Privacy preserving data publishing using slicing with marginal publication

A.LOGESWARI
Post Graduate Student - CSE
Dr.Mahalingam College of Engineering and Technology,Pollachi.
Email: logeswariathai@gmail.com

Mr.K.THIRUKUMAR M.E
Assistant Professor (SG) - CSE
Dr.Mahalingam College of Engineering and Technology,Pollachi.
Email:thirukumar@drmcet.ac.in

***Abstract*— Privacy preservation has become a major issue in many data analysis applications. Protecting the individuals privacy is a vital activity in microdata publishing. Anonymization is a technique to reduce the re-identification of person specific information. When a data set is released to data recipient for data analysis by data publisher, some privacy-preserving techniques are often required. Therefore, data publisher releases individual's information to reduce the possibility of identifying sensitive information about individuals. In order to protect sensitive information, the simplest solution is not to disclose the information. The data publisher can transform the data in such a way that the modified data must guarantee privacy and also retains sufficient utility before it is released to data recipient. In the existing system, a novel anonymization technique for privacy preserving data publishing, Slicing is implemented. This technique preserves better utility and privacy by using partitioning strategies. But it provides less utility correlations. In the proposed system, overlapping slicing is implemented using one of the soft clustering techniques called fuzzy clustering, which provides better utility correlations. Fuzzy clustering partitions the data which may overlap in more than one cluster. It achieves overlapping clustering using membership function. It is expected that fuzzy clustering algorithm gives better performance in terms of both cardinality and dimensionality.**

***Index Terms*— Privacy preservation, Anonymization, Fuzzy Algorithm.**

I. INTRODUCTION

Anonymization is the process of removing or modifying the identifying variables contained in the dataset. Identifying those variables contains the characteristics of an individual. The issue is how to publish the data in such a way that the privacy of individuals can be preserve. Various anonymization methods can be applied to preserve the sensitive information, they are: generalization, suppression, randomly swapping some attributes in the original data records, permutations or perturbative masking. Although these anonymization methods are used to increase protection, minimizing the disclosure risk, but they also decrease the quality of the data (i.e.) it's utility. These techniques include methods such as randomization, k-anonymity, and l-diversity.

The goal of Privacy Preserving Data Publishing is to transform the table, such that individuals may not be linked to specific tuples with high certainty. At the same time, the published data should be useful to the data publisher and anonymize the data such that a certain degree of privacy is preserved while data utility is maximized.

Anonymity is an important concept for privacy and it can embed privacy protection in data itself. for example, no one can tell to whom a data record is related (referred to as identity privacy) or no one can learn about a particular property of individuals (referred to as attribute privacy) from observing an anonymous dataset. The traditional approach of releasing the data tables without breaching the privacy of individuals in the table is to de-identify records by removing the identifying fields such as name, address, and social security number. However, joining this de-identified table with a publicly available database (like the voters database) on attributes like race, age, and zip code (usually called quasi-identifier) can be used to identify the individuals.

A. ANONYMITY MODELS

K-Anonymity :

In this technique each record within an anonymized table must be identical with at least k-1 other record within the dataset, with respect to a set of QI attributes [9]. In particular, a table is k-anonymous if the QI attributes values of each record are identical to those of at least k-1 other records. To achieve the k-anonymity requirement, generalization or suppression could be used. It is very difficult for a database owner to determine which of the attributes are available or not available in external tables.

L-Diversity:

This technique was proposed to solve the homogeneity attack of k-anonymity technique that emphasizes not only on saving the minimum size of k group but also considers saving the variety of the sensitive attributes of each group [10].It treats all values of a given attribute in a same way regardless of its distribution in the data and it is insufficient to prevent attribute disclosure.

T-Closeness:

Each group of records that have identical QI attribute values is called an equivalence class [12]. An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t. A table is said to have t-closeness if all equivalence classes have t-closeness. T-closeness uses the Earth Mover Distance (EMD) function to measure the sensitive attribute frequency distribution of publicly available data with the distribution of quasi group and requires the closeness to be within it. It requires the distribution of sensitive values to be the same in all quasi identifier groups.

B.ANONYMIZATION APPROACHES

Generalization and Suppression:

Generalization technique substitutes the values of a given attribute with more general values whereas suppression replaces some values with a special value (e.g., "*" or "Any") also indicating that the replaced values are not disclosed.

Bucketization and Randomization:

Bucketization partitions the original dataset into non-overlapping groups (or buckets) and for each group, its provides a clear separation between quasi- attributes and sensitive attributes and the bucketized data consists of a set of buckets with permuted sensitive attribute values.By randomization ,anonymized data could be created by randomly perturbing the attribute values.

Privacy models typically include three types of disclosure: identity, attribute, and membership disclosure. Preventing adversaries from learning whether one's record is included in the published data set is called membership disclosure. Identity disclosure occurs if an individual is linked to a particular record in the released table. Attribute disclosure occurs if new information about some individuals is revealed

II. RELATED WORK

T. Li and N. Li analyzed the fundamental characteristics of privacy and utility, and show that it is inappropriate to directly compare privacy with utility [1]. Also, it describes the tradeoff between privacy and utility in micro data publishing.

The direct comparison methodology evaluates privacy loss and utility gain. Therefore privacy loss is measured as the adversary's accuracy improvement in guessing the sensitive attribute value of an individual and utility gain is measured as the researcher's accuracy improvement in building a classification model for the sensitive attribute.

Some of the fundamental characteristics are:

(1)Privacy concerns information about specific individuals and utility contributes to aggregate information about large populations.

(2)Privacy should be enforced for each individual and utility accumulates all useful knowledge.

The idea of trade-off metrics is to consider both the privacy and information requirements at every anonymization operation and to determine an optimal trade-off between the two requirements.

Thomas and Wangmo proposed modification to the fuzzy c-means algorithm to overcome the limitations of it in calculating the new cluster centers and in finding the membership values with natural data [7]. Clustering algorithms maps a new data item into one of several known clusters. Membership of a data item in a cluster can be determined by measuring the distance from each cluster center to the data point. The most popular fuzzy clustering technique is fuzzy C-means algorithm.

The algorithm has difficulty in handling outlier points and it has inability to calculate the membership value if the distance of a data point is zero.

Some of the modifications are:

i. C-means with Modified Distance Function

Klawonn and Keller have proposed a modified C-means algorithm with new distance function which is based on dot product instead of the conventional Euclidean distance [11].

ii. Modified C-means for MRI Segmentation

Jiang and Yang have formulated modified C-means by modifying the objective function of the standard fuzzy C-means (FCM) method to compensate for intensity inhomogeneities.

iii. Adaptive Fuzzy Clustering

Krisnapuram and Keller proposed a modified version of the c-means clustering [11]. The adaptive fuzzy clustering algorithm is efficient in handling data with outlier points. In comparison with C-means algorithm, it gives only very low membership for outlier points.

A modified version of fuzzy c-means algorithm is presented and the new algorithm is applied on a natural data set and its performance is compared with that of classical fuzzy C-means algorithm and found that the new method gives better performance in defining cluster centers.

Chen and Hu proposed a new overlapping cluster algorithm which differs from traditional clustering algorithms [5]. The new clustering algorithm achieved overlapping, so clusters are allowed to overlap with one another.

The traditional partitioning cluster methods use the shortest distances to determine which cluster an object should belong to and an object is assigned to the nearest cluster.

The author has developed overlapping partitioning cluster algorithm (the OPC algorithm) and it adopts the same framework as that of K-Medoids algorithm.

The algorithm has two input parameters k and s. The number k specifies the number of clusters and s is the threshold for similarity. Since the algorithm is a heuristic method, the process should be repeated until a satisfactory result is obtained.

