International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

A Better Approach for Privacy Preserving Data Publishing by Slicing

Mohd Faquroddin[1], G. Kiran Kumar[2]

[1]M. Tech Student, Department of CSE, Anurag Group of Institutions, Hyderabad, India

[2]Assistant Professor, Department of CSE, Anurag Group of Institutions, Hyderabad, India

Abstract: ***In order to preserve Micro data publishing, several anonymization techniques, such as generalization and bucketization, have been designed. Recent work has shown that generalization loses considerable amount of information, especially for high dimensional data. 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, a novel technique called slicing is presented, which partitions the data both horizontally and vertically. The 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 and also slicing can be used for attribute disclosure protection and develop an efficient algorithm for computing the sliced data that obey the diversity requirement. Several experiments confirm that slicing preserves better utility than generalization and is more effective than bucketization in workloads involving the sensitive attribute and also demonstrate that slicing can be used to prevent membership disclosure.***

Keywords: Data Anonymization, Privacy Preservation, Data publishing, Data Security, Generalization, Bucketization.

1. Introduction

Private companies and government sectors are sharing micro data to facilitate pure research and statistical analysis. Individuals' privacy should be protected. Micro data contains sensitive values of record owners. Generally, microdata stored in table format (T). Adversaries (attackers) associates more than two dataset and apply their background knowledge for deducing the sensitive information. Certain attributes are associates with external knowledge to identify the individual's records indirectly. Such attributes are called Quasi Identifiers(QI). Quasi identifiers are associated with sensitive attribute(S) which should not be disclosed. Data leakage occurs by association of quasi identifiers and background knowledge.

There are two types of disclosure namely attribute disclosure and identity disclosure. Privacy-preserving publishing of microdata has been studied extensively in recent years. Microdata contains records each of which contains information about an individual entity, such as a person, a household, or an organization. Several microdata anonymization techniques have been proposed.

Data Anonymization is a technology that converts clear text into a non-human readable form. Data Anonymization technique for privacy-preserving data publishing has received a lot of attention in recent years. Detailed data (also called as microdata) contains information about a person, a household or an organization. Most popular Anonymization techniques are Generalization and Bucketization. Data Anonymization enables the transfer of information across a boundary, such as between two departments within an agency or between two agencies, while reducing the risk of unintended disclosure, and in certain environments in a manner that enables evaluation and analytics post-Anonymization.

Generalization is one of the commonly anonymized approaches, which replaces quasi-identifier values with values that are less-specific but semantically consistent. Then, all quasi-identifier values 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. 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. To perform data analysis 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 significantly reduces the data utility of the generalized data.

Bucketization The term bucketization, is to partition the tuples in T into buckets, and then to separate the sensitive attribute from the nonsensitive ones by randomly permuting the sensitive attribute values within each bucket. The sanitized data then consists of the buckets with permuted sensitive values.

2. Literature Survey

The privacy preserving data mining problem has gained considerable importance in recent years be- cause of the vast amounts of personal data about individuals stored at different commercial vendors and organizations. In many cases, users

are willing to divulge information about them only if the privacy of the data is guaranteed. Thus methods need to be proposed to mask the sensitive information in the records. This creates the natural challenge of mining the data in an effective way with a limited data representation. A variety of techniques have recently been proposed both to represent and mine the data without loss of privacy. Some important techniques for privacy include methods such as perturbation, k-anonymity, condensation, and data hiding with conceptual reconstruction. In this paper, we will analyze the k-anonymity approach for the high dimensional case. The idea behind this class of approaches is that many of the fields in the data can be treated as pseudo-identifiers or quasi-identifiers which can be matched with publically known data in order to identify individuals. For example, a commercial database containing birthdates, gender and zip-codes can be matched with voter registration lists in order to identify the individuals precisely. Another related class of methods to deal with the issue of k-anonymity is the k-in distinguish ability approach. The k-anonymity and k-in distinguish ability approaches are briefly discussed below:

In the k-anonymity approach, generalization techniques are applied in order to mask the exact values of attributes. For example, a quantitative attribute such as the age may only be specified to a range. This is referred to as attribute generalization. By defining a high enough level of generalization on each attribute, it is possible to guarantee k-anonymity. On the other hand, attribute generalization also leads to a loss of information. In the k-in distinguish ability model, clustering techniques are used in order to construct indistinguishable groups of k records. The statistical characteristics of these clusters are used to generate pseudo-data which is used for data mining purposes. While such pseudo-data does not rep-resent the true data records, it is useful for most modeling purposes, since it reflects the original distribution of the records. There are some advantages in the use of pseudo-data, in that it is more resistant to hacking, and it does not require any modification of the underlying data representation as in a generalization approach.

