An Efficient Sliced Data Algorithm Design for Data Protection

[1] **G. Hima Bindhu,** [2] **Dr. S. Sai Satyanarayana Reddy**

[1]M.Tech, CSE, LBRCE, Mylavaram, India

[2] Professor, CSE, LBRCE, Mylavaram, India

Abstract - Today, most enterprises are actively collecting and storing data in large databases. Privacy has become a key issue for progress in data mining. Maintaining the privacy of data mining has become increasingly popular because it allows sharing of privacy-sensitive data for analysis. Privacy-preserving data mining is used to safeguard sensitive information from unsanctioned disclosure. Privacy is an important issue in data publishing years because of the increasing ability to store personal data about users. Privacy-preserving data publishing (PPDP) provides methods and tools for publishing useful information while preserving data privacy. A number of techniques such as bucketization, generalization have been proposed to perform privacy-preserving data mining. Recent work has shown that generalization not support for high- dimensional data. Bucketization cannot prevent membership disclosure and does not apply for data that do not have a clear separation between quasi-identifying attributes and sensitive attributes. A new technique is introduced that is known as slicing, which partitions the data both horizontally and vertically. Slicing provides better data utility than generalization and can be used for membership disclosure protection. Slicing can handle high dimensional data. Also slicing can be used for attribute disclosure protection and develop an efficient algorithm for computing the sliced data that obey the l-diversity requirement. Slicing is more effective than bucketization in workloads involving the sensitive attribute. Another advantage of slicing can be used to prevent membership disclosure.

Keywords - *Data publishing, Generalization, Bucketization, Slicing.*

1. Introduction

There are number of techniques in data mining to manufacture easily and interestingly to get the data which consists of sensitive information also from the large amount of database. Confidentiality and privacy obligations are compromised by uncovering the information which is required while data preparation. The increaseof the disclosure risks of sensitive data because of data aggregation. When the data are collected data aggregation is used, it is analysed because it is taken from different sources and placed together. It is not enough to protect only private sensitive data secure but also it must be provided to openly known data. Especially when the data is anonymous, an individual's privacy comes under a struggle when the data previously composed,that data may roots the data miner or newly created datasets are able to make out specific individuals. The process of collecting the data from the record owners is data collection and the process of proving collected data publically for data recipient is data publishing. The process of protecting publishing information from the attackers is privacy preserving [1].

Privacy preserving publishing of microdata has been studied of great extent in recent years. The researchers and analysersget useful data from data publishing. The increasing ability to store personal data about users has made difficulty of privacy preserving data mining as more important issue. Now a day almost all the organizations are demanded for microdata publishing. Microdata contain records each of which contains information about an individual entity, such as a person or a household. There are many anonymization techniques have been proposed and most popular ones are generalization with k-anonymity and bucketization with l-diversity. The attributes in both methods are divided into three categories, some of them are identifiers that can be uniquely identified such as name or security number, some are quasi identifiers. These are the set of attributes are those that in the combination can be linked with the external information to re-identify such as birthdate, sex, zipcode and third category is sensitive attributes. These kinds of attributes are unknown to the opponent and are considered sensitive such as disease and salary. These are the three categories of attributes in microdata.

First identifiers are removed from the data and then partitions the tuples into buckets in both techniques. The generalization technique transforms the quasi identifying values in each bucket into less specific and semantically constant so that tuples in the same bucket cannot be distinguished by their quasi identifier values. One separates the sensitive attribute values from the quasi identifier values randomly permuting the SA values in the bucket in bucketization technique. The anonymized data consists of a set of buckets with permuted sensitive attribute values. When the patient data is shared, identity of patients must be protected. Earlier time we used techniques using k-anonymity and l-diversity. Subsisting works mainly considers datasets with a single sensitive attribute while patient data consists multiple sensitive attributes such as treatment and diagnosis. For preserving patient data both techniques are not so efficient. Therefore, here we are introducing a new technique for preserving patient data and publishing by slicing the data both horizontally and vertically. Data slicing can also be used to prevent membership disclosure and is efficient for high dimensional data and preserves better data utility.

2. Related Work

There are two main privacy preserving paradigms have been entrenched: k-anonymity [2], which prevents identification of individual records in the data, and l-diversity [3], which prevents the association of an individual record with the sensitive attribute value.

2.1. K-Anonymity

The database is said to be k-anonymous where attributes are suppressed or generalized until each row is identical with at least k-1 other rows. Thus k-anonymity prevents definite database linkages. It guarantees that the data released is accurate. This k-anonymity focuses on two techniques which are generalization and suppression. K-anonymity model was developed to protect released data from linking attack which causes the information disclosure. One of the emerging concepts in microdata protection is k-anonymity, which has been recently proposed as a property that captures the protection of microdata table with respect to possible re-identification of respondents to which the data refer. K-anonymity demands that every tuple in the microdata table released be indistinguishably related to no fewer than k respondents. One of the interesting aspects of this principle is its association with protection techniques that preserve the truthfulness of the data.

