used to partition the tuples into the bucket.

Algorithm-Tuple Partition (Tu,l)

Qu defines the queue value where the data set is available. Buc is a bucket, Tu is a tuple and CBu is carved bucket.

```
1. Qu = {Tu}; CBu = ∅.
2. While Qu is not empty
 3. Remove bucket Bu from Qu; Qu = Qu − {Buc}.
4. Split Buc into two buckets Buc1 and Buc2.
5. If diversity check for (Tu, Qu ∪ {Buc1, Buc2} ∪ CBu, ℓ)
6. Qu = Qu ∪ {Buc1,Buc2}.
7. Else CBu = CBu ∪ {Buc}.
8. Return CBu.
```

Initially, queue (Qu) contains tuples and carving bucket that is empty. When the queue is not null bucket (Bu) is removed from the queue. The bucket is partitioned into bucket1 (Bu1) and bucket2 (Bu2). Both the buckets are united with queue if diversity check satisfies the condition. Else bucket is combined with a carved bucket.

Intersect carving is proposed to prevent the drawbacks of the existing methods and provides better data utility than all the other existing methods. It does not entail clear separation between QI- values and delicate values. It supports large data set. Intersect slicing group column values and duplicate a column value with other column values.

TABLE.1 INTERSECT CARVING

Age, Gender Disease	Pin code, Disease
32, M, Dyspepsia 52, F, Flu	67906, Dyspepsia 47915, Flu
70, M, Gastritis 64, F, Flu	67304, Gastritis 57302, Flu

V. Data Analysis

Performance Evaluation:

In this chapter, the performance of the algorithm is evaluated by using Graph representation. The comparison result shows the duplicated data value for each method in table 2.

TABLE 2: COMPARISON TABLE

DATA	ORIGINAL	BUCKETIZATION	SLICING	INTERSECT CARVING
EQUAL DATA	24	18	12	6
NON-EQUAL DATA	0	6	12	18

Inference

The Intersect Carving method overcomes the drawback of all the existing methods. The result of the intersect carving method is shown in table 2. The following is the comparison result of all methods.

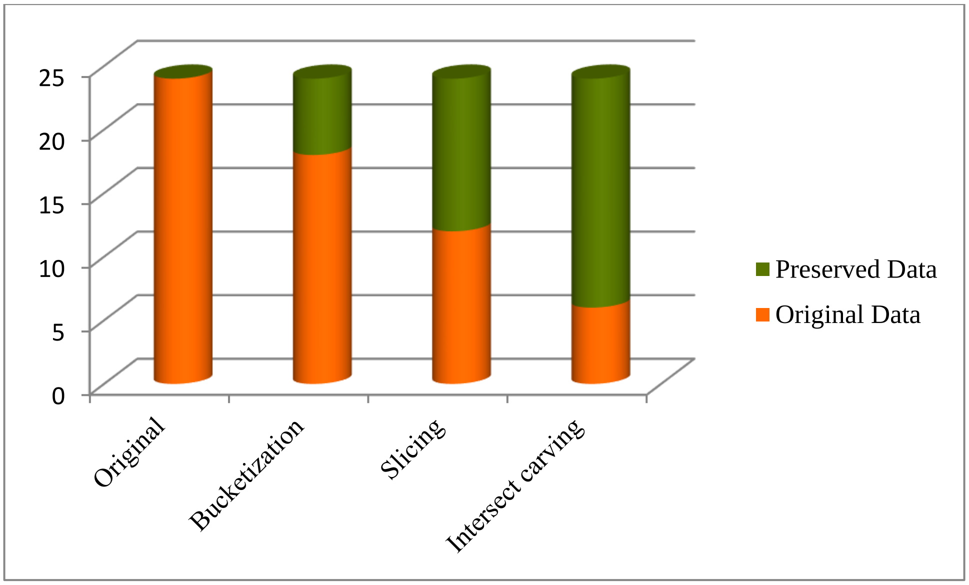

Figure 5. Graphical representation

Inference

Out of 192 data, only 48 attributes are duplicated in the bucketization method. Hence the utility and privacy-preserving percentage are 25. The slicing method duplicates 96 attributes and utility is 50 %. Intersect carving result shows that 75 % of data utility and it is better than the other existing methods.

VI. Conclusion

In this research intersect carving is introduced to improve data access and privacy of sensitive information. In the existing system, there is no guarantee for high data utility and privacy for secret information. It does not support large data set as well. Intersect carving duplicates a column value with other columns to make data secrecy and utility better. It always supports high dimensional data. The performance of the proposed method is discussed and compared with existing above and the proposed technique is an extended method of slicing where generalization and bucketization drawbacks were removed. Intersection carving produces a high rate of protection for sensitive information and obscures the intruders who want to reveal the sensitive data of an individual is the purpose of this research.

Acknowledgment

I thank the almighty to grace this publication to become a reality from my dream. I deeply adore and express my gratefulness to my husband and parents for their immortal support and encouragement throughout my work. I am extremely grateful to Mr.Sairam (late) for his eternal motivation and guidance. His advice and guidance have motivated me to complete my publication work successfully.

Furthermore, I thank the International Training Institute, Papua New Guinea, who enthused me to do the publications. Besides, I acknowledge the corespondents and the management of the hospitals that I have visited for gathering the data and other sources. I was indeed one among them all. Last but not least I express my sincere thanks to my friends for their support and cooperation.

References

[1]. C.Aggarwal. (2008). On k-Anonymity and the Curse of Dimensionality," Proc. Int'l Conf. Very Large Data Bases.

[2]. C.Dwork. (2008). Differential Privacy: A Survey of Results, Theory, and Applications of Models of Computation.

[3]. D.J Marting, D. A. (2007). Worst-Case Background Knowledge for Privacy-Preserving Data Publishing.

[4]. K. LeFevre, D. D. (2006). Mondrian multidimensional k-anonymity. In IEEE ICDE,.

[5]. L.Sweeney. (2002). Achieving k-Anonymity Privacy Protection Using Generalization and Suppression," Int'l J. Uncertainty Fuzziness and Knowledge-Based Systems, .

[6]. M.Terrovitis, N. a. (n.d.). Privacy-Preserving Anonymization of set-valued data. 2008.

[7]. P.Samarati. (2001). Protecting Respondents Privacy in Microdata Release.

[8]. Tiancheng Li, N. L. (2012). "Slicing A new approach for privacy-preserving Data Publishing.

[9]. Y.Tao, X. a. (2006). Anatomy Simple and Effective Privacy Preservation.

[10]. Y.Xu, K. A. (2008). Anonymizing Transaction Databases for Publication," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining.