Cluster Based Attribute Slicing: A New Approach for Privacy Preservation

Vanita Babanne[1], Neha N. Jamdar[2]

[1]M. E. Computer Engineering Department, RMDSSOE, Pune, India

[2]Profecessor, Computer Engineering Department, RMDSSOE, Pune, India

Abstract: ***Various anonymization techniques have been proposed for publishing a microdata. Anonymization techniques hide the sensitive data from the attackers. Examples of these techniques are slicing, bucketization and generalization. Generalization hides sensitive data but it loses lot of data. Bucketization make the difficult to detect sensitive attribute by randomizing sensitive attribute but it does not prevent relation between them. Slicing is better technique amongst all remaining technique because it cannot lose information and it is used for maintain the relation between attributes. In case of slicing, equal-width discretization is used to convert continuous attribute into categorical attribute. Equal-width discretization has very time consuming technique. To solve this problem we propose cluster based attribute slicing algorithm. Proposed technique does not take more to sort a data.***

Keywords: Privacy Preservation, Data Mining, Slicing Algorithm, Cluster Based Attribute Slicing Algorithm

1. Introduction

Data mining is a technique of finding or collecting useful and meaningful information from huge amount of database. Data collection is task in which data holder collects the data from data owners. Data publishing is the task in which releasing data for any user to use publicly which is collected by data holders. Privacy preservation is process in which private information is protected from attackers. Microdata is records which gives the information about on individual entity. Attributes are uniquely recognizing an entity known as identifier (ID). Attributes taken together, they recognize an entity known as Quesi-identifiers (QI). Attributes are unknown to attackers known as Sensitive attribute (SA). Table shows the microdata table. Generalization [1, 3, 5, 8] and bucketization [2, 4, 6, 7] are two anonymization techniques.

Table 1: Microdata Table

ID	QI			SA
Name	Age	Sex	Zipcode	Salary
Ajay	25	F	416306	35000
Vijay	36	F	416305	35000
Raj	52	F	416305	20000
Rani	54	M	416002	35000
Rahul	68	M	416002	25000

2. Slicing Algorithm

Slicing algorithm [10] divides attributes both horizontally and vertically. First vertical partition is done after that horizontal partition is performed. Slicing algorithm [9] first partitions attributes into columns that means vertical partition. Each column contains a subset of attributes. Slicing also partition tuples into buckets that means horizontal partition. Each bucket contains a subset of tuples. Slicing algorithm contains two phases: Attribute partition, and tuple partition.

1) Attribute Partitioning

In this phase, highly correlated attributes are grouped together into one column. Attribute partitioning also contains three steps:

A. Equal-width partitioning

Attributes are two types' categorical and continuous attributes. So, in this step continuous attributes are converted into categorical attributes. In equal –width, first divides the range into N intervals of equal size: uniform grid if A and B are the lowest and highest values of the attribute, the width of intervals will be: W = (B-A)/N.

B. Measures of Correlation

In this step, we calculate the relation between two attributes. To calculate relation between attributes mean-square contingency coefficient formula is used.

Let two attributes A_1 and A_2 with domains $\{V_{11}, V_{12}, \ldots\ldots V_{1n1}\}$ and $\{V_{21}, V_{22}, \ldots\ldots V_{2n2}\}$ respectively. There domain sizes are n_1 and n_2. The mean-square contingency coefficient between A_1 and A_2 is defined as

$$\varphi^2(A_1, A_2) = \frac{1}{\min\{n_1, n_2\}-1} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \frac{(f_{ij}-f_{i}f_{j})^2}{f_{i}f_{j}} \quad (1)$$

Here f_{ij} is the fraction of co-occurrences of V_{1i} and V_{2j} in the data. f_i and f_j are the fraction of occurrences of V_{1i} and V_{2j} in the data, respectively. It can shown that $0 \leq \varphi^2(A_1, A_2) \leq 1$.

C. Attribute Clustering

In this step, *k*-medoid clustering algorithm is used to partition attributes into columns and is as follows:

- Initialize: select randomly any *k* of the *n* data points as the medoids.