The OPC algorithm contains three parts,

- The first part is the preprocessing work.
- The second part is a smart method, which randomly selects the initial k cluster center-objects.
- The last part adjusts the clusters by iteratively changing cluster center-objects.

A new overlapping partitioning cluster algorithm is developed by modifying the traditional K-Medoids algorithm [5]. Finally, a simulation is designed to evaluate the effectiveness and the efficiency of the OPC algorithm. The results indicate that the algorithm is efficient and can generate satisfactory clustering results.

III. EXISTING SYSTEM

The basic idea of slicing is to break the association across columns and to preserve the association within each column [6]. Slicing groups highly correlated attributes together and preserves the correlations between such attributes and breaks the associations between uncorrelated attributes.

Slicing partitions the dataset in two ways,

- Vertical partitioning
- Horizontal partitioning

The main goal of slicing is to preserve utility and privacy. Therefore, it preserves utility by grouping highly correlated attributes together and provides privacy by breaks the associations between uncorrelated attributes.

Compared to generalization, Slicing is better approach in which, it groups correlated attributes together in one column and preserves their correlation.

Compared to bucketization, Slicing can be used without a clear separation between QI and sensitive attributes and also it can be used to prevent membership disclosure.

A.ATTRIBUTE PARTITIONING (VERTICAL PARTITIONING)

In this phase, it involves two steps:
(i)Compute the correlations between pairs of attributes
(ii)Cluster attributes based on their correlations

Correlation Measure:

Mean-square contingency coefficient is a chi-square measure of correlation between two categorical attributes.

$$\varphi^2(A_1,A_2)=\frac{1}{\min\{d_1,d_2\}-1}\sum_{i=1}^{d_1}\sum_{j=1}^{d_2}\frac{(f_{ij}-f_{i.}f_{.j})^2}{f_{i.}f_{.j}}$$

Equation is used to calculate Chi-square correlation measure
Where,

A_1, A_2 → two categorical attributes
d_1,d_2 →domain size of attributes
$[\{v_{11},v_{12},\ldots\ldots\ldots v_{1d1}\},\{v_{21},v_{22},\ldots\ldots\ldots v_{2d2}\}]$
$f_{i.}$, $f_{.j}$ → fraction of occurances of v_{1i} and v_{2j} in the data
f_{ij} → fraction of cooccurances of v_{1i} and v_{2j} in the data

Attribute Clustering:

K-medoids algorithm (PAM) is used for attribute clustering. The distance between two attributes in the clustering space is defined as $d(A_1, A2) = 1-\Phi^2(A_1, A_2)$. Two attributes that are strongly correlated will have a smaller distance between the corresponding data points in our clustering space.
PAM steps are:

i. Initially, selecting k data points as the initial medoids.
ii. In each subsequent step, PAM chooses one medoid point and one non-medoid point and swaps them as long as the cost of clustering decreases.
iii. Cost of clustering - sum of the distance from each data point in the cluster to the medoid point of the cluster.

Result of this algorithm is shown in fig 1.

B. COLUMN GENERALIZATION

Column generalization may be required for identity/membership disclosure protection. If a column value is unique in a clumn (i.e., the column value appears only once in the column), a tuple with this unique column value can only have one matching bucket. Column generalization ensures that one column satisfies the k-anonymity requirement

