CARVING: A Novel Technique to Prevent Membership Disclosure

G.Siva Kumar[#1], K.V.Srinivasa Rao[*2]

[#]M.Tech Scholar
Department of CSE ,
Prakasam Engineering College,
Kandukur ,Andhra Pradesh,India.
[1]*sivacse1988@gmail.com*

[*]Assistant Professor
Department of CSE,
Prakasam Engineering College,
Kandukur ,Andhra Pradesh,India.
[2]srinivasa_rao_kalva@yahoo.co.in

Abstract

We introduced a very new technique called carving, which provides better and accurate result than generalization and bucketization techniques, it can also partition the data in both horizontal and vertical directions and also it can handle very high-dimensional data. In this paper carving technique can be used for attribute disclosure protection and develop an algorithm for computing the carved data that obey the ℓ-diversity requirement. It also demonstrates that carving used to prevent membership disclosure.

Keywords

Carving, Bucketization, Generalization, Computing, Attribute Disclosure.

1. Introduction

Microdata is a type of records each of which contains information about an individual entity, such as a person, an item, an diseases name, a household, or an organization. Several microdata anonymization techniques have been proposed till to date. The most popular ones are generalization [1, 2] which was used for k-anonymity [2] and bucketization [3, 4, 5] which was used for ℓ-diversity [6]. In both approaches, the attributes are partitioned into following 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), and which, can potentially identify an individual, e.g., Birth date, Sex, and Zip code;

3) Some attributes are Sensitive Attributes (SAs), which are unknown to the adversary and are considered sensitive, such as Disease and Salary.

It has been clearly shown [7, 8, 3] that generalization for k- anonymity technique greatly

losses considerable amount of information, especially for high-dimensional data. This is due to the three reasons. First, the reason for great loss of generalization for k-anonymity is it suffers from the curse of dimensionality. For generalization to be effective in its nature, records present in the same bucket must be very close and near to each other so that generalizing the records would not lose too much of information. However, when we go with high- dimensional data, most data points will have similar distances with each other, forcing a great amount of generalization to satisfy k-anonymity even for relative small k's records. Secondly, in order to perform data analysis operation or data mining tasks on the generalized table, the data analyst has to make the uniform distribution assumption that every value in a generalized interval/set is equally possible, as no other distribution assumption can be justified. This reduces the data utility of the generalized data. Third, because each attribute is generalized separately, correlations between different attributes are lost, while bucketization [3, 4, 5] has better data utility than generalization, it has several limitations. First, bucketization does not prevent membership disclosure [9]. Second, bucketization requires a clear separation between QIs and SAs Third, by separating the sensitive attribute from the QI attributes, bucketization breaks the attribute correlations between the QIs and the SAs.

In this paper, we introduce an new anonymization technique called carving technique. It partitions the whole dataset both vertically and horizontally. Vertical partitioning of data set is done by grouping attributes present in that data set into columns based on the correlations among the attributes. Each column contains a subset of attributes that are correlated. Horizontal partitioning of data set attributes is done by grouping tuples present in that table into buckets. Finally, within each bucket present in that table, values in each column are randomly taken and sorted to break the linking between different columns. We finally show that carving can be used for preventing attribute disclosure, based on the privacy requirement of ℓ-diversity.

