Volume : 2 | Issue : 11 | Nov 2013 • ISSN No 2277 - 8160

Research Paper Engineering

REPLICATION OF ATTRIBUTES FOR BETTER UTILITY IN DATA PUBLISHING

K.J.Chakitha Mitra M Tech, CSE Department, JNTUACEA, Andhra Pradesh, India

ABSTRACT

Knowledge and Data Engineering is a prominent field which is also known as Data Mining. It involves in preserving the person specific and sensitive information like medical records [2], student records etc. It provides privacy and better data utility in maintaining huge data. The most important thing is to develop the methods and tools for publishing the data even in critical environment. The published data must contain privacy for every individual record in it and at the same time it must be useful in every aspect. Privacy preserving data publishing is the study of eliminating privacy threats and preserving the useful information. The work done here is maintaining the utility [3] of data in better way by replicating the attributes which are present in the original data regarding patient records.

KEYWORDS : Data anonymization, data publishing, utility, data mining, k-anonymity.

1. Introduction:

In recent years privacy preserving and data publishing of micro records has been studied extensively. The micro record refers to a single individual entity such as organization, person or a household. In order to handle this micro record several anonymization techniques have been proposed. But most important is given to generalization and bucketization. Where generalization is used for k-anonymity [8] and bucketization is used for l-diversity.

These anonymization techniques [1] are proposed to protect ones private information while publishing.

In order to preserve the privacy the attributes are classified into three categories:

- Identifiers (ID): This uniquely identifies an individual. E.g.: Name.
- Quasi-identifiers (QI): These are already known and when together considered can identify an individual potentially. E.g.: Zip code, gender, date of birth etc.
- Sensitive Attributes (SA): These are the fields which should be considered as sensitive as to avoid being associated to a single specific person. E.g.: Disease, salary.

General anonymization removes the identifiers from the data and then the tuples are partitioned into buckets up to this procedure is common from here it differs for each technique. Generalization involves recoding a value which is semantically consistent value with a less specific one.

2. Related Work:

In the related work the tuple partition and the diversity check plays major role. It satisfies the k-anonymity and l-diversity. It maintains both privacy and utility than generalization and bucketization.

2.1 K-Anonymity:

Let T(A1,…An) be a table and the quasi-identifier which is associated with it be QIT. k-anonymity is satisfied when each sequence of values in T[QIT] in T is appears with at least k occurrences in T [QIT].Usually in bucketization SA are separated from QI by performing random permutations on the SA value in each bucket. So, anonymized data contains a set of buckets including the permuted SA values. This requires the separation between the SA and QI attributes for that it breaks the correlation between them.

The main reason for not considering the l-diversity here is that it is insufficient for providing the attribute disclosure. Even there are two attacks on l-diversity:

Skewness Attack: where l-diversity doesn't prevent the attribute disclosure when the overall distribution is skewed.

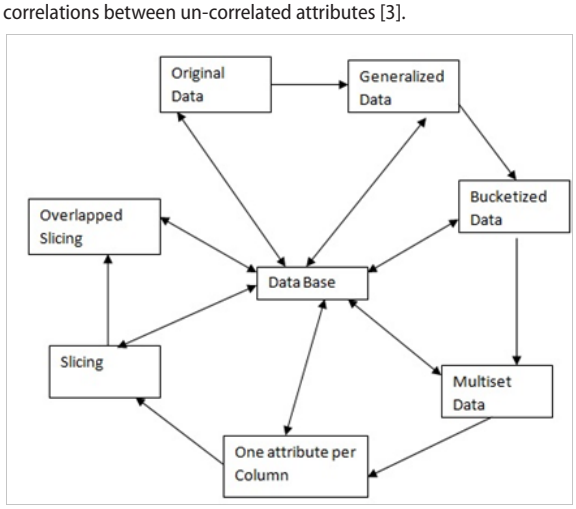
Similarity Attack: In equivalence class when the sensitive attribute values are distinct and semantically equal the adversary will learn the information.

In order to get better privacy preserving while data publishing we go for SLICING [1]. Slicing generally partitions the data set horizontally and vertically. Vertical partitioning is grouping the individual attributes into columns based on the attribute correlations. Horizontal slicing is done by grouping the tuples into buckets. The main idea behind the slicing is that to break the association across columns, but to preserve the association with each column. By which the dimensionality of the data is reduced and preserves better utility than the bucketization and generalization [1]. The utility is best preserved by slicing because it groups correlated attributes together, and the correlation between such attributes is preserved.

