The Most Prominent Technique for Privacy Preservation in Mining Micro Data

C. Shyamala Gowri[1]

[1]Assistant Professor, Department of IT, Syed Ammal Engineering College, Ramanathapuram

Abstract

The data are increasingly being collected and used. Privacy preserving data mining tries to strike a balance between two opposing forces: the objective of discovering valuable information and knowledge, verse the responsibility of protecting individual's privacy. Several anonymization techniques, such as generalization and bucketization, have been designed for preserving privacy in micro data publishing. But generalization loses considerable amount of data. On the other side, bucketization does not prevent membership disclosure and there is no clear isolation of quasi identifiers and sensitive attributes. We present a novel data anonymization technique called slicing which partitions the data both horizontally and vertically. Our empirical result shows that slicing protects individual entity with high degree of data utility than generalization and suppression. Slicing also provides attribute and membership disclosure protection. And our algorithm satisfies ℓ-diverse requirement.

Index Terms: *Data anonymization, privacy preserving, data security, micro data, ℓ-diversity.*

1. Introduction

In recent years, the phenomenal advance technological developments in information technology have led to an increase in the capability to store and record personal data about customers and individuals. The use of published organizational micro data for variety of purposes has the chance of violation of leakage of individual secret information. Micro data contains records which contains information about individual entity, such as a person, a household, or an organization. For example, micro data are collected and used by various Government Agencies (*e.g., U.S. Census Bureau and Department of Motor Vehicles*) and by many commercial companies (*e.g., health organizations, insurance companies and retailers*). Data Mining is a common methodology to retrieve and discover useful hidden knowledge and information from the personal data. This has led to concerns that the personal data may be breached and misused. Therefore it is necessary to protect personal data through some privacy preserving techniques before conducting data mining.

Several micro data anonymization techniques have been proposed. The most used ones are generalization [10], [11] for k-anonymity [11] and bucketization [12], [8], [6] for ℓ-diversity [7]. In both the approaches, attributes are partitioned into three categories: 1) identifiers that uniquely identify an individual, such as *Name, Social Security Number*; 2) Quasi Identifiers (QI), can be linked with external data to uniquely identify at least one individual in the general population, such as *Birthdate, Sex, and Zip code*; 3) Sensitive Attributes (SA) is an attribute whose value for any particular individual must be kept secret from adversary such as *salary and disease*.

1.1 Need For Slicing

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. In Suppression, the values are not released at all.

It has been shown that [4], [5], [12] that generalization for k-anonymity losses considerable amount of information, especially for high dimensional data. This is due to the notion of k-anonymity is susceptible to homogeneity and background knowledge attacks. While bucketization [12], [8], [6] has better data utility than generalization and suppression, it has several limitations. First, bucketization does not prevent membership disclosure [9]. 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. As shown in [11], 87 percent of the individuals in the United States can be uniquely identified using only three attributes (*Birthdate, Sex, and Zip code*). A micro data (*e.g., census data*) usually contains many other attributes besides those three attributes. This means that the membership information of most individuals can be inferred from the bucketized table. Second, it does not show clear separation between QI attributes and Sensitive attributes. Third, bucketization breaks the attribute correlation between the QIs and the SAs by separating them.

This lead to design a novel data anonymization technique called *slicing* to improve the current state of 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 sorted to break the linking between different columns.

The prime 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, suppression and bucketization. Slicing preserve 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. On the other hand slicing can group some QI attributes with the SA, thus preserving correlations with the sensitive attribute. The key notion is that the slicing process ensures that for any tuple, there are generally multiple matching buckets.

1.2 Contribution & Organization

In this paper, we present a novel data anonymization technique for privacy preserving micro data publishing. Our contributions include the following.

First, we introduce privacy threats involved in publishing micro data in Section 2. Second, in Section 3, the architectural design for the proposed system is provided. Third, In Section 4 we show several methods used to preserve privacy against threats and framework for anonymizing technique.

Fourth, we show slicing as a new method for privacy preserving micro data publishing in Section 5. Slicing can be effectively used for preventing attribute disclosure, based on the privacy requirement of ℓ-diversity. Fifth, in Section 6 we formalize slicing algorithm and compered it with generalization and bucketization.