- Associate each data point to the closest medoid. ("Closest" here is defined using Euclidean distance.)
- For each medoid *m*
- For each non-medoid data point *p*
- Swap *m* and p and compute the total cost of the configuration
- Choose the configuration with the minimum value cost.
- Repeat 2-4 steps until there is no change in the medoid.

2) Tuple Partitioning

In this step, tuples are divided into buckets. Following is the algorithm for tuple-partitioning.

Algorithm tuple-partition (T, l)
a) P ={T}; SB=∅;
b) while P is not empty.

- Remove first bucket B from P;P=P-{B}.
- Split B into two buckets B_1 and B_2
- if diversity-check(T,P Ụ {B_1 , B_2} Ụ SB, l)
- P=P Ụ {B_1 ,B_2}.
- else SB=SB Ụ {B}.
- return SB.

In above algorithm, P is a queue of buckets and SB is a sliced buckets. Also in this algorithm l-diversity is check.

3. Cluster Based Attribute Slicing Algorithm

In existing system, equal-width discretization is used so it cannot handle skew data properly. So to solve this problem we proposed a new algorithm. In proposed method, we use the cluster based attribute algorithm for convert the continuous attribute into categorical attribute. The algorithm shows the follows:

Input: Vector of real valued data a = (a_1,a_2 ...., a_n) and the number of clusters to be determined k.
Goal: To find a partition of the data in k distinct clusters.
Output: The set of cut points t_0,t_1,...,t_k with t_0<t_1<...<t_k that defines the discretization of the adom(A).

Algorithm

- Compute a_{max} = max{a_1,a_2,...,a_n} and a_{min} = min{a_1,a_2,...,a_n}.
- Choose the centers as the first k distinct values of the attribute A.
- Arrange them in increasing order i. e. such that C[1] < C[2] < … < C[k].
- Define boundary points b_0 = a_{min} , b_j =(C[j] +C[j+1])/2 for j=1 to k-1, b_k = a_{max}
- Find the closest cluster to a_i .
- Recompute the centers of the clusters as the average of the values in each cluster.
- Find the closest cluster to a_i from the possible clusters {j-1, j, j+1}.
- Determination of the cut points
    - t_0 = a_{min}
    - for i = 1 to k-1
    - do
    - t_i = (C[i] +C[i+1])/2
    - endfor
    - t_k = a_{max}

After this next step is apply the formula of measures of correlation and attribute clustering. And finally apply the attribute partition algorithm same in slicing algorithm.

4. Dataset

For our proposed system we use the Adult dataset which is available UC Irvine machine learning repository. It conations total 11858 tuples and 15 attributes. Table 2shows the attributes of adult dataset.

Table 2: Adult dataset attribute

Sr. No.	Attribute	Types
1	Age	Continuous
2	Work-class	Categorical
3	Final-weight	Continuous
4	Sex	Categorical
5	Capital-gain	Continuous
6	Capital-loss	Continuous
7	Hours/week	Continuous
8	Country	Categorical
9	Salary	Categorical
10	Education	Categorical
11	Education-Num	Continuous
12	Marital-Status	Categorical
13	Occupation	Categorical
14	Relationship	Categorical
15	Race	Categorical

5. Results

Figure 1 and figure 2 shows the Comparison between various anonymization techniques and Comparison between slicing and cluster based attribute slicing algorithm respectively.

