Privacy Preservation Data Publishing using Slicing

Vishakha Kamble[1] Sheetal Jadhav[2] Nilesh Dolse[3]
[1,2,3]Department of Information Technology
[1,2,3]Atharva College of Engineering (INFT)

Abstract— In earlier days we have techniques like Bucketization and Generalization for Privacy preservation. Since these two techniques have some disadvantages like, for high dimensional data we can't use Generalization as it loses considerable amount of information. Also Bucketization on the other hand, does not prevent membership disclosure and does not apply for data that do not have a clear separation between quasi-identifying attributes and sensitive attributes. Hence, to overcome above disadvantages, we are introducing new technique called as Slicing, which partitions the data both horizontally and vertically. Slicing preserves better data utility than generalization and can be used for membership disclosure protection. Another important advantage of slicing is that it can handle high-dimensional data. We show how slicing can be used for attribute disclosure protection. Our experiments confirm that slicing preserves better utility than generalization and is more effective than Bucketization involving the sensitive attribute. Our experiments also demonstrate that slicing can be used to prevent membership disclosure.

Key words: Slicing, Anonymization, Bucketization

I. INTRODUCTION

Most organizations (medical, travel, flight and insurance agencies) are experiencing an exponential growth in data collection that may contain person specific, unaggregated information called microdata. Failure to provide proper protection within a release may create situations that harm the public. To avoid such privacy breach, the uniquely identifying information are removed from the table before disclose. Privacy preservation provides a limitation in linking the unveiled data to a particular individual.

There are three types of person specific data that are relevant to privacy preservation.
They include:

A. Identifiers (IDs):

Identifiers are attributes that are used individually to identify a tuple. (e.g.: Social security number, passport number). Hence IDs should always be removed to protect privacy.

B. Quasi-Identifiers (QIs):

Sets of attributes (like gender, birth date, and zip code) that can be combined with external data to act as IDs are called quasi-identifiers.

C. Sensitive Attributes (SAs):

Sensitive attributes (such as criminal offence, disease, salary) are fields that should be hidden to avoid being associated to specific persons.

Publishing information may cause threat to person's private information. A number of approaches are proposed for data anonymization in order to protect the privacy of individual. Sweeney proposed k-anonymity model in 2002. The kanonymity protection model forms the basis on which the realworld systems known as data fly, km-anonymity, (a, k) anonymization.

II. LITERATURE REVIEW

Fig. 1: Data Collection and Publishing Scenario

In the data collection phase, the data publisher collects data from record owners (e.g., Alice and Bob). In the data publishing phase, the data publisher releases the collected data to a data miner or to the public, called the data recipient, who will then conduct data mining on the published data. For example, a hospital collects data from patients and publishes the patient records to an external medical center. In this example, the hospital is the data publisher, patients are record owners, and the medical center is the data recipient. The data mining conducted at the medical center could be anything from a simple count of the number of men with diabetes to a sophisticated cluster analysis. We have to preserve the privacy of sensitive data.[9]

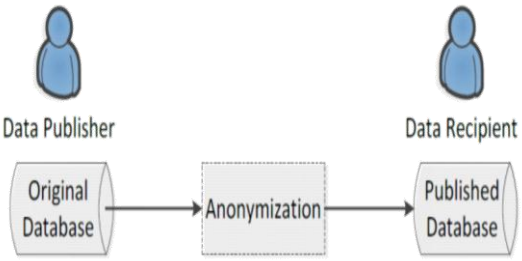

Fig. 2: Anonymization

Privacy Preserving publishing of micro data has been studied extensively in recent years. Micro data contain records each of which contains information about an individual entity, such as a person, a household, or an organization. Several micro data anonymization techniques have been proposed. The most popular ones are generalization for k-anonymity and bucketization for diversity. In both approaches, attributes are partitioned into three categories:

- Some attributes are identifiers that can uniquely identify an individual, such as Name or Social Security Number.

