Ravindra S. Wanjari, Prof. Devi Kalpna/ International Journal of E ngineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com
Vol. 3, Issue 4, Jul-Aug 2013, pp. 119-122

Improving the implementation of new approach data privacy preserving in data mining using slicing

Ravindra S. Wanjari[1], Prof. Devi Kalpna[2]
[1](Vivekanand Institute of Technology and Science , Karimnagar , Andhra Pradesh)
[2](Assistant Professor, Computer Science and Engineering Department Vivekanand Institute of Technology and Science , Karimnagar , Andhra Pradesh)

Abstract
Several anonymization 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 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. We show how slicing can be used for attribute disclosure protection and develop an efficient algorithm for computing the sliced data that obey the ℓ-diversity requirement. Our workload experiments confirm that slicing preserves better utility than generalization and is more effective than bucketization in workloads involving the sensitive attribute. Our experiments also demonstrate that slicing can be used to prevent membership disclosure.

Keywords:-Privacy preservation, data anonymization, data publishing, data security..

I. Introduction

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. The most popular ones are generalization for k-anonymity and bucketization [17] for 'ℓ-diversity [25].

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 Zipcode;
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.

It has been shown [1], [16], that generalization for k anonymity losses considerable amount of information, especially for high-dimensional data. This is due to the following three reasons. First, generalization for k-anonymity suffers from the curse of dimensionality. 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, forcing a great amount of generalization to satisfy k-anonymity even for relatively small k's. Second, 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. Third, because each attribute is generalized separately, correlations between different attributes are lost. In order to study attribute correlations on the generalized table, the data analyst has to assume that every possible combination of attribute values is equally possible. This is an inherent problem of generalization that prevents effective analysis of attribute correlations. While bucketization [26], [17] has better data utility

than generalization, it has several limitations. First, 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.

II. Proposed Method

In this paper, we present a novel technique called slicing for privacy-preserving data ublishing. Our contributions include the following.First, we introduce slicing as a new technique for privacy preserving data publishing. Slicing has several vantages when compared with generalization and bucketization. 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.

Second, we show that slicing can be effectively used for preventing attribute disclosure, based on the privacy requirement of ℓ -diversity. We introduce a notion called ℓ- diverse slicing, which ensures that the adversary cannot learn the sensitive value of any individual with a probability greater than 1/ℓ, We develop an efficient algorithm for computing the sliced table that satisfies ℓ-diversity. Our algorithm partitions attributes into columns, applies column generalization, and partitions tuples into buckets. Attributes that are highly correlated are in the same column; this preserves the correlations between such attributes. The associations between uncorrelated attributes are broken; this provides better privacy as the associations between such attributes are less frequent and potentially identifying. Fourth, we describe the intuition behind membership disclosure and explain how slicing prevents membership disclosure. A bucket of size k can potentially match k^C tuples where c is the number of columns. Because only k of the k^C tuples are actually in the original data, the existence of the other k^C - k tuples hides the membership information of tuples in the original data.

Finally, 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. Our experiments also show the limitations of bucketization in membership disclosure protection and slicing remedies these limitations. We also evaluated the performance of slicing in anonymizing the Netflix Prize data set.

III. Proposed techniques used

In the proposed work we have used slicing technique and compared it to generalization and bucketization

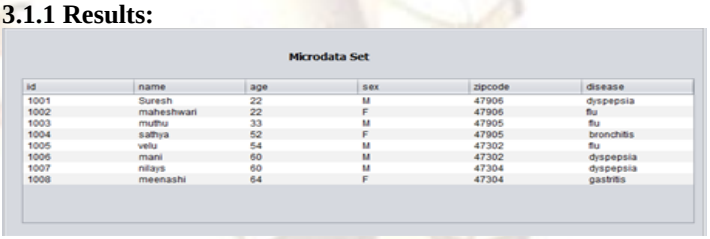
3.1 Slicing

Slicing first partitions attributes into columns. Each column contains a subset of attributes. This vertically partitions the table. For example, the sliced table in Table 6 contains two columns: the first column contains { Age; Sex} and the second column contains {Zipcode; Disease}. The sliced table shown in Table 5 contains four columns, where each column contains exactly one attribute. Slicing also partition tuples into buckets. Each bucket contains a subset of tuples. This horizontally partitions the table. For example, both sliced tables in Tables 5 and 6 contain two buckets, each containing four tuples.

Within each bucket, values in each column are randomly permutated to break the linking between different columns.

For example, in the first bucket of the sliced table shown in Table 6, the values {(22,M) , (22,F) (33,F) , (52,F)} are randomly permutated and the values{(47906,dyspepsia), (47906,flu), (47905, flu),(47905, bronchitis)} are randomly permutated so that the linking between the two columns within one bucket is hidden.