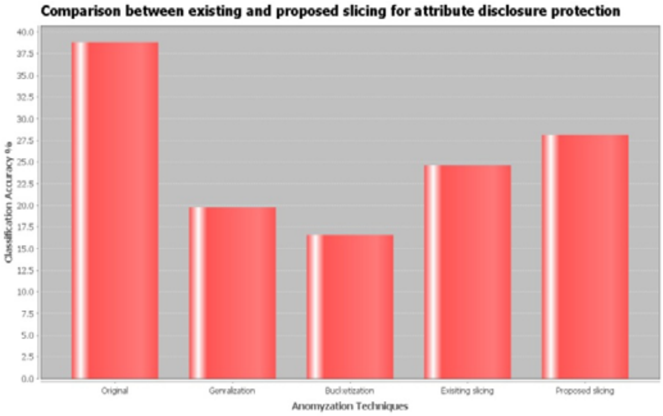

Figure 1: Comparison between various anonymization techniques

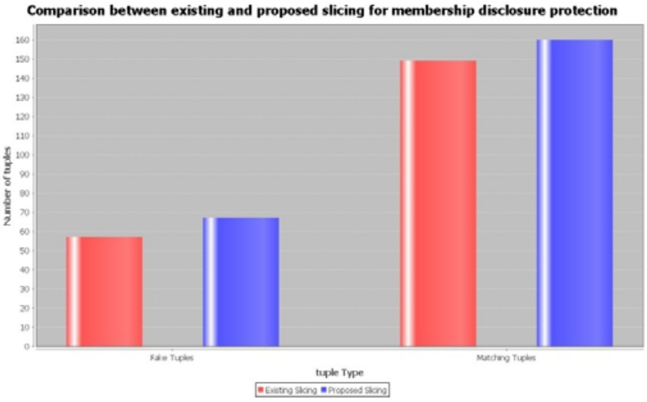

Figure 2: Comparison between slicing and cluster based attribute slicing algorithm

6. Conclusion

This proposed system developed a new technique for privacy preservation. This proposed algorithm means cluster based attribute slicing algorithm is better than slicing algorithm. It gives the more effective result than slicing algorithm. Also it takes less time to convert continuous attribute to categorical attribute. In cluster based attribute slicing skew data is handling properly.

7. Acknowledgment

The authors would like to express heartfelt gratitude towards the people whose help was very useful to complete this dissertation work on the topic of “Cluster Based Attribute Slicing for Privacy Preservation in data mining.” It is great privilege to express sincerest regards to P.G. Guide Prof Vanita S. Babanne and, for her valuable inputs, able guidance, encouragement, whole-hearted cooperation and constructive criticism throughout the duration of this work. Also thanks to H.O.D. Prof. D. N. Rewadkar for his valuable inputs and whole-hearted. Author would like to pay Sincere thank to ICCICCt 2014 organizing committee, for encouraging and allowing P.G. Students to present the dissertation for the partial fulfillment of the requirements leading to the award of M.E. degree. Author is also thankful to Director Dr. P. M. Patil and the management. He would also like to thank all the faculties who have cleared all the major concepts that were involved in the understanding of the techniques behind this work. The Dissertation is based on research work in “Slicing: A New Approach for Privacy Preserving Data Publishing” by Tiancheng Li, Ninghui Li, Senior Member, IEEE, Jian Zhang, Member, IEEE. Author is very much thankful to them for such a precious work.

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] 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.

[3] 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.

[4] X. Xiao and Y. Tao, “Anatomy: Simple and Effective Privacy Preservation,” *Proc. Int’l Conf. Very Large Data Bases (VLDB),* pp. 139-150, 2006.

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

[6] Tiancheng Li, Ninghui Li, Jian Zhang, and Ian Molloy”Slicing: A New Approach for Privacy Preserving Data Publishing” *Prof. IEEE Transactions on Knowledge and Data Engineering,* pp. 561-574, 2012.

[7] Daniela Joiţa “Unsupevised Static Discretization Methods in Data Mining” Titu Maiorescu University, Bucharest.

[8] 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.

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

[10] Tiancheng Li, Ninghui Li, Jian Zhang “Slicing: A New Approach for Privacy Preserving Data Publishing,” Proc. IEEE Transaction on knowledge and data mining engineering, VOL. 24, NO. 3, pp 561-574, 2012.

Author Profile

Neha Jamdar received the B.E. degree in Computer Science and Engineering from KIT COE Kolhapur in 2011.Currently appearing M E 2nd year Computer Engineering in RMD SSOE Pune and also working as Assistant Professor of Computer Engineering Department in ISB&M SOT Pune, India

Vanita Babane received the B.E. and M.E degrees in Computer Engineering from MBES COE Ambajogai and VIT Pune in 2005 and 2013, respectively. Currently she is working as Assistant Professor of Computer Engineering Department in RMD SSOE Pune, India