– Some attributes are Quasi Identifiers (QI), which the adversary may already know (possibly from other publicly available databases) and which, when taken together, can potentially identify an individual, e.g., Birthdate, Sex, and Zip code.[9]
– Some attributes are Sensitive Attributes (SAs), which are unknown to the adversary and are considered sensitive, such as Disease and Salary. In both generalization and bucketization, one first removes identifiers from the data and then partitions tuples into buckets.

The two techniques differ in the next step.

Generalization transforms the QI-values in each bucket into "less specific but semantically consistent" values so that tuples in the same bucket cannot be distinguished by their QI values.

In Bucketization, one separates the SAs from the QIs by randomly permuting the SA values in each bucket. The anonymized data consist of a set of buckets with permuted sensitive attribute values.

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 such attributes. Slicing protects privacy because it breaks the associations between uncorrelated attributes, which are infrequent and thus identifying. Note that when the dataset contains QIs and one SA, bucketization has to break their correlation; slicing, on the other hand, can group some QI attributes with the SA, preserving attribute correlations with the sensitive attribute. Slicing also has a drawback when more number of similar attribute value and the sensitive value may present in the different tuples may give the original tuple while performing the random permutation. The utility of the dataset is lost by generation the fake tuples. Thus enhanced slicing models have designed to overcome the drawbacks of slicing. The suppression slicing is done by suppressing any one of the attribute value in the tuples and then perform the slicing. Thus utility is maintained with minimum loss by suppressing only very few values and privacy is maintained by random permutation. The next model is enhanced slicing in this the random permutation is done with all the buckets not within the single bucket. Thus same utility of the original dataset is maintained. We are also going to implement overlapping slicing. This releases more attribute correlation and compare with normal slicing. Enhanced technique also helps in avoiding random grouping.[10]

III. EXISTING SYSTEM

Several micro data anonymization techniques have been proposed. The most popular ones are generalization for k-anonymity and bucketization for ‘-diversity. In both approaches, attributes are partitioned into three categories.

1) Some attributes are identifiers that can uniquely identify an individual, such as Name or Social Security Number.
2) Some attributes are Quasi Identifiers (QI), which the adversary may already know (possibly from other publicly available databases) and which, when taken together, can potentially identify an individual, e.g., Birthdate, Sex, and Zipcode.
3) Some attributes are Sensitive Attributes (SAs), which are unknown to the adversary and are considered sensitive, such as Disease and Salary.

In both generalization and bucketization, one first removes identifiers from the data and then partitions tuples into buckets. The two techniques differ in the next step. Generalization transforms the QI-values in each bucket into "less specific but semantically consistent" values so that tuples in the same bucket cannot be distinguished by their QI values. In bucketization, one separates the SAs from the QIs by randomly permuting the SA values in each bucket.

A. Disadvantages of Existing System:

1) Generalization for k-anonymity losses considerable amount of information, especially for high-dimensional data.
2) Bucketization does not prevent membership disclosure. Because bucketization publishes the QI values in their original forms, an adversary can find out whether an individual has a record in the published data or not.

IV. PROPOSED SYSTEM

We introduce a novel data anonymization technique called slicing 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.

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 such attributes. Slicing protects privacy because it breaks the associations between uncorrelated attributes, which are infrequent and thus identifying. Note that when the data set contains QIs and one SA, bucketization has to break their correlation; slicing, on the other hand, can group some QI attributes with the SA, preserving attribute correlations with the sensitive attribute. The key intuition that slicing provides privacy protection is that the slicing process ensures that for any tuple, there are generally multiple matching buckets.

A. Advantages of Proposed System:

1) It preserves better data utility than generalization. It preserves more attribute correlations with the SAs than bucketization.
2) It can also handle high-dimensional data and data without a clear separation of QIs and SAs.slicing can be effectively used for preventing attribute disclosure, based on the privacy requirement of ‘-diversity.

