A Tabu Search Method for K-anonymity in database privacy protection

Cui Run*, Hyoung Joong Kim**
*Dept. of Information Management and Security, Korea University
**Dept. of Information Management and Security, Korea University

e-mail : cuirun@korea.ac.kr

Abstract

In this paper, we introduce a new Tabu method to get K-anonymity character in database information privacy protection. We use the conception of lattice to form the solution space for K-anonymity Character and sear ch the solution area in this solution space to achieve the best or best approach modification solution for the information in the database. We then compared the Tabu method with other traditional heuristic method and our method show a better solution in most of the cases.

1. Introduction

A number of organizations have to publish micro-data for special purposes such as demographic and public health research. In order to protect individual privacy, known identifiers (e.g., Name and Social Security Number) must be removed. In addition, this process must account for the possibility of combining certain other attributes with external data to uniquely identify individuals. For example, an individual might be re-identified by joining the released data with another (public) database on Age, Sex, and Zip-code.

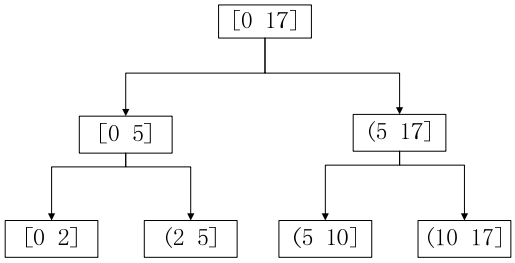
K-anonymity method is one of the most popular way s to solve this privacy problem such as in [1], [2] and [3]. In this paper, we introduce lattice conception to for m a solution space and search in this solution space. Di fferent node means different way of data modification. T abu method is used here to bring about a search proced ure. We use a greedy method to get an initial solution and search all the neighbors.

Also there are many different kinds of ways to analy sis the k-anonymity characters, such in [6], [7], [8]. So me basic idea about this area are introduced in [9], [10], [11] and [12].

2. Solution space in lattice

In k-anonymity method, the most important point is how to modify the data, and in our paper, we use a conceptio n of lattice to show the possible solution for the databa se we use, for example [4], [5]. Here is an example of how to form the lattice we use.

Firstly we focus on one attribute column such as age, and the range of the age is [0, 17]. Then we get the classification tree as follows:



(Fig 1) Classification Tree

For each level in the tree, we can assign a mark to, such a s A1, A2, A3 and A4 (the height of the tree is 4, we ignore th e original data in Fig 1). For other attributes columns, we ado pt the same procedure. Assume we have the colum n information as follows:

Age: A1 to A4,
No.: B1 to B3,
Level: C1 to C5.

Then we get the lattice L = [A, B, C], here the valu e of A B C is chosen from the corresponding marks ab ove. So in our lattice, we have totally 4*3*5 = 60 node s. And each node represents a modification way in class ification method.

3. Information Loss Evaluation

There are many different kinds of Information Loss Matrix. This is no common rule for that. So users can choose any kinds of traditional Information Loss Matrix or design by themselves. In this experiment, to achieve a non-monotone character, we adopt the Discern-ability Metric as following:

$$DM = \sum_{f_i \geq k} (f_i)^2 + \sum_{f_i \leq k} (n \times f_i) \qquad (1)$$

where f_i is the size of equivalence classes.

If DM is bigger, it means that we can distinguish dat a more easily and less information. In this case, we do fewer modifications in the data.

4. Initial solution achieved by greedy method.

In this paper, we adopt a greedy method to achieve a n initial solution which is used as a start point

As we want to adopt a Tabu search to the lattice space, we need a start point, which is a node in the lattice space. The way we get it is a greedy procedure. Form the top to the bottom, for each node; it has some in-node in the lattice graph. From these nodes, we choose the smallest information loss node k-anonymity node. Repeat the procedure, until there is no k-anonymity node.

Then the k-anonymity nodes will forms a sharp curve division in the lattice space.This solution may be optimal or local optimal.

5. Tabu search method in lattice space

The Tabu search method in this paper is shown as fo llows:

Tabu part
{
1. Initial Solution
2. Initial Tabu table and other parameter with the greedy solution above.
3. Set repeat time T
4. While T>0
   Candidate chosen;
   Update Tabu Table;
   End while
5. Return the final result.

}

Initial Solution
{
1. Node = Top node
2. Get lower neighbors of Node.
   If No Neighbors
   Break;
3. Evaluation all the neighbors, among all K-Anonymity nodes, find the one with largest DM value. Assign it to Node
4. Goto 2

}

Candidate chosen
{
1. Find all the neighbors of Node
2. Remove the neighbors in the Tabu table.
3. Evaluate each neighbors remained, keep the best answer as a global answer.
4. Based on the information loss and k-Anonymity state, assign a possibility to each neighbor.
5. Randomly choose a neighbor and save it into Node Based on the neighbors K-state and the DM values.