3. Proposed System:

By observing the three Anonymization techniques we can observe that Slicing can be better utilized for maintaining privacy while publishing the data compared to generalization and bucketization.

For privacy preserving and data publishing several anonymization techniques like generalization, bucketizations were designed but they have several limitations. So, slicing is introduced [1]. The main idea of slicing is to accomplish a better trade-off between privacy and utility by preserving correlations between correlated attributes and breaking correlations between un-correlated attributes [3].



In slicing process, the data is divided into rows & columns to improve the current state of the data. Columns are created by rouping them based on the attribute correlations. Rows are created by grouping tuples into buckets. Partitioning the tuples and checking its diversity plays a major role here.

Slicing protects privacy and also preserves utility. But while publishing

some data the privacy and utility defers in its ratio. That is, when there is a gain in privacy then utility will be lost and when utility is gained then privacy will be lost. Both the privacy and utility are vice-versa [3]. So, this paper concentrates only on maintaining utility while publishing data.

Fig (a): Architecture of the Proposed System.
Above architecture shows that the overlapped slicing, slicing, one attribute per column, multiset data, bucketized and generalized data all depends on the original data. The overlapped slicing is the extension of slicing. And all these are connected to the database. When slicing is compared to all other previous anonymization techniques which were observed in the fig (a), it is better in maintaining privacy and utility. In overlapping slicing the maintenance of utility is only concentrated.

This paper shows utility of the patients records who have general diseases [2].The region where the attribute must be replicated is first checked and it must be accepted as trusted one [5]. Then the attribute is replicated.

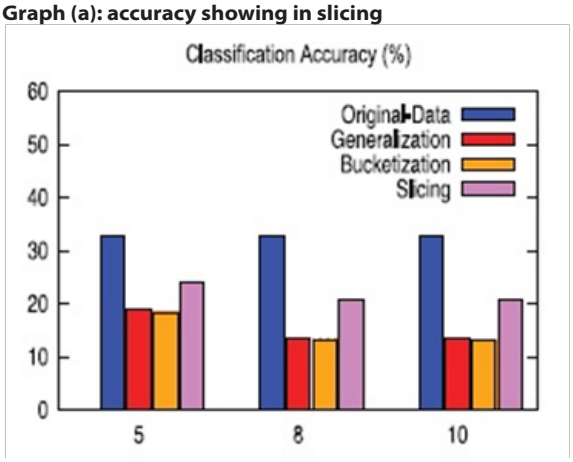
4. Experimental Results:

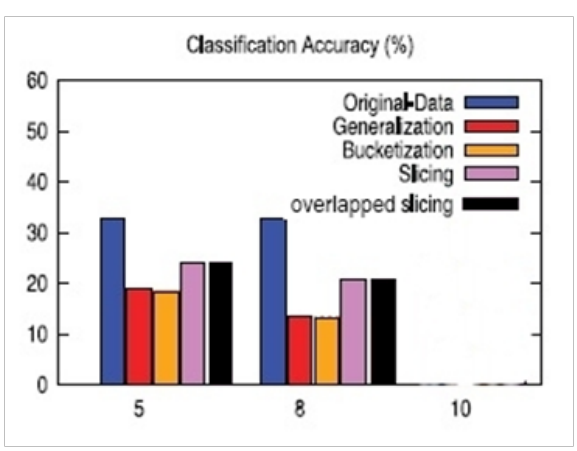
Some patient's records were used to test the proposed methodology on that data and compared the results with the results of the work done by [1].

Our results shows increase in the utility compared to previous works.

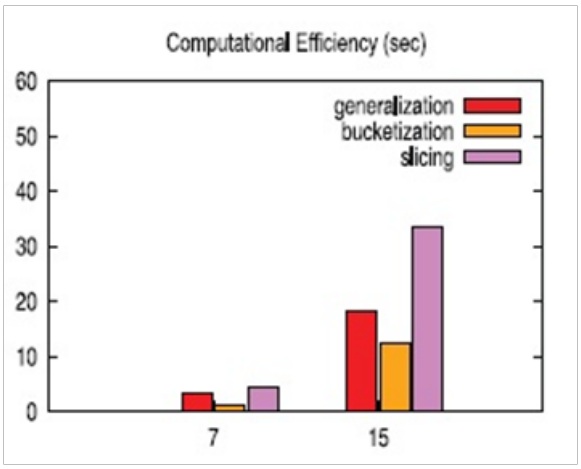
By calculating the accuracy of the data we can observe that the accuracy is maintained in the replicated data as same as in sliced data.