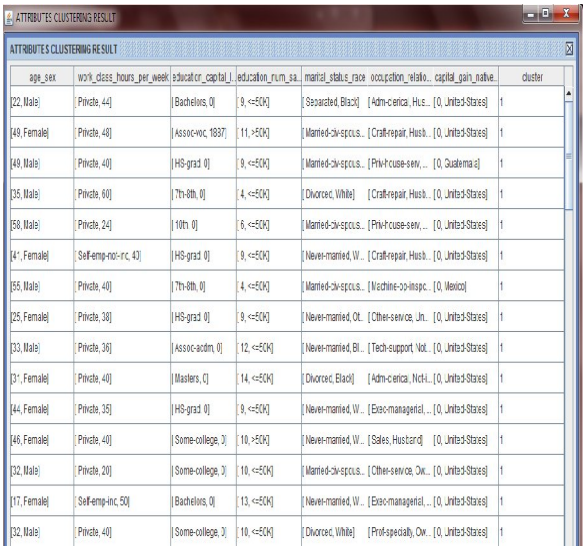

Fig 1. Attribute clustering

C. HORIZONTAL PARTITIONING (TUPLE PARTITIONING)

In the phase, tuples are partitioned into buckets using modified mondrain algorithm as in fig 2.

Tuple partitioning algorithm steps are:

The algorithm maintains two data structures: a queue of buckets Q and a set of sliced buckets SB.Initially, Q contains only one bucket which includes all tuples and SB is empty. In each iteration, the algorithm removes a bucket from Q and splits the bucket into two buckets.

If the sliced table after the split satisfies l-diversity, then the algorithm puts the two buckets, at the end of the queue Q. Otherwise, bucket cannot be splited anymore and the algorithm puts the bucket into SB. When Q becomes empty, sliced table have been computed. The set of sliced buckets is denoted as SB. When Q becomes empty, sliced table have been computed. The set of sliced buckets is denoted as SB.

The main part of the tuple-partition algorithm is to check whether a sliced table satisfies l-diversity. For each tuple t, the algorithm maintains a list of statistics about t's matching buckets. Each element in the list contains statistics about one matching bucket B: the matching probability and the distribution of candidate sensitive values.

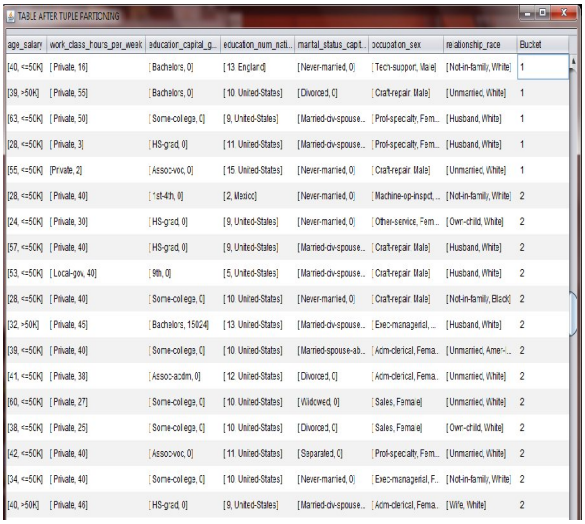

Fig 2. Tuple partitioning

IV. PROPOSED SYSTEM

Marginal publication can be viewed as a special case of slicing which does not have horizontal partitioning. Therefore, correlations among attributes in different columns are lost in marginal publication. It is similar to overlapping vertical partitioning and it also termed as overlapping slicing, which duplicates an attribute in more than one columns. This releases more attribute correlations.

Clustering is an unsupervised learning of unlabeled data. As the training is unsupervised in clustering algorithms, these can be safely used on a data set without much knowledge of it. Easy tackling of noisy data and outliers and the ability to deal with the data having various types of variables are the two important benefits of clustering. Several fuzzy clustering algorithms had been proposed by various researchers. Those algorithms include fuzzy ISODATA, fuzzy C-means, fuzzy K-nearest neighborhood algorithm, potential-based clustering, etc.

Fuzzy C-means (FCM) algorithm, one of the most popular fuzzy clustering techniques, was originally proposed by Dunn and had been modified by Bezdek.FCM is able to determine, and in turn, iteratively update the membership values of a data point with the pre-defined number of clusters. Thus, a data point can be the member of all clusters with the corresponding membership values.

