International Journal of Multidisciplinary Research and Development
Online ISSN: 2349-4182 Print ISSN: 2349-5979, Impact Factor: RJIF 5.72
www.allsubjectjournal.com
Volume 3; Issue 6; June 2016; Page No. 54-56

Slicing techniques: A new approach to privacy preserving data publishing

Jeevanandhini M, * Ruby gnanaselvam C

M.phil, Full time, computer data mining, bharathiar university, Coimbatore, India

Abstract
Several K-anonymity techniques, such as generalization and bucketization, have been designed for privacy preserving microdata 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 overlapping Slicing Techniques, which partitions the data both horizontally and vertically. We show that overlapping Slicing Techniques preserves better data utility than generalization and can be used for membership disclosure protection. Another important advantage of overlapping Slicing Techniques is that it can handle high-dimensional data storage.

Keywords: K-anonymity, generalization, Bucketization, Slicing Techniques, high-dimensional data storage

Introduction

We show how overlapping Slicing Techniques can be used for attribute disclosure protection and develop an efficient algorithm for computing the sliced data that obey the ‘diversity requirement. Our workload experiments confirm that overlapping Slicing Techniques preserves better utility than generalization and is more effective than bucketization in workloads involving the sensitive attribute. Our experiments also demonstrate that overlapping Slicing Technique scan be used to prevent membership disclosure. We consider the collaborative data publishing problem for anonym zing horizontally partitioned data at multiple data providers. We consider a new type of “insider attack” by colluding data providers who may use their own data records (a subset of the overall data) in addition to the external background knowledge to infer the data records contributed by other data providers. The paper addresses this new threat and makes several contributions. First, we introduce the notion of m-privacy, which guarantees that the K-anonymity data satisfies a given privacy constraint against any group of up to m colluding data providers. Second, we present heuristic algorithms exploiting the equivalence group monotonicity of privacy constraints and adaptive ordering techniques for efficiently checking m-privacy given a set of records. Finally, we present a data provider-aware K-anonymity algorithm with adaptive m-privacy checking strategies to ensure high utility and m-privacy of K-anonymity data with efficiency. Experiments on real-life datasets suggest that our approach achieves better or comparable utility and efficiency than existing baseline algorithms while providing m-privacy guarantee. self-identify when requested by a law enforcement officer. In recent years, due to increase in ability to store personal data about users and the increasing sophistication of data mining algorithms to leverage this information the problem of privacy preserving data mining has become more important. A number of anonymization techniques have been researched in order to perform privacy-preserving data mining. Data anonymization technique for privacy-preserving data publishing has received a lot of attention in recent years. Detailed data (also called as micro data) contains information about a person, a household or an organization. Most popular anonymization techniques are *Generalization and Bucketization.* There are number of attributes in each record which can be categorized as 1) *Identifiers* such as *Name or Social Security Number* are the attributes that can be uniquely identify the individuals. 2) Some attributes may be Sensitive Attributes (SAs) such as *disease* and *salary* and 3) some may be Quasi-Identifiers (QI) such as *zip code*, *age*, and *sex* whose values, when taken together, can potentially identifying individual.

Existing system

Several micro data K-anonymity 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 Zip code;
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.

Future work

We introduce a novel data K- anonymity technique called overlapping Slicing Techniques to improve the current state of the art. Overlapping Slicing Techniques 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 overlapping Slicing Techniques 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. Overlapping Slicing Techniques preserves utility because it groups highly correlated attributes together, and preserves the correlations between such attributes. Overlapping Slicing Techniques 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; overlapping Slicing Techniques, on the other hand, can group some QI attributes with the SA, preserving attribute correlations with the sensitive attribute. The key intuition that overlapping Slicing Techniques provides privacy protection is that the overlapping Slicing Techniques process ensures that for any tuple, there are generally multiple matching buckets.