To perturb the input (the data) before it is mined is the first approach toward privacy protection in data mining. The drawback of the perturbation approach is that it lacks a formal framework for proving how much privacy is guaranteed. At the same time, a second branch of privacy preserving data mining was developed, using cryptographic techniques. Thus, it falls short of providing a complete answer to the problem of privacy preserving data mining. One definition of privacy which has come a long way in the public arena and accepted today by both legislators and corporations is that of k-anonymity [4]. The guarantee given by k-anonymity is that no information can be linked to groups of less than k individuals. The generalization for k-anonymity losses considerable amount of information especially for the high dimensional data.
The limitations of k-anonymity are: it does not hide whether a given individual is in the database, it reveals individual's sensitive attribute, it does not protect against attacks based on background knowledge, mere knowledge of the k-anonymity algorithm can violate privacy, it cannot be applied to high dimensional data without complete loss of utility.

2.2. L-Diversity

L-diversity is the next concept which is presented from the limitation of k-anonymity. The constraints can be putted on minimum number of distinct values by the l-diversity which can be seen within an equivalence class for any sensitive attribute. When there is l or more well represented values for the sensitive attribute then it is an equivalence class for the sensitive attribute then it is an equivalence class of l-diversity. This may prevent the extraction of useful information from the data, compromising utility.

Suppose say you have a group of k different records that all share a particular quasi identifier. That's good, in that an attacker cannot identify the individual based on the quasi identifier. But what if the value they are interested in, (for example the individual's medical diagnosis) is the same for every value in the group. The distribution of the target values within a group is referred to as l-diversity [5].

The limitations of l-diversity are, while the l-diversity principle represents an important step with respect to k-anonymity in protecting against attribute disclosure, it has several drawbacks. It is very difficult to achieve l-diversity and it also may not provide sufficient privacy protection.

3. Anonymization Techniques

There are two widely popular data anonymization techniques used. They are generalization [5], [6]and bucketization [7], [8]. The main difference between the two techniques lies in that bucketization does not generalize the quasi identifier attributes.

3.1. Generalization

Generalization is one of the commonly anonymized approaches, which replaces the quasi identifier values with the values that are less-specific but semantically consistent. Then, all quasi identifiers in a group would be generalized to the entire group extent in the QID space. If at least two transactions in a group have distinct values in a certain column (i.e., one contains an item and the other does not), then all information about that item in the current group is lost. The QID used in this process includes all possible items in the log.

There are three types of encoding schemes have been introduced for generalization. They are global recording, regional recording and local recording. The property gifted to global recording is that generalized value can be replaced with the multiple occurrences of the same value. Regional record partitions the domain space into non-intersect regions and data points in the same region are represented by the region they are in. Local recoding allows different occurrences of the same value to be generalized differently and does not have the above constraints. Due to the high dimensionality of the quasi identifier, with the number of possible items in the order of thousands, it is likely that any generalization method would incur extremely high information loss, rendering the data useless. In order for generalization to be effective, records in the same bucket must be close to each other so that generalizing the records would not lose too much information. However, in high dimensional data, most data points have similar distances with each other possible. This is an inherent problem of generalization that prevents effective analysis of attribute correlations. Generalization maintains the correctness of the data at the record level. The main problems with generalization are that it fails on high dimensional data due to curse dimensionality and it causes too much information loss due to uniform distribution assumption.

3.2. Bucketization

Bucketization partitions tuples in the table into buckets and then separates the quasi identifiers (QI) with the sensitive attributes by randomly permuting the sensitive attribute values in each bucket. A set of buckets with permuted sensitive attribute values called as anonymized data. In particular, it is used for anonymizing high dimensional data. Its main aim is to separation between quasi identifiers and sensitive attributes. In addition, because the exact values of all quasi identifiers are released, membership information is disclosed. While bucketization has better data utility than generalization, it has several limitations. They are as follows firstly 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. Secondly, it requires a clear separation between quasi identifiers and sensitive attributes. However, in many datasets it is unclear that which attributes are quasi identifiers and which are sensitive attributes. Thirdly, by separating sensitive attributes from the quasi identifier attributes, bucketization breaks the attribute correlations between them.

4. Slicing Algorithm

To improve the current state of art in this paper,we introduce a novel data anonymization techniquecalled slicing. Slicing partitions data set both horizontally and vertically. Vertical partitioning is done by grouping the attributes into columns based on correlation 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 bucket 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 among such attributes.

Slicing can handle high dimensional data and data without clear separation of quasi identifiers and sensitive attributes. It can be effectively used based on the privacy requirement of l-diversity for preventing attribute disclosure. L-diverse slicing ensures that the adversary cannot learn the sensitive value of any individual with a probability greater than 1/l. 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 sensitive attributes, preserving attribute correlation 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. Slicing first partitions attributes into columns. Each column contains a subset of attributes. Slicing also partitions tuples into buckets. Each bucket consists of subset of tuples. This horizontally partitions the table. Within each bucket, values in each column are randomly permutated to break the linking between different columns.