Graph (a): accuracy showing in slicing

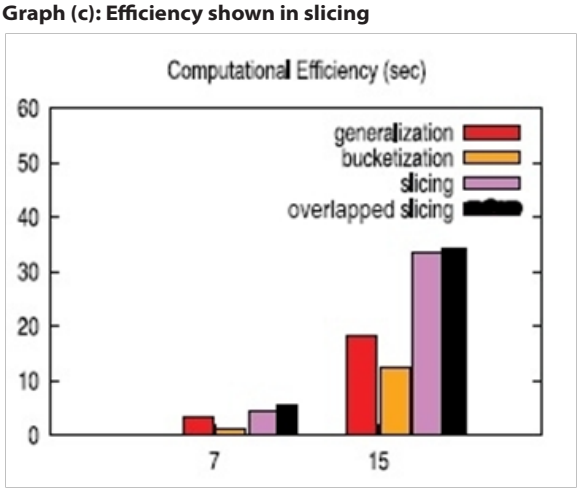




Graph (b): accuracy showing in overlapped slicing
The following graphs show the difference between the data efficiency in both the sliced data and the replicated data.



Graph (c): Efficiency shown in slicing



Graph (d): Efficiency shown in overlapped slicing
In the above graph we can observe the black color bar which shows the improvement in the efficiency of the replicated data than the sliced data.

5. Conclusion:

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. The experiments done in [1] show that preservation of data utility in slicing is better than in generalization and bucketization. The work done here is maintaining the utility of data in better way by replicating the attributes which are present in the original data regarding patient records. The same has proved by conducting the experiments on huge amount of patient records and have been successful throughout the testing. The results are shown by comparing with the previous methods done in [1]. Further study concentrates on improving both the privacy as well utility at the same time using combination of slicing and other anonymization techniques.

REFERENCES

[1]. Tiancheng Li, Ninghui Li, Senior Member, IEEE, Jain Zhang, Member, IEEE, and Ian Molloy,"Slicing: A New Approach for Privacy Preserving Data Publishing", IEEE Transactions on Knowledge and Data Engineering, Vol. 24, No. 3, March 2012. | [2]. "Privacy Preserving Medical Data Publishing", James Johnson Gardner,M.S. Computer Science, Emory University, Atlanta, 2007. | [3]. 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. | [4]. D. Kifer and J. Gehrke, "Injecting Utility into Anonymized DataSets," Proc. ACM SIGMOD Int'l Conf. Management of Data(SIGMOD), pp. 217-228, 2006. | | | [5]. "A review of trust region algorithms for optimization", Ya-xiang Yuan, State Key Laboratory of Scientific and Engineering Computing,Institute of Computational Mathematics and Scientific/Engineering Computing. | | [6]. "Limiting Attribute Disclosure in Randomization Based Microdata Release", Ling Guo, Xiaowei Ying, and Xintao Wu*University of North Carolina at Charlotte, NC 28223, USA. | | [7]. C. Aggarwal, "On k-Anonymity and the Curse of Dimensionality,"Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 901-909, 2005. | | [8]. S.Kiruthika, M.Mohamed Raseen M.S, "Suppression Slicing-using l-diversity", year of publication 2013. | | [9]. V. Nivedhitha and S. Kiruthika, 2012 Privacy preservation using l-diversity, proc. Int'l conf. On computational intelligence and information technology, pp-DMB065-71. | | [10]. R. C. -W. Wong, J. Li, A. W. -C. Fu, and K. Wang, 2006 (?, k)-Anonymity: "An Enhanced k-Anonymity Model for Privacy Preserving Data Publishing", Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 754-759. | | [11]. Benjamin C.M. Fung, Ke Wang, Rui Chen and Philip S. Yu, "Privacy-Preserving Data Publishing: A Survey of Recent Developments" ACM Computing Surveys, Vol. 42, No. 4, Article 14, Publication date: June 2010. |