Finally, we conclude the paper and discuss future research work in Section 7.

2 MICRO DATA PUBLISHING

In this section, we discuss the risks involved in publishing micro data which has its wide applications in research.

2.1 Information Disclosure Risks

There are two types of privacy disclosure threats in publishing micro data. The first type is membership disclosure and the second type is attribute disclosure.

2.1.1 Membership Disclosure

When the data to be published is selected from a larger population and the selection criteria are sensitive (*e.g., when publishing datasets about diabetes patients for research purposes*), it is important to prevent an adversary from learning whether an individual's record is in the data or not.

2.1.2 Attribute Disclosure

Attribute disclosure occurs when new information about some individual is revealed, i.e., the released data makes it possible to infer the characteristics of an individual more accurately than it would be possible before the data release. In some scenarios, the adversary is assumed to know who is and who is not in the data, i.e., the membership information of individuals in the data. The adversary tries to learn additional sensitive information about the individuals.

3 ARCHITECTURE DESIGN

The architectural diagram for this paper is as follows.

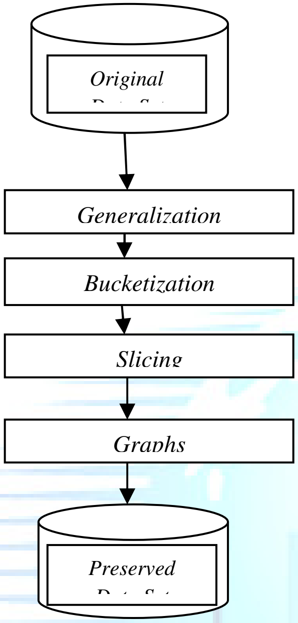

Figure 1: Architectural Design

4 ANONYMIZATION FRAMEWORKS

The various privacy models and anonymization methods are used and they are described below.

4.1 Privacy Models

A number of privacy models have been proposed. It includes k-anonymity and ℓ-diversity.

4.1.1 K-Anonymity

K-anonymity as the property that each record is indistinguishable with at least k-1 other records with respect to the quasi identifier. It requires that each QI group contains at least k records. The protection provided by k-anonymity is simple and easy to understand. If a table satisfies k-anonymity for some value k, then anyone who knows only the quasi identifier values of one individual with confidence greater than 1/k.

4.1.2 ℓ -Diversity

K-anonymity does not provide sufficient protection against attribute disclosure. Two attacks were identified: the homogeneity attack and the background knowledge attack. A QI group is said to have ℓ -diversity if there are at least ℓ "well represented" values for the sensitive attribute.

4.2 Anonymization Methods

Several popular anonymization methods are used at the earliest. In this section, we describe three methods.

4.2.1 Generalization and Suppression

Generalization replaces a value with a "less specific but semantically consistent" value that can be shown in table 1(b). Tuple suppression removes an entire record from the table.

4.2.2 Bucketization

Another anonymization method is bucketization (*also known as anatomy or permutation based anonymization*). The bucketization method first partitions tuples in the table into buckets and then separates the quasi identifiers with the sensitive attribute by randomly permuting the sensitive attribute values in each bucket is shown in table 1(c). The anonymized data consists of a set of buckets with permuted sensitive attribute values. It provides membership disclosure protection but does not preserve the data utility. The bucketization is based on the ℓ-diversity model. It preserves membership disclosure protection.

5 SLICING ALGORITHM

Our proposed technique is slicing which improves a step ahead the current state of art.

Given micro data table T and two parameters c and ℓ, the algorithm computes the sliced table that consists of c columns and satisfies the privacy requirement of ℓ-diversity. The algorithm consists of three phases: attribute partitioning, column generalization, and tuple partitioning. The sliced table is shown in table 1(d). We now describe the three phases.

5.1 Attribute Partitioning

The algorithm partitions attributes so that highly correlated attributes are in the same column. It provides privacy by breaking the associations between uncorrelated attributes. First, it measures the correlation between the attributes. Second, we use clustering to partition attributes into columns. Third, we calculate correlations between the sensitive attribute SA and each QI attribute. Then we rank the QI attributes by the decreasing order of their correlation with SA.