Slicing algorithm here we compare with the generalization and bucketization. Generally in privacy preservation there is loss of security. The privacy protection is impossible due to thepresence of adversary's background knowledge in the real life application. Data in its original form contains sensitive information about individuals.This data when published can violate the privacy. The current practice in data publishing relies mainly on policies and guidelines as to what type of data can be published and on agreements on the use of published data. The approach alone may leads to excessive data distortion or insufficient protection.

Privacy preserving data publishing provides methods and tools for publishing useful information while preserving data privacy. Many algorithms like bucketization, generalization have tried to preserve privacy however they exhibit attribute disclosure. So to overcome this problem an algorithm called slicing is used. An efficient algorithm is developed for computing sliced table that satisfies l-diversity.This algorithm consists of three phases: attribute partitioning, column generalization and tuple partitioning.

4.1. Attribute Partitioning

This algorithm partitions attributes so that highly correlated attributes are in the same column. This is good for both utility and privacy. In terms of data utility, grouping highly correlated data preserves the correlations among those attributes. In terms of privacy, the association of uncorrelated attributes presents higher identification risks than the association of highly correlated attributes because the associations of uncorrelated attributes is much less frequent and thus more identifiable. Therefore, it is better to break the associations between uncorrelated attributes in order to protect privacy.

4.2. Column Generalization

First column generalization may be required for identity or membership disclosure protection. If a column value is unique in a column, a tuple with this unique column value can only have one matching bucket. This is not good for privacy protection as in the case of the generalization or bucketization where each tuple can belong to only one equivalence class or bucket.

4.3. Tuple Partitioning

In tuple partitioning phase tuples are partitioned into buckets. The algorithm consists of two data structures: a queue of buckets Q and a set of sliced buckets SB. Initially, Q contains only one bucket which includes all the tuples and SB is empty. For each iteration, the algorithm removes a bucket from Q and splits it into two buckets. Ifthe sliced table after split satisfies the l-diversity, then the algorithm puts the two buckets at the end of the queue Q, otherwisewe cannot split the bucket anymore and the algorithm puts the bucket into SB. When the queue Q becomes empty, it is clear that we have computed the sliced table. The set of sliced buckets is SB.

```
Algorithm tuple-partition (T, ℓ)
$Q = \{T\}; SB = \emptyset$.
While Q is not empty
Remove the first bucket B from Q; $Q = Q - \{B\}$.
Split B into two buckets B1 and B2, as in Mondrian.
if diversity-check (T, $Q \cup \{B1, B2\} \cup SB$, ℓ)
$Q = Q \cup \{B1, B2\}$.
else $SB = SB \cup \{B\}$.
return SB.
```

```
Algorithm diversity-check (T,T0, ℓ)

for each tuple $t \in T$, $L[t] = \emptyset$.
for each bucket B in T0
record f(v) for each column value v in bucket B.
for each tuple $t \in T$
calculate p(t,B) and find D(t,B).
$L[t] = L[t] \cup \{hp(t,B),D(t,B)i\}$.
for each tuple $t \in T$
calculate p(t, s) for each s based on L[t].
if $p(t, s) \geq 1/\ell$, return false.
return true.
```

5. Conclusion

The slicing strategy overcomes the limitations of generalization and bucketization methods. It preserves better utility while protecting against privacy threats where each attribute is exactly in one column. An extension of slicing is overlapping slicing which duplicates an attribute in more than one column. The proposed tuple grouping algorithm is optimized l-diversity check algorithm which obtains more effective tuple grouping and provides the secure data. Another advantage of slicing is that it can handle high dimensional data.Its future work can be as privacy preservation as the big issue, large number of datasets is increasing security to such data must be available. Therefore, as the term privacy entered encryption and decryption and compression can further be done for such databases.

References

[1] Brickell.J and Shmatikov, "The Cost of Privacy: Destruction of Data Mining Utility in Anonymized Data Publishing", Proc.ACM SIGKDD int'l conf. Knowledge Discovery Data Mining (KDD), 2008.

[2] D. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke and J. Halphern, " Worst-Case Background Knowledge for Privacy preserving data publishing." In ICDE, 2006.

[3] A. Machanavajjhala, D. Kifer, J. Gehrke and M. Venkitasubramaniam, "l-diversity: Privacy Beyond k-anonymity" in ICDE, 2007.

[4] L. Sweeney, "k-anonymity: A Model For Protecting Privacy", Int'l J. Uncertainty Fuzziness and Knowledge-Based Systems, vol. 10, no. 5, pp. 557-570, 2002.

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

[6] He.Y and Naughton.J, "Anonymization of set-valued Data via Top-Down, local generalization," Proc. IEEE 25th Int"l Conf.Data Engineering (ICDE), 2009.

[7] Aggarwal.C, "On K-Anonymity and the Curse of Dimensionality," Proc. Int"l Conf.Very Large Databases (VLDB), 2005.

[8] Li.N, Li.T, "Slicing: The new Approach for privacy Preserving Data Publishing", Proc.ACM SIGKDD Int"l Conf.Knowledge Discovery and Data Mining (KDD), 2009.