V. SLICING ALGORITHM

(1) Step 1: In the initial stage we consider a queue of buckets Q and a set of sliced buckets SB. Initially Q contains only one bucket which includes all tuples and SB is empty. So Q={T};SB=∅.
(2) Step 2: In each Iteration the algorithm removes a bucket from Q and splits the bucket into two buckets. Q=Q-{B}; for l-diversity check (T,Q∪{B1,B2}∪SB,l);The main part of tuple partitioning algorithm is to check whether a sliced table satisfies l- diversity.
(3) Step 3: In the diversity check algorithm for each tuple t, it maintains a list of statistics L[t] contains Statistics about one matching bucket B. t∈T,L[t]=∅.The matching probability p(t,B) and the distribution of candidate sensitive values D(t,B). Step 4: Q=Q∪{B1,B2} here two buckets are moved to the end of the Q
(4) Step 5: else SB=SB∪{B} in this step we cannot split the bucket more so the bucket is sent to SB
(5) Step 6: Thus a final result return SB, here when Q becomes empty we have Computed the sliced table. the set of sliced buckets is SB .So, finally Return SB

VI. CONCLUSION

In this project we present the detailed study about the privacy preserving data mining and briefly review the techniques Data Modification and Secure Multiparty Computation. We tried to present the comparative study of privacy preserving techniques which is helpful to understand that which technique is better in which scenario or environment. Privacy preserving data publishing is an important need for all the organization because every organization has their own personal data and they care about their data. All the methods discussed here are only approximate to our goal of privacy preservation now we need to further refine those approaches or develop some efficient techniques. For considering these issues, following problem should be widely studied.

VII. ACKNOWLEDGEMENT

It gives us great pleasure in presenting this project report titled:"Privacy Preservation Data Publishing Using Slicing"

On this momentous occasion, we wish to express our immense gratitude to the range of people who provided invaluable support in the completion of this project. Their guidance and encouragement has helped in making this project a great success.

We express our gratitude to our project guide Prof. Jyoti Arun, who provided us with all the guidance and encouragement and making the lab available to us at any time. We also would like to deeply express our sincere gratitude to Project coordinators.

We are eager and glad to express our gratitude to the Head of the Information Technology Dept. Prof. Neelima Pathak, for her approval of this project. We are also thankful to her for providing us the needed assistance, detailed suggestions and also encouragement to do the project.

We would like to deeply express our sincere gratitude to our respected principal Prof. Dr. Shrikant Kallurkar and the management of Atharva College of Engineering for providing such an ideal atmosphere to build up this project with well-equipped library with all the utmost necessary reference materials and up to date IT Laboratories

We are extremely thankful to all staff and the management of the college for providing us all the facilities and resources required.

REFERENCES

[1] C. Aggarwal, "On k-Anonymity and the Curse of Dimensionality," Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 901-909, 2005.
[2] A. Blum, C. Dwork, F. McSherry, and K. Nissim, "Practical Privacy: The SULQ Framework," Proc. ACM Symp. Principles of Database Systems (PODS), pp. 128-138, 2005.
[3] J. Brickell and V. Shmatikov, "The Cost of Privacy: Destruction of Data-Mining Utility in Anonymized Data Publishing," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 70-78, 2008.
[4] B.-C. Chen, K. LeFevre, and R. Ramakrishnan, "Privacy Skyline: Privacy with Multidimensional Adversarial Knowledge," Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 770-781, 2007.
[5] H. Cramt'er, Mathematical Methods of Statistics. Princeton Univ. Press, 1948.
[6] I. Dinur and K. Nissim, "Revealing Information while Preserving Privacy," Proc. ACM Symp. Principles of Database Systems (PODS), pp. 202-210, 2003.
[7] C. Dwork, "Differential Privacy," Proc. Int'l Colloquium Automata, Languages and Programming (ICALP), pp. 1-12, 2006.
[8] J.H. Friedman, J.L. Bentley, and R.A. Finkel, "An Algorithm for Finding Best Matches in Logarithmic Expected Time," ACM Trans. Math. Software, vol. 3, no. 3, pp. 209-226, 1977.
[9] International Journal of Emerging Research in Management &Technology ISSN: 2278-9359 (Volume-2, Issue-6)
[10] Tiancheng Li, Ninghui Li, Jian Zhang, Ian Molloy Purdue University, West Lafayette, IN 47907 {li83,ninghui}@cs.purdue.edu