Fuzzy C-means algorithm steps are:

Step 1: Initialize $U= [u_{ij}]$ matrix , $U^{(0)}$

Step 2: At k-step : calculate the centers vectors $C^{(k)}=[c_i]$ with $U^{(k)}$

$$\mathbf{c}_i = \frac{\sum_{j=1}^{n} u_{ij}^{m} \mathbf{x}_j}{\sum_{j=1}^{n} u_{ij}^{m}}$$

Step 3: Update $U^{(k)}$, $U^{(k+1)}$

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d_{ij}}{d_{kj}} \right)^{2/(m-1)}}$$

Where, $d_{ij} = \| x_i - x_j \|$

Step 4: If $\| U^{(k+1)} - U^{(k)} \| < \varepsilon$,then STOP ; otherwise return to step 2

By using this algorithm, overlapping is introduced between attributes. As a result, compared to existing system, better utility is achieved.

V. PERFORMANCE EVALUATION

The performance evaluation of the proposed system is measured in terms of utility and privacy. Utility refers to usefulness of the anonymized data which is released to the public. Privacy refers to how much security is imparted among released data. By using soft clustering techniques, our proposed system will be improved using these performance measures.

VI. CONCLUSION

Privacy preserving data publishing using slicing works well for high-dimensional data. It partitions the input dataset based on correlation coefficient followed by hard clustering of attributes .Furthermore, tuples are partitioned using l-diversity check algorithm. The overlapping slicing using soft clustering technique is expected to produce more utilitycorrelations among attributes.

VII. ACKNOWLEDGEMENT

I would like to express my gratitude to Mr.K.Thirukumar M.E., Assistant Professor (SG), Department of CSE, Dr.Mahalingam College of Engineering and Technology for his useful comments, remarks and engagement through the learning process of this project. Furthermore I would like to thank Ms.G.Anupriya M.E., Assistant Professor (SG), Department of CSE, Dr.Mahalingam College of Engineering and Technology for introducing me to the topic as well as for the support. Also, I would like to thank my family members, who have supported me throughout entire process, both by keeping me harmonious and helping me. I will be grateful forever for their love.

REFERENCES

[1] Li T, Li N, (2009), "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.

[2] Koudas N, Srivastava D ,Yu T, and Zhang Q, (2007), "Aggregate Query Answering on Anonymized Tables", Proc. IEEE 23rd Int'l Conf. Data Eng. (ICDE), pp. 116-125.

[3] Wong W K, Mamoulis N, Cheung D, (2006), "Non-homogeneous generalization in privacy preserving data publishing", SIGMOD Conference, pp. 747-758.

[4] Binu Thomas, Raju G, Sonam Wangmo, (2009), "A Modified Fuzzy C-Means Algorithm for Natural Data Exploration", World Academy of Science, Engineering and Technology, pp. 478-481.

[5] Chen Y L, Hu H L, (2006), "An Overlapping Cluster Algorithm To Provide Non-Exhaustive Clustering" , European Journal of Operational Research 173, pp.762–780.

[6] Tiancheng Li, Ninghui Li, Jian Zhang, and Molloy, (2012), "Slicing: A New Approach for Privacy Preserving Data Publishing", IEEE Transactions on Knowledge and Data Engineering, vol.24, no.3, pp. 561-574.

[7] Cox E, (2005), "Fuzzy Modeling and Genetic Algorithms for Data Mining and Exploration", Morgan Kaufmann Publishers of Elsevier.

[8] LeFevre K, DeWitt D and Ramakrishnan R, (2006), "Mondrian multidimensional K-anonymity ", International Conference on Data Engineering (ICDE), pp. 1-12.

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

[10] Machanavajjhala A, Gehrke J, Kifer D and Venkitasubramaniam M, (2006),"L-diversity: Privacy beyond k-anonymity", International Conference on Data Engineering (ICDE), pp. 1-12.

[11] Tao C W, (2002), "Unsupervised fuzzy clustering with multi-center clusters", Fuzzy Sets and Systems 128 (3), pp. 305–322.

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