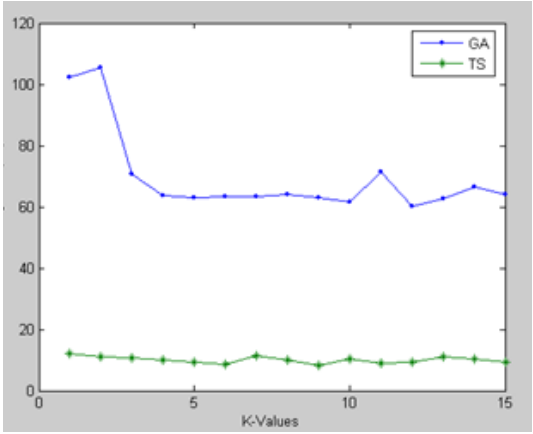
}

Update Tabu Table
{
1. Reduce the time mark of each record in the table.
2. If Time mark is 0, delete the related record
3. Check if the Tabu Table is full

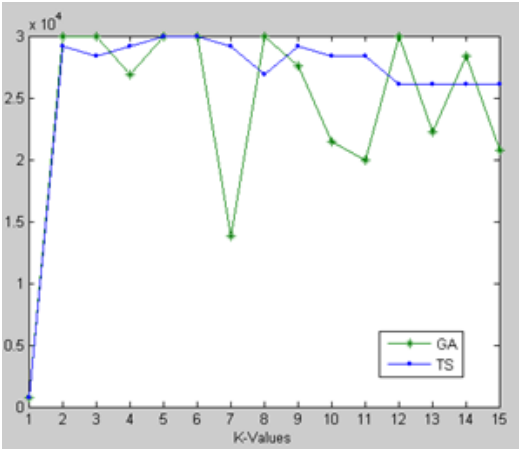
If full, delete the one entering the table first.
Add the new record: the solution node checking.
}

Experiment result and analysis

In this paper, we run the algorithms above and comp ared the result with the genetic algorithm. The database we use is Pima Indians Diabetes Database. The compar ed result is shown as follows:



(Fig 2) Compared result of Processing Time



(Fig 3) Compared result of DM values

From the result above, we can see that our Tabu met hod runs much faster and achieve a lower DM value in most of the cases.

In future work, we can modify the rules in the Tabu method and find new start point category to improve t he result.

Acknowledgements

This research is supported by Ministry of Culture, Spor ts and Tourism (MCST) as Korea Culture Content Agen cy (KOCCA) in the Culture Technology (CT) Research & Development Pro-gram 2011.

Reference

[1] Kristen LeFevre. “Mondrian Multidimensional K-Anonymity”
[2] Kristen LeFevre. “Incognito: Efficient Full Domain K-Anonymity”
[3] Hua Zhu. “Achieving k-Anonymity Via a Density-Based Clustering Method”
[4] WANG Zhi-Hui. “Clustering-Based Approach for Data Anonymization”
[5] KHALED EL EMAM. “A Globally Optimal k-anonymity Method for the De-Identification of Health Data”
[6] LATANYA SWEENEY. “K-anonymity: A model for protecting privacy”
[7] Nergiz, M.E. “MultiRelational k-Anonymity”
[8] Arik Friedman. “Providing k-Anonymity in Data Mining”
[9] Roberto J. “Data privacy through optimal k-anonymization”
[10] Charu C. “On k-anonymity and the curse of dimensionality”
[11] Gagan Aggarwal. “Approximation algorithms for k-anonymity”
[12] Avrim Blum “Practical privacy: The SuLQ framework”

Biography

Cui Run is a research scholar in Graduate School of Information Management and Security, Korea University, Korea. He received his B.S. in Harbin Institute of Technology in 2008. His research interest includes database Security, parallel and distributed computing and data mining.

Hyoung Joong Kim received his B.S., M.S., and Ph.D. degrees from Seoul National University, Korea, in 1978, 1986, and 1989, respectively. He joined the faculty member of Kangwon National University, Korea, in 1989. He is currently a Professor of Korea University, Korea. He published numerous technical papers including more than 40 peer-reviewed journal papers covering distributed computing and multimedia computing. He served Guest Editor of several journals including IEEE Transactions on Circuits and Systems for Video Technology. He is a Vice Editor-in-Chief of the LNCS Transactions on Data Hiding and Multimedia Security. His main research interests include security engineering.