3. System Architecture & Problem Statement

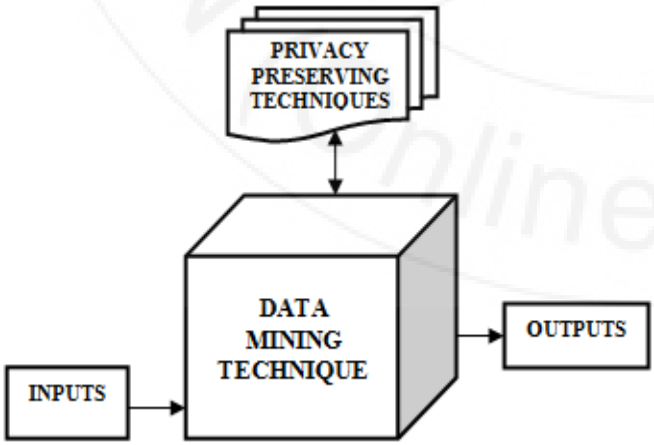

Figure : Architecture of Privacy Preserving in Data Mining

The privacy protection is impossible due to the presence of the adversary's background knowledge in real life application. Data in its original form contains sensitive information about individuals. These data when published violate the privacy. The current practice in data publishing relies mainly on policies and guidelines as to what types of data can be published and on agreements on the use of published data. Several anonymization techniques, such as generalization and bucketization, have been designed for privacy preserving micro data publishing. Recent work has shown that generalization loses considerable amount of information, especially for high dimensional data.

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. In this paper, we present a novel technique called slicing, which partitions the data both horizontally and vertically. We show that 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. Generalization for k-anonymity losses considerable amount of information, especially for high-dimensional data.

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. Bucketization requires a clear separation between QIs and SAs. However, in many data sets, it is unclear which attributes are QIs and which are SAs.

4. Proposed System

In A novel data anonymization technique called slicing is introduced 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. It preserves better data utility than generalization. It preserves more attribute correlations with the SAs than bucketization. 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.

5. Implementation

Original Data:

We conduct extensive workload experiments. Our results confirm that slicing preserves much better data utility than generalization. In workloads involving the sensitive attribute, slicing is also more effective than bucketization. In some classification experiments, slicing shows better performance than using the original data.

Table 1: Original Table

AGE	SEX	ZIPCODE	DISEASE
27	**F**	56805	CANCER
27	**M**	56805	FLU
38	**M**	56804	THYROID
57	**M**	56804	CANCER
55	**F**	56201	TYPHOID
60	**F**	56201	CANCER
60	**F**	56203	FLU
65	**M**	56203	FLU

Generalized Data

Generalized Data, in order to perform data analysis 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 significantly reduces the data utility of the generalized data.

Table 2: Generalized Table

AGE	SEX	ZIPCODE	DISEASE
[27-57]	*	5680*	CANCER
[27-57]	*	5680*	FLU
[27-57]	*	5680*	THYROID
[27-57]	*	5680*	CANCER
[55-65]	*	5620*	TYPHOID
[55-65]	*	5620*	CANCER
[55-65]	*	5620*	FLU
[55-65]	*	5620*	FLU

Bucketized Data

we show the effectiveness of slicing in membership disclosure protection. For this purpose, we count the number of fake tuples in the sliced data. We also compare the number of matching buckets for original tuples and that for fake tuples. Our experiment results show that bucketization does not prevent membership disclosure as almost every tuple is uniquely identifiable in the bucketized data.

Table 3: Bucketized Table

AGE	SEX	ZIPCODE	DISEASE
27	F	56805	FLU
27	M	56805	CANCER
38	M	56804	CANCER
57	M	56804	THYROID
55	F	56201	CANCER
60	F	56201	FLU
60	F	56203	FLU
65	M	56203	TYPHOID

Multiset-based Generalization Data

We observe that this multiset-based generalization is equivalent to a trivial slicing scheme where each column contains exactly one attribute, because both approaches preserve the exact values in each attribute but break the association between them within one bucket.

Table 4: Multiset-Based Genaralized table

AGE	SEX	ZIPCODE	DISEASE
27:2,38:1,57:1	F:1,M:3	56805:2,56804:2	CANCER
27:2,38:1,57:1	F:1,M:3	56805:2,56804:2	FLU
27:2,38:1,57:1	F:1,M:3	56805:2,56804:2	THYROID
27:2,38:1,57:1	F:1,M:3	56805:2,56804:2	CANCER
55:1,60:2,65:1	F:3,M:1	56201:2,56203:2	TYPHOID
55:1,60:2,65:1	F:3,M:1	56201:2,56203:2	CANCER
55:1,60:2,65:1	F:3,M:1	56201:2,56203:2	FLU
55:1,60:2,65:1	F:3,M:1	56201:2,56203:2	FLU

One-attribute-per-Column Slicing Data

We observe that while one-attribute-per-column slicing preserves attribute distributional information, it does not preserve attribute correlation, because each attribute is in its own column. In slicing, one groups correlated attributes together in one column and preserves their correlation.

For example, in the sliced table shown in Table correlations between Age and Sex and correlations between Zipcode and Disease are preserved. In fact, the sliced table encodes the same amount of information as the original data with regard to correlations between attributes in the same column.