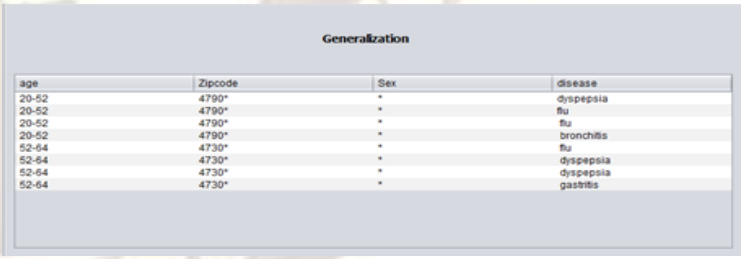
3.1.1 Results:



Microdata Set

id	name	age	sex	zipcode	disease
1001	Suresh	22	M	47906	dyspepsia
1002	maheshwari	22	F	47906	flu
1003	muthu	33	M	47905	flu
1004	sathya	52	F	47905	bronchitis
1005	velu	54	M	47302	flu
1006	mani	60	M	47302	dyspepsia
1007	nilays	60	M	47304	dyspepsia
1008	meenashi	64	F	47304	gastritis

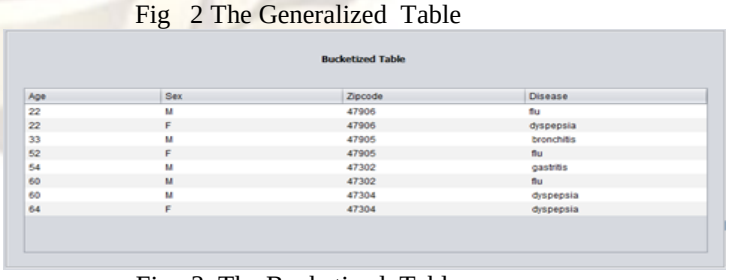
Fig 1 Original Microdata Table



Generalization

age	Zipcode	Sex	disease
20-52	4790*	*	dyspepsia
20-52	4790*	*	flu
20-52	4790*	*	flu
20-52	4790*	*	bronchitis
52-64	4730*	*	flu
52-64	4730*	*	dyspepsia
52-64	4730*	*	dyspepsia
52-64	4730*	*	gastritis

Fig 2 The Generalized Table



Bucketized Table

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

Fig 3 The Bucketized Table

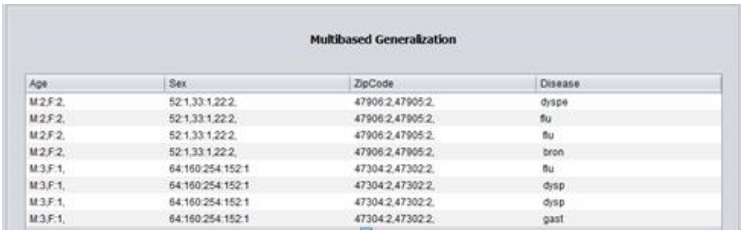


Fig 4 Multiset based generalization

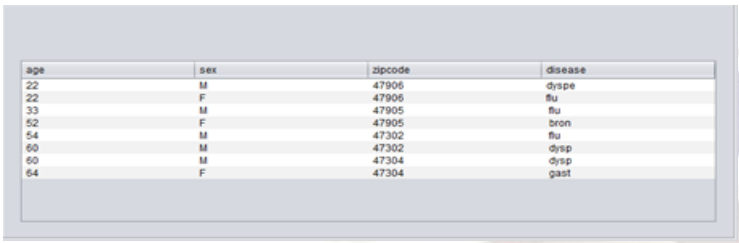


Fig 5 One attribute per column slicing

Fig 6 The sliced Table

IV. Comparitative Results

Two popular anonymization techniques are generalization and bucketization. Generalization replaces a value with a "less-specific but semantically consistent" value.

The main problems with generalization are: 1) it fails on high-dimensional data due to the curse of dimensionality and it causes too much information loss due to the uniform-distribution assumption .

Bucketization 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. The anonymized data consist of a set of buckets with permuted sensitive attribute values. In particular, bucketization has been used for anonymizing high-dimensional data [. However, their approach assumes a clear separation between QIs and SAs. In addition, because the exact values of all QIs are released, membership information is disclosed.

V. Conclusion

This paper presents a new approach called slicing to privacy preserving microdata publishing. Slicing overcomes the limitations of generalization and bucketization and preserves better utility while protecting against privacy threats.

We illustrate how to use slicing to prevent attribute disclosure and membership disclosure. Our experiments show that slicing preserves better data utility than generalization and is more effective than bucketization in workloads involving the sensitive attribute.

The general methodology proposed by this work is that before anonymizing the data, one can analyze the data characteristics and use these characteristics in data anonymization. The rationale is that one can design better data anonymization techniques when we know the data better. In [22], [24], we show that attribute correlations can be used for privacy attacks.