5.2 Column Generalization

Although column generalization is not a required phase, it can be useful in several aspects. If a column value is unique in a column, a tuple with this unique column value can only have one matching bucket. The main problem is that this unique column value can be identifying. In this case, it would be useful to apply column generalization to ensure that each column value appears with at least some frequency.

Age	Sex	Zip code	Disease
22	M	47906	Heart Disease
22	F	47906	Cancer
33	F	47905	Cancer
52	F	47905	Viral Infection
54	M	47302	Cancer
60	M	47302	Heart Disease
60	M	47304	Heart Disease
64	F	47304	Viral Infection

Table 1(a): Original Micro Data Table

5.3 Tuple Partitioning

In tuple partitioning phase, tuples are partitioned into buckets. The main part of the tuple partition algorithm is to check whether a sliced table satisfies ℓ-diversity.

Age	Sex	Zip code	Disease
[20-52]	*	4790*	Heart Disease
[20-52]	*	4790*	Cancer
[20-52]	*	4790*	Cancer
[20-52]	*	4790*	Viral Infection
[54-64]	*	4730*	Cancer
[54-64]	*	4730*	Heart Disease
[54-64]	*	4730*	Heart Disease
[54-64]	*	4730*	Viral Infection

Table 1(b): Generalization and Suppression

6 MODULES DESCRIPTION

In this section, we first give an example to illustrate slicing. We then formalize slicing, compare it with generalization and bucketization.

6.1 Formalization of Slicing

Slicing first partitions attributes into columns. Each column contains a subset of attributes. Vertical partitions the table. It also partition tuples into buckets. Each bucket contains a subset of tuples. This horizontally partitions the table. Within each bucket, values in each column are randomly permuted to break the linking between different columns. Column generalization ensures that one column satisfies the k-anonymity requirement. It can be used as an additional step in slicing. Each column contains much fewer attributes than the whole table, attribute partition enables slicing to handle high dimensional data. A key notion of slicing is that of matching buckets. Consider only one sensitive

Age	Sex	Zip code	Disease
22	M	47906	Cancer
22	F	47906	Heart Disease
33	F	47905	Viral Infection
52	F	47905	Cancer
54	M	47302	Viral Infection
60	M	47302	Cancer
60	M	47304	Heart Disease
64	F	47304	Heart Disease

Table 1(c): Bucketization

Age	Sex	Zip code	Disease
22	F	47906	Cancer
22	M	47905	Cancer
33	F	47906	Heart Disease
52	F	47905	Viral Infection
54	M	47302	Heart Disease
60	M	47302	Viral Infection
60	M	47304	Heart Disease
64	F	47304	Cancer

Table 1(d): The Sliced Table

attribute S, if the data contains multiple sensitive attributes, one can either consider them separately or consider their joint distribution. Exactly one of the c columns contains S. Without loss of generality, let the column that contains S be the last column Cc. This column is also called the sensitive column. All other columns $\{C_1, C_2, \ldots, C_{c-1}\}$ contain only QI attributes.

6.2 Comparison with Generalization

With Generalization and Suppression, Slicing preserves better data utility and has the ability to handle high dimensional data. By partitioning attributes into columns, slicing reduces the dimensionality of the data. Each column of the table can be viewed as a sub table with a lower dimensionality. Figure 2(a) shows the classification accuracy in percentage which used the target as sensitivie attribute, *disease*. Figure 2(b) shows the classification accuracy which used the learning attribute as QI, *Age, Sex and Zipcode*.

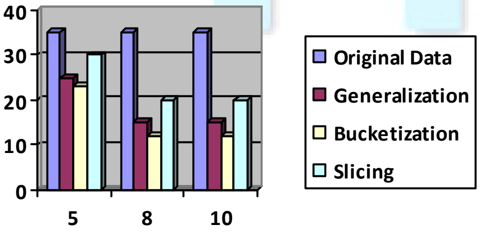

Figure 2(a): Classification Accuracy(%) based on SA

6.3 Comparison with Bucketization

To compare slicing with bucketization can be viewed as a special case of slicing, where there are exactly two columns: one column contains only the SA, and the other contains all the QIs. The advantages of slicing over bucketization can be understood as follows: First, by partitioning attributes into more than two columns, slicing can be used to prevent membership disclosure. An empirical evaluation on a real data set shows that bucketization does not prevent membership disclosure. Second, unlike bucketization, which requires a clear separation of QI attributes and the sensitive attribute, slicing can be used without such a separation. For dataset such as the census data, one often cannot clearly separate QIs from SAs because there is no single external public database that one can use to determine which attributes the adversary already knows. Slicing can be useful for such data.