Table 5: One Attribute per Column Slicing

AGE	SEX	ZIPCODE	DISEASE
27	M	56805	FLU
27	F	56804	THYROID
38	M	56805	CANCER
57	M	56804	CANCER
55	F	56201	CANCER
60	M	56203	FLU
60	F	56201	FLU
65	F	56203	TYPHOID

Sliced Data

Another important advantage of slicing is its 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. Slicing is also different from the approach of publishing multiple independent sub-tables in that these sub-tables are linked by the buckets in slicing.

Table 6: The Sliced Table

AGE,SEX	ZIPCODE,DISEASE
(27,F)	(56805, FLU)
(27,M)	(56805,CANCER)
(38,M)	(56804,CANCER)
(57,M)	(56804,THYROID)
(55,F)	(56201,CANCER)
(60,F)	(56201,FLU)
(60,F)	(56203,FLU)
(65,M)	(56203,TYPHOID)

6. Conclusion

Slicing is a promising technique for handling high-dimensional data. By partitioning attributes into columns, We protect privacy by breaking the association of uncorrelated attributes and preserve data utility by preserving the association between highly-correlated attributes. For example, slicing can be used for anonymizing transaction databases, which has been studied recently.

Finally, while a number of anonymization techniques have been designed, it remains an open problem on how to use the anonymized data. In our experiments, we randomly generate the associations between column values of a bucket. This may lose data utility. Another direction to design data mining tasks using the anonymized data computed by various anonymization techniques. In future an extension is the notion of overlapping slicing, which duplicates an attribute in more than one columns. This releases more attribute correlations. For example, in Table, one could choose to include the Disease attribute also in the first column.

That is, the two columns are {Age,Sex,Disease} and {Zipcode,Disease}. This could provide better data utility, but the privacy implications need to be carefully studied and understood. It is interesting to study the tradeoff between privacy and utility.

References

[1] J. Brickell and V. Shmatikov. The cost of privacy: destruction of data-mining utility in anonymized data publishing. In KDD, pages 70–78, 2008.
[2] G. Ghinita, Y. Tao, and P. Kalnis. On the anonymization of sparse high-dimensional data. In ICDE, pages 715–724, 2008.
[3] 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.
[4] L. Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. Int. J. Uncertain. Fuzz., 10(6):571–588, 2002.
[5] X. Xiao and Y. Tao. Anatomy: simple and effective privacy preservation. In VLDB, pages 139–150, 2006.
[6] X. Xiao and Y. Tao. Output perturbation with query relaxation. In VLDB, pages 857–869, 2008.
[7] Y. Xu, K. Wang, A. W.-C. Fu, and P. S. Yu. Anonymizing transaction databases for publication. In KDD, pages 767–775, 2008.
[8] Benjamin C.M.Fung,KE Wang,Ada Wai-Chee Fu and Philip S. Yu,(2010) "Introduction to Privacy- Preserving Data Publishing Concepts and techniques", ISBN:978-1-4200-9148-9,2010
[9] Raymond Wong, Jiuyong Li,Ada Fu and Ke wang,(2009), "(α,k)-anonymous data publishing", Journal Intelligent Information System, pp209-234.
[10] L. Sweeney,(2002) "An Achieving k-Anonymity Privacy Protection Using Generalization and Suppression", International Journal of Uncertainty, Fuzziness and Knowledge-Based System,pp571- 588
[11] Gabriel Ghinita, Member IEEE, Panos Kalnis, Yufei Tao," *Anonymous Publication of Sensitive Transactional Data"* in Proc. Of IEEE Transactions on Knowledge and Data Engineering February 2011 (vol. 23 no. 2) pp. 161-174.
[12] G.Ghinita, 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.
[13] T. Li, N. Li, J. Zhang, and I. Molloy, "Slicing: A new approach for privacy preserving data publishing," IEEE *Transactions on Knowledge and Data Engineering*, vol. 24, pp. 561 – 574, 2012.
[14] P. Samarati, "Protecting respondents' identities in microdata release," *IEEE Trans. on Knowl. and Data Eng.*, vol. 13, no. 6, pp. 1010–1027, 2001.
[15] G.Ghinita,Y.Tao,and P.Kalnis.On the anonymization of sparse high-dimensional data.In ICDE,pages 715-724,2008.
[16] A.Inan,M.Kantarcioglu,&E.Bertino.Using anonymized data for classification.In ICDE,2009 .
[17] J.Brickell and V,Shmatikov.The cost of privacy:destruction of data-mining utlity in anonymized data publishing.In KDD,pages 70-78,2008.

Author Profile

Mohd Faquroddin received the B.E degree in Information Technology from Osamnia University in 2012 and pursuing M.Tech. degree in Computer science and Engineering from Anurag Group of Institutions (Formerly CVSR College of Engineering) JNTU Hyderabad.

G. Kiran Kumar working as assistant professor in Computer science and Engineering from Anurag Group of Institutions (Formerly CVSR College of Engineering) JNTU Hyderabad.ate.