2. Carving Technique

In this section, we mainly discuss about carving technique and then, compare it with other existing techniques like generalization and bucketization, and we also discuss privacy threats that carving can address. Table 1 clearly shows an example for Microdata table and its anonymized versions using various anonymization techniques The original table is shown in Table 1(a). The three QI attributes present in that table are {Age, Sex, Zip code}, and the sensitive attribute SA among them is Disease. A generalized table that satisfies 4-anonymity principle is shown in Table 1(b), a bucketized table that satisfies 2-diversity is shown in Table 1(c), a generalized table where each attribute value is replaced with the values in the bucket is shown in Table 1(d), and two carved tables are shown in Table 1(e) and 1(f). Carving first partitions the attributes into columns. Each column contains a subset of attributes. This vertically partitions the table. For example, the carved table in Table 1(f) contains 2 columns: the first column contains {Age, Sex } and the second column contains {Zipcode, Disease}. The carved table shown in Table 1(e) contains 4 columns, where each column contains exactly one attribute. Carving also partition tuples into buckets. Each bucket contains a subset of tuples. This horizontally partitions the table. For example, both carved tables in Table 1(e) and Table 1(f) contain 2 buckets, each containing 4 tuples. In each bucket, the values in each column are randomly permutated to break the linking between different columns.

3. Attribute Disclosure Protection Mechanism

In this phase we show how slicing can be used to prevent attribute disclosure, based on the privacy requirement of ℓ- diversity.

3. 1 Example

Here an example illustrating how carving satisfies ℓ-diversity [6] where the attribute is "Disease". The sliced table shown in Table 1(f) satisfies 2-diversity. Consider tuple t1 with QI values (22,M, 47906). In order to determine t1's sensitive value, one has to examine t1's matching

buckets. By examining the first column (Age,Sex) in Table 1(f), we know that t1 must be in the first bucket B1 because there are no matches of (22,M) in bucket B2. Therefore, one can conclude that t1 cannot be in bucket B2 and t1 must be in bucket B1. Then, by examining the Zipcode attribute of the second column (Zipcode,Disease) in bucket B1, we know that the column value for t1 must be either (47906, dyspepsia) or (47906, flu) because they are the only values that match t1's zipcode 47906. Note that the other two column values have zipcode 47905. Without additional knowledge, both dyspepsia and flu are equally possible to be the sensitive value of t1. Therefore, the probability of learning the correct sensitive value of t1 is bounded by 0.5. Similarly, we can verify that 2-diversity is satisfied for all other tuples in Table 1(f). corrupted (no matching hash of the payload), dropped, or delayed (entry is not matched within γ).

Table 1: An original micro data table and its anonymized versions.

Age	Sex	Zipcode	Disease
[20-52]	*	4790*	dyspepsia
[20-52]	*	4790*	flu
[20-52]	*	4790*	flu
[20-52]	*	4790*	bronchitis
[54-64]	*	4730*	flu
[54-64]	*	4730*	dyspepsia
[54-64]	*	4730*	dyspepsia
[54-64]	*	4730*	gastritis

(b) The generalized table

Age	Sex	Zipcode	Disease
22	M	47906	flu
22	F	47906	dyspepsia
33	F	47905	bronchitis
52	F	47905	flu
54	M	47302	gastritis
60	M	47302	flu
60	M	47304	dyspepsia
64	F	47304	dyspepsia

(c) The bucketized table

Age	Sex	Zipcode	Disease
22:2,33:1,52:1	M:1,F:3	47905:2,47906:2	dysp.
22:2,33:1,52:1	M:1,F:3	47905:2,47906:2	flu
22:2,33:1,52:1	M:1,F:3	47905:2,47906:2	flu
22:2,33:1,52:1	M:1,F:3	47905:2,47906:2	bron.
54:1,60:2,64:1	M:3,F:1	47302:2,47304:2	flu
54:1,60:2,64:1	M:3,F:1	47302:2,47304:2	dysp.
54:1,60:2,64:1	M:3,F:1	47302:2,47304:2	dysp.
54:1,60:2,64:1	M:3,F:1	47302:2,47304:2	gast.

(d) Multiset-based generalization

Age	Sex	Zipcode	Disease
22	F	47906	flu
22	M	47905	flu
33	F	47906	dysp.
52	F	47905	bron.
54	M	47302	dysp.
60	F	47304	gast.
60	M	47302	dysp.
64	M	47304	flu

(e) One-attribute-per-column slicing

(Age,Sex)	(Zipcode,Disease)
(22,M)	(47905,flu)
(22,F)	(47906,dysp.)
(33,F)	(47905,bron.)
(52,F)	(47906,flu)
(54,M)	(47304,gast.)
(60,M)	(47302,flu)
(60,M)	(47302,dysp.)
(64,F)	(47304,dysp.)