We consider the collaborative data publishing setting with horizontally partitioned data across multiple data providers, each contributing a subset of records Ti. As a special case, a data provider could be the data owner itself who is contributing its own records. This is a very common scenario in social networking and recommendation systems. Our goal is to publish a K-anonymity view of the integrated data such that a data recipient including the data providers will not be able to compromise the privacy of the individual records provided by other parties.

Techniques

When the data are distributed among multiple data providers or data owners, two main settings are used for K-anonymity. One approach is for each provider to anonymize the data independently (anonymize -and-aggregate, Figure 1A), which results in potential loss of integrated data utility. A more desirable approach is collaborative data publishing which anonymizes data from all providers as if they would come from one source (aggregate-and-anonymize, Figure 1B), using either a trusted third-party(TTP) or Secure Multi-party Computation (SMC) protocols to do computations. Generalization module performs 2-K-anonymity process. In generalization approach we use the identifiers data and Quasi Identifiers. Here the attribute age is Identifiers, and gender is Quasi Identifiers. The generalization data can be retrieved from an original data. The dataset data's are stored into two buckets. Bucketization module can be performs 2-diversity process. In generalization approach we use the Quasi Identifiers. Here the attribute work class is attribute. The bucketization data can be retrieved from an original data. The dataset data's are stored into two buckets. Multi-set generalization module performs 2-K-anonymity process. In multi-set generalization approach we use the identifiers data and Quasi Identifiers. Here the attribute age is Identifiers, and gender, work class are Quasi Identifiers. The multi-set generalization data can be retrieved from an original data. The dataset data's are stored into two buckets.

Overlapping Slicing Techniques

Overlapping Slicing Techniques partitions the data set both vertically and horizontally. Overlapping Slicing Techniques preserves better data utility than generalization and can be used for membership disclosure protection. Here we using the following sub modules a data recipient, e.g. P0, could be an attacker and attempts to infer additional information about the records using the published data (T∗) and some background knows - edge (BK) such as publicly available external data. Each data provider, such as P1 in Figure 1, can also use K-anonymity data T∗ and his own data (T1) to infer additional information about other records. Compared to the attack by the external recipient in the first attack scenario, each provider has additional data knowledge of their own records, which can help with the attack. This issue can be further worsened when multiple data providers collude with each other.

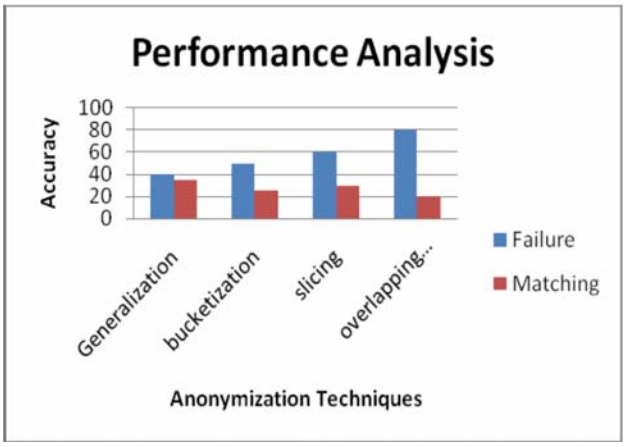
T_a^*

Provider	*Name*	Age	Zip	Disease
P_1	*Alice*	[20-30]	*****	Cancer
P_1	*Emily*	[20-30]	*****	Asthma
P_3	***Sara***	**[20-30]**	*****	**Epilepsy**
P_1	*Bob*	[31-35]	*****	Asthma
P_2	*John*	[31-35]	*****	Flu
P_4	*Olga*	[31-35]	*****	Cancer
P_4	*Frank*	[31-35]	*****	Asthma
P_2	*Dorothy*	[36-40]	*****	Cancer
P_2	*Mark*	[36-40]	*****	Flu
P_3	*Cecilia*	[36-40]	*****	Flu

