Expressions of K-Anonymity with Integer Programming

*Cui Run **H. J. Kim

Center for Information Security Technologies(CIST), Korea University

silenceofcr@gmail.com

Abstract

In this paper, we introduce a new kind of expressions for privacy protection techniques in database, such as K-anonymity L-diversity and t-closeness. With such kind of expressions, we provide a new way to solve the privacy protection problems, such as Linear programming, Non-linear programming, Integer programming and so on. Also most of the heuristic techniques are also efficient to be adopted under the expressions given.

1. Introduction

Sometimes organizations have to show their secret micro-data to the public for some kinds of special usages, such as medical research and statistics. With the publication of private information, dangers come. The attackers can do the analysis to the published data with data mining method to get the deep relations hiding in the micro-data. With quasi-identifiers from the data, they can locate the records to unique people. To avoid being attacked, information protection techniques are becoming more and more important.

K-anonymity model is just one of the most popular ways to solve the privacy protection problem above. It provides modification to the tuples in the database to remove the Quasi-identifiers. Modification in suitable level is one of the more popular way. Also l-diversity and t-closeness are also well-known method.

In this paper, we will show how to change the method of K-anonymity into Integer Programming, which provide us another way to solve the modification problem without the traditional one.

2. Expressions for K-anonymity

2.1 Binary expression

The dataset is numbers in tables which are stored in the databases. The first process to the data is binary. An example is shown in Figure 1.

We can see after the binary, there are many same columns. We can move such kinds of redundant informations.

30	1 1110
26	1 1010
28	1 1100
29	1 1101

Figure 1. Binary processes to the data.

2.2 Subgroup drawing

As K-anonymity is a NP problem, it can not be solved directly up to now. As a result, we must find a suitable way to reduce the complexity of the computation. Here we choose to draw subgroup in a suitable length to implement it.

Here we draw subgroups with an adaptive scheme as follows:

Assume K is the anonymity command and N is the member number of sub-group, R is total record number in the database table, then:

if K < R/4, N = 4K
else if K < R/3, N = 3K
else if K < R/2, N = 2K
else N = K

Then we devide the database into the sub-groups with the length above. we can use any kind of classification method to implement the procedure. The residual part can be added to one of the sub-groups chosen.

2.3 Bits weighted

Different bits in the table represent different weights in its own attribute values. So we can assign a weight vector to different bits in the table as their meanings, an example is shown as follows:

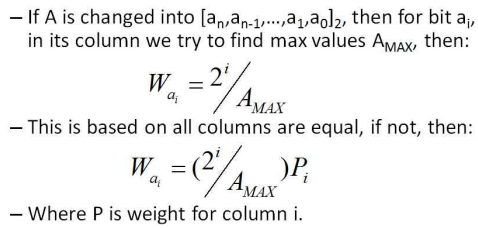

Figure 2. An example of weight processes.

2.4 Formulations

With the processing above, we can get the following 0-1 matrices:

W(M by M): weight matrix. The diagonal is weight value, others are 0.
A(M by N): 0-1 coefficients from the data table.
X(N by 1): solution vector.
Ks(N by 1): anonymity vector, all the values in it are K.

Here M is the column number of the table.

Then the K-anonymity for the subgroup can be expressed as follows:

$$Min: abs(WAX - Ks) * e_N$$
$$\begin{cases} \sum_{i=1}^{N} X_i = K \\ X_i \in \{0,1\} \end{cases}$$

Figure 3. Expressions for K-anonymity with 0-1 programming.

Here the condition for summation of X can be expended into >=,instead of =.

3. Expressions for other anonymity method

3.1 L-diversity

For l-diversity technique, the expressions can be achieved with modification to the formulations in Figure 3. The result is shown as follows:

$$Min: abs(WAX - Ks) * e_N$$
$$\begin{cases} \sum_{i=1}^{N} X_i = K \\ X_i \in \{0,1\} \end{cases} + Number(SensitiveColumns) \geq L$$

Figure 4. Expressions for l-diversity with 0-1 programming.

3.2 t-closeness

The target of t-closeness method is to minimize the distance between the distribution above and the distribution of modified data. So for a subgroup, we can modify the target function.

$$Min: \text{ distance(original data distribution, modified data distribution)}$$
$$\begin{cases} \sum_{i=1}^{N} X_i = K \\ X_i \in \{0,1\} \end{cases}$$

Figure 5. Expressions for t-closeness with 0-1 programming.

4. Conclusion

In this paper, we introduce a new way of expression for the anonymity method of privacy data protection in the database.

There are three kinds of expression models are shown in this paper. Such kind of models can be easily solved by many different methods. According to different commands in practice, we can modify the target function or constrains to reduce the computation complexity or simple the problem model to be solved. It provides a new way to solve the anonymity problems in the database.

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] Gagan Aggarwal. “Approximation algorithms for k-anonymity”
[6] Iyengar V. “Transforming data to satisfy privacy constraints”. In Proc. Of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 279-288, Edmonton, Alberta, Canada.
[7]Monte Lunacek. “A crossover operator for the k-anonymity problem”. In Genetic and Evolutionary Computation Conference.
[8] Rhonda Chaytor, “A Better Problem

Representation for k -Anonymity". In Proc. 1st ACM SIGKDD Int'l Work. on Privacy, Security, and Trust in KDD
[9] Roberto J. "Data privacy through optimal k-anonymization"
[10] Charu C. "On k-anonymity and the curse of dimensionality"
[11] KHALED EL EMAM. "A Globally Optimal k-anonymity Method for the De-Identification of Health Data"
[12] www.wikipedia.org