(f) The sliced table

4. Carving Algorithms

It consists of three phases: attribute partitioning, column generalization, and tuple partitioning.

4.1 Attribute Partitioning

In attribute partitioning, we find the correlations between pairs of attributes and then cluster attributes .

4.2 Column Generalization

In this phase, tuples are generalized to satisfy frequency requirement. It can be applied on the only the attributes in one column to provide the anonymity requirement.

4.3 Tuple Partitioning

In this phase, tuples are partitioned into buckets. We use the Mondrian [10] algorithm for tuple partition. No generalization is applied to the tuples.

5. Membership Disclosure Protection

To protect membership information, it is required that, in the anonymized data, a tuple should have a similar frequency as a tuple that is not in the original data. Otherwise it can differentiate tuples in the original data from tuples not in the original data. Let E the set of tuples in the original data and let F be the set of tuples that are not in the original data. Ds be the sliced data. Given Ds a tuple t, the goal of membership disclosure is to determine whether $t \in E$ or $t \in F$. In order to distinguish the tuples we find their differences. If $t \in E$, t must have at least one matching buckets in Ds. To protect membership information, we must sure that at least some tuples in E should also have matching buckets. Otherwise, the tuples can be differentiated.

6. Conclusion

The main purpose of this paper is to before anonymizing the data, one can analyze the data characteristics and use these characteristics in data anonymization. In this paper, we consider each carving attribute is in exactly one column which duplicates an attribute in more than one columns. This release more attributes correlations. Second, is membership disclosure protection . Our experiments show that random grouping is not very effective. Third, carving is a promising technique for handling high dimensional data. By partitioning attributes into columns, we protect privacy by breaking the association of uncor - related attributes and preserve data utility by preserving the association between highly-correlated attributes. Finally, while a number of anonymization techniques have been designed, it remains an open problem on how to use the anonymized data.

7. References

[1] P. Samarati. Protecting respondent's privacy in microdata release. TKDE, 13(6):1010–1027, 2001.

[2] L. Sweeney. K-anonymity: A model for protecting privacy. Int. J. Uncertain. Fuzz, 10(5):557–570, 2002.

[3]X. Xiao and Y. Tao. Anatomy: simple and effective privacy preservation. In VLDB, pages 139–150, 2006.

[4] D. J. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J. Y. Halpern. Worst-case background knowledge for privacy-preserving data publishing. In ICDE, pages 126–135, 2007.

[5] N. Koudas, D. Srivastava, T. Yu, and Q. Zhang. Aggregate query answering on anonymized tables. In ICDE, pages 116–125, 2007.

[6] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. ℓ-diversity: Privacy beyond k-anonymity. In ICDE, page 24, 2006.

[7] C. Aggarwal. On k-anonymity and the curse of

dimensionality. In VLDB, pages 901–909, 2005.

[8] D. Kifer and J. Gehrke. Injecting utility into anonymized datasets. In SIGMOD, pages 217–228, 2006.

[9] M. E. Nergiz, M. Atzori, and C. Clifton. Hiding the presence of individuals from shared databases. In SIGMOD, pages 665–676, 2007.

[10] K. LeFevre, D. DeWitt, and R. Ramakrishnan. Mondrian multidimensional k-anonymity. In CDE, page 25, 2006.

[11] A. Inan, M. Kantarcioglu, and E. Bertino. using anonymized data for classification. In ICDE, 2009.

8. About the Authors

Siva Gullipalli is currently pursuing his M.Tech in Computer Science and Engineering at Prakasam Engineering College, Kandukur, Andhra Pradesh. His area of interests include Networks, Data Mining