Performance Analysis

Graph generation module can be used to find the classification accuracy between Original data, Generalization, Bucketization and Overlapping Slicing Techniques. Overlapping Slicing Techniques shows better accuracy than generalization. When the target attribute is the sensitive attribute, overlapping Slicing Techniques even performs better than bucketization. In this module Doctor can see all the patients details and will get the background knowledge(BK),by the chance he will see horizontally partitioned data of distributed data base of the group of hospitals and can see how many patients are affected without knowing of individual records of the patients and sensitive information about the individuals. In this module Admin acts as Trusted Third Party (TTP). He can see all individual records and their sensitive information among the overall hospital distributed data base. Anonymation can be

done by this people. He / She collected information from various hospitals and grouped into each other and make them as an anonymizes data.



Conclusion

In this paper, we considered a new type of potential at-tackers in collaborative data publishing – a coalition of data providers, called m-adversary. To prevent privacy disclosure by any m-adversary we showed that guaranteeing m-privacy is enough. We presented heuristic algorithms exploiting equivalence group monotonicity of privacy constraints and adaptive ordering techniques for efficiently checking m-privacy. We introduced also a provider-aware K-anonymity algorithm with adaptive m-privacy checking strategies to ensure high utility and m-privacy of K-anonymity data. Our experiments confirmed that our approach achieves better or comparable utility than existing algorithms while ensuring m-privacy efficiently. There are many remaining research questions. can provide strong and robust privacy protection to individuals in published or shared databases without sacrificing much utility of the data. Anonymity is very powerful technique for protecting privacy. This paper presents a new approach for privacy preservation called Slicing. Slicing is promising technique for handling high-dimensional data.

References

1. Tiancheng Li, Ninghui Li, Senior Member, IEEE, Jia Zhang, Member, IEEE, and Ian Molloy Slicing: A New Approach for Privacy Preserving Data Publishing Proc. IEEE Transactions on Knowledge and Data Engineering, 2012, 24(3).
2. 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; 23(2):161-174.
3. Ghinita G, Tao Y, Kalnis P. On the Anonymization of Sparse High Dimensional Data Proc. IEEE 24th Int'l Conf. DataEng. (ICDE), 2008, 715-724.
4. Martin DJ, Kifer D, Machanavajjhala A, Gehrke J, Halpern JY. Worst-Case Background Knowledge for Privacy-Preserving Data Publishing," Proc. IEEE 23rd Int'l Conf. DataEng. (ICDE), 2007, 126-135.
5. Samarati P. Protecting Respondent's Privacy in Micro dataRelease, IEEE Trans. Knowledge and Data Eng 2001; 13(6):1010-1027.
6. Inan A, Kantarcioglu M, Bertino E. Using Anonymized Data for Classification, Proc. IEEE 25th Int'l Conf. Data Eng. (ICDE), 2009, 429-440.
7. Golab L, Ozsu M. Issues in data stream management, ACM Sigmod Rec 2003; 32(2):5-14.
8. Babocock B, Babu S, Datar M, Motwani R, Widom J. Models and issues in data stream systems, in Proc., 21st ACM SIGMOD-SIGACT-SIGART Symp. Principles Database Syst, 2002, 1-16.
9. Nehme R, Rundensteiner E, Bertino E. A security punctuation framework forenforcing access control on streaming data, in Proc, IEEE 24th Int. Conf Data Eng, 2008, 406-415.
10. Carminati B, Ferrari E, Cao J, Tan K. A framework to enforce access control over data streams, ACM Trans. Inf. Syst. Security 2010; 13(3):28.
11. Samarani P. Protecting respondents' identities in microdata release, IEEE Trans. Knowl. Data Eng 2001; 13(6):1010-1027.
12. Machanavajjhala A, kifer D, Gehrke J, Venkitasubramaniam M. I-diversity: Privacy beyond k-anonymity," ACM Trans. Knowl. Discov. Data 2007; 1(1):3.
13. Cao J, Carminati B, Ferrari E, Tan K. Castle: Continuously anonymizing data streams,"IEEE Trans.Dependable Secure Comput 2011; 8(99):337-352.