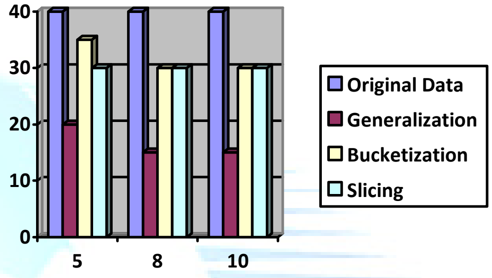

Figure 2(b): Classification Accuracy(%) based on QI

7 CONCLUSIONS AND FUTURE WORK

The general methodology proposed by this work is that: before anonymizing the data, one can analyze the data characteristics and use the characteristics in data anonymization. To ensure privacy in micro data publishing, a new technique slicing is introduced. It overcomes the limitations of generalization and bucketization and preserves better data utility while protecting against privacy threats.

This work motivates several directions for research work. First, in this paper we consider slicing where each attribute is in exactly one column. An extension is the notion of overlapping slicing, which duplicates an attribute in more than one column. Another direction is to design data mining tasks using anonymized data [2] computed by various anonymization techniques.

REFERENCES

[1] "Slicing: A New Approach for Privacy Preserving Data Publishing", Tiancheng Li, Ninghui Li, Senior Member, IEEE, IEEE Transactions on Knowledge and Data Mining, Vol 24, No 3, March 2012.

[2] A. Inan, M. Kantarcioglu, and E. Bertino, "Using Anonymized Data for Classification", Proc. IEEE 25th Int'l Conf. Data Eng. (ICDE), pp. 429-440, 2009.

[3] G. Chinita, Y. Tao, and P. Kalnis, "On the Anonymization of Sparse High Dimensional Data", Proc. IEEE 24th Int'l Conf Data Eng. (ICDE), pp. 715-724, 2008.

[4] C. Aggarwal, "On k-Anonymity and the Curse of Dimensionality", Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 901-909, 2005.

[5] D. Kifer and J. Gehrke, "Injecting Utility into Anonymized Data Sets", Proc. ACM SIGMOD Int'l Conf. Management of Data(SIGMOD), pp. 217-228, 2006.

[6] N. Koudas, D. Srivastava, T. Yu, and Q. Zhang", Aggregate Query Answering on Anonymized Tables", Proc. IEEE 23rd Int'l Conf. Data Eng. (ICDE), pp. 116-125, 2007.

[7] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "ℓ-Diversity: Privacy Beyond k-Anonymity", Proc. Int'l Conf Data Eng. (ICDE), p. 24, 2006.

[8] D.J. Martin, D. Kifer, A. Machanavajjhala, J.Gehrke, and J.Y. Halpern, "Worst-Case Background Knowledge for Privacy Preserving Data Publishing", Proc. IEEE 23rd Int'l Conf. Data Eng. (ICDE), pp. 126-135, 2007.

[9] M.E. Nergiz, M. Akzori, and C. Clifton, "Hiding the Presence of Individuals from Shared Databases", Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), pp. 665-676, 2007.

[10] P. Samarati, "Protecting Respondent's Privacy in Micro data Release", IEEE Transaction Knowledge and Data Eng., Vol. 13, no. 6, pp. 1010-1027, Nov./Dec. 2001.

[11] L. Sweeney, "k-Anonymity: A model for protecting privacy", Int'l J. Uncertainity Fuzziness and knowledge-Based Systems, vol. 10, no. 5, pp. 557-570, 2002.

[12] X. Xiao and Y. Tao, "Anatomy: Simple and Effective Privacy Presentation", Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 139-150, 2006.