VI. Future Scope

While a number of anonymization echniques have been designed, it remains an open problem on how to use the anonymized data. In our experiments, we randomly generate the associations between c column values of a bucket. This may lose data utility. Another direction is to design data mining tasks using the anonymized data [14] computed by various anonymization techniques.

References

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

[2] *A. Blum, C. Dwork, F. McSherry, and K. Nissim, "Practical Privacy: The SULQ Framework," Proc. ACM Symp. Principles of Database Systems (PODS), pp. 128-138, 2005.*

[3] *J. Brickell and V. Shmatikov, "The Cost of Privacy: Destruction of Data-Mining Utility in Anonymized Data Publishing," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 70-78, 2008.*

[4] *B.-C. Chen, K. LeFevre, and R. Ramakrishnan, "Privacy Skyline: Privacy with Multidimensional Adversarial Knowledge," Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 770-781, 2007.*

[5] *Cramt'er, Mathematical Methods of Statistics. Princeton Univ. Press, 1948.*

[6] *I. Dinur and K. Nissim, "Revealing Information while Preserving Privacy," Proc. ACM Symp. Principles of Database Systems (PODS), pp. 202-210, 2003.*

[7] *C. Dwork, "Differential Privacy," Proc. Int'l Colloquium Automata, Languages and Programming (ICALP), pp. 1-12, 2006.*

[8] *C. Dwork, "Differential Privacy: A Survey of Results," Proc. Fifth Int'l Conf. Theory*

and Applications of Models of Computation (TAMC), pp. 1-19, 2008.

[9] *C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating Noise to Sensitivity in Private Data Analysis," Proc. Theory of Cryptography Conf. (TCC), pp. 265-284, 2006.*

[10] *J.H. Friedman, J.L. Bentley, and R.A. Finkel, "An Algorithm forFinding Best Matches in Logarithmic Expected Time," ACM Trans. Math. Software, vol. 3, no. 3, pp. 209-226, 1977.*

[11] *B.C.M. Fung, K. Wang, and P.S. Yu, "Top-Down Specialization for Information and Privacy Preservation," Proc. Int'l Conf. Data Eng. (ICDE), pp. 205-216, 2005.*

[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] *Y. He and J. Naughton, "Anonymization of Set-Valued Data via Top-Down, Local Generalization," Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 934-945, 2009.*

[14] *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.*

[15] *L. Kaufman and P. Rousueeuw, "Finding Groups in Data: An Introduction to Cluster Analysis," John Wiley & Sons, 1990.*

[16] *[16] 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.*

[17] *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.*

[18] *K. LeFevre, D. DeWitt, and R. Ramakrishnan, "Incognito: Efficient Full-Domain k-Anonymity," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), pp. 49-60, 2005.*

[19] *K. LeFevre, D. DeWitt, and R. Ramakrishnan, "Mondrian Multidimensional k-Anonymity," Proc. Int'l Conf. Data Eng. (ICDE), p. 25, 2006.*

[20] *K. LeFevre, D. DeWitt, and R. Ramakrishnan, "Workload-Aware Anonymization," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 277-286, 2006.*

[21] *N. Li, T. Li, and S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and '-Diversity," Proc. IEEE 23rd Int'l Conf. Data Eng. (ICDE), pp. 106-115, 2007.*

[22] *T. Li and N. Li, "Injector: Mining Background Knowledge for Data Anonymization," Proc. IEEE 24th Int'l Conf. Data Eng. (ICDE), pp. 446-455, 2008.*

[23] *T. Li and N. Li, "On the Tradeoff between Privacy and Utility in Data Publishing," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 517-526, 2009.*

[24] *T. Li, N. Li, and J. Zhang, "Modeling and Integrating Background Knowledge in Data Anonymization," Proc. IEEE 25th Int'l Conf. Data Eng. (ICDE), pp. 6-17, 2009.*

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

[26] *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.*

[27] *M.E. Nergiz, M. Atzori, 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.*

[28] *P. Samarati, "Protecting Respondent's Privacy in Microdata Release," IEEE Trans. Knowledge and Data Eng., vol. 13, no. 6, pp. 1010-1027, Nov./Dec. 2001.*

[29] *L. Sweeney, "Achieving k-Anonymity Privacy Protection Using Generalization and Suppression," Int'l J. Uncertainty Fuzziness and Knowledge-Based Systems, vol. 10, no. 6, pp. 571-588, 2002.*

[30] *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.*

[31] *M. Terrovitis, N. Mamoulis, and P. Kalnis, "Privacy-Preserving Anonymization of Set-Valued Data," Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 115-125, 2008. [32] R.C.-W. Wong, A.W.-C. Fu, K. Wang, and J. Pei, "Minimality Attack in Privacy Preserving Data Publishing," Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 543-554, 2007.*