A T-similarity Sensitive Information Protection Method Based on Sensitive Information Gradient Partition

Linping Su\textsuperscript{1}, Zhipeng Liu\textsuperscript{*}

\textsuperscript{1}Department of control and Computer Engineering, North China Electric Power University, Beijing, China

\textsuperscript{*}email: 12019227238@ncepu.edu.cn

Abstract. Nowadays, data technology has been widely used. Researchers have also made an in-depth exploration on how to effectively protect sensitive information in massive data, including k-anonymity, l-diversity and t-similarity. In this paper, the main ideas of the three technologies are described, and the characteristics of the three technologies are analyzed. By introducing the concept of sensitive gradient when calculating the similarity distance, the defect of not being sensitive to the sensitive attribute value in t-similarity technology is improved and supplemented. Algorithm for t-similarity technology can form more effective protection for sensitive attribute values with high privacy requirements under acceptable time overhead differences, and can meet the needs of practical applications.

1. Introduction

With the development and application of data mining and data analysis technology in various fields, data has been highly valued in shopping, medical treatment, living payment and finance. There are also a large number of data sets desensitized to sensitive information and recorded individuals on the network [1]. However, for data records, simply removing the identity information can not effectively protect individual privacy. The background information obtained by attackers in other ways can be located to a record in the data table [2]. According to the research conclusion published in 2000 by latanya Sweeney, a professor at the Institute of quantitative social sciences at Harvard University, it is 87% possible to uniquely identify an American only through three information \{zip code, date of birth and gender\}. The disclosure of private information will not only affect personal interests, but also threaten organizational interests, and may even endanger national security. Because of the high risk and harm of privacy information disclosure, how to reduce the probability of a specific individual's privacy information being attacked has become a very important research content.

2. Anonymous privacy protection technology

Jiang et al. proposed k-anonymity (k-anonymity) [3], l-diversity (l-diversity) [4] proposed by Machanavajhala, and t-closeness (t-closeness) [5] to set desensitization requirements on the data tables to be published, which makes it more difficult for attackers to obtain specific individual information. To further illustrate the three types of anonymous technologies, we first introduce several basic definitions:

Define 1.1 identifier attributes: Identifier attributes allow you to directly identify a specific individual or record in a data table.

Define 1.2 Quasi-identifier attributes: Quasi-identifier attributes generally have one or more attributes. A specific record cannot be determined by a quasi-identifier or a set of quasi-identifiers. However, by combining with external information, specific records in a data table can be locked, such
as the gender, profession, and zip code attributes.

Define 1.3 Sensitive Attribute: This attribute is the data that an individual in the data table consciously wants to be protected, does not want to be associated with the individual, and cannot be published at the same time as the identifier attribute, such as the professional attribute.

Define a 1.4 equivalent class: set the published data set to $T\{T_1, T_2 \ldots T_i\}$, where $T_i$ is the first record in data set $T$; Quasi-table identifier attribute set $S\{S_1, S_2 \ldots S_i\}$, where $S_i$ is the ith quasi-identifier attribute in $S$ if $T_i, T_j$ have exactly the same value on $S$, $T_i,T_j$ are in an equivalent class.

Definition 1.5 Anonymous: Anonymous is the processing of data values to hide sensitive and private information. The main methods include generalization, suppression, disruption, clustering, etc.

Next, we will take \{name and ID\} as identifier attributes, \{score, sex and zipcode\} as quasi identifier attributes and \{major\} as sensitive attributes to introduce the three anonymous technologies and analyze their existing problems.

2.1. K-anonymity Technology
After desensitization of the data set, the number of corresponding records in each equivalent class is at least k, ensuring that each record is indistinguishable from other k-1 records, it is said to meet the k-anonymity requirements.

After the identifier \{name, ID\} attribute is proposed in the published data table, the \{score, sex, zipcode\} attribute is generalized to varying degrees, and the number of records in each equivalent class is 3 records.

However, k-anonymity technology cannot resist homogenization attack and background knowledge attack. If the attacker can lock the record in the first equivalence class, it can be determined that the required individual major must be "computer science and technology"; Or if the attacker locks the target in the fourth equivalence class and learns from the background that the target is a science and engineering major, it can also determine that the target major attribute value is "electrical engineering and automation".

2.2. L-diversity Technology
L-diversity technology hopes to make up for the deficiency in k-anonymity by setting the number of sensitive attributes in the equivalence class that can well resist background knowledge attacks. When an attacker needs to master L-1 background knowledge to associate a record with a specific individual, We say that the sensitive attribute in this equivalence class has L values that can well resist background knowledge attacks. When all equivalence classes have at least l such attribute values, we say that the published data satisfies l-diversity. Assuming that the values of sensitive attributes in Table 1 are values that can resist background knowledge attacks to a high extent, the three equivalent classes 2, 3 and 4 in Table 2 meet 2-diversity.

However, l-diversity cannot resist skew attack and similarity attack. If the attacker determines that the target record is in equivalence class 4, the target major can be judged as "financial management" through 66.7% probability; Or in an equivalent class, the major attribute value is \{computer science and technology, computer science\}, and the attacker can determine that the target major is a computer major.

In addition, in order to meet l-diversity, there are high requirements for sensitive attributes and high resource overhead, which is not necessary for some scenarios.

2.3. T-similarity technique
For the non desensitized data table, the value distribution of sensitive attributes is $p$, and the distribution of sensitive attributes in equivalent classes is $Q$. t-similarity limits the gap between $P$ and $Q$. When the distance between them is less than a given t, it is said to meet t-similarity. When all equivalent classes in the data table meet t-similarity, it is said that the distributed data table meets t-similarity requirements.

T-similarity technology considers the dual effects of the effectiveness of quasi identifier attributes and the privacy of sensitive information, finds a balance between them, and controls the information loss degree of quasi identifier attributes through the set distance t, which also increases the difficulty of
attackers in analyzing the overall and local data distribution, although it can not completely resist homogeneous attacks. Similarity attack and background knowledge attack, but it is still a better way to process data tables. However, the distance control in t-similarity technology considers the equivalence class and the data table as a whole. The privacy requirements of sensitive attribute values can not enhance the degree of desensitization for some sensitive attribute values.

### 3. T-similarity method based on gradient partition of sensitive information

In the t-similarity method based on the gradient division of sensitive information, the semantics of different sensitive attribute values are considered to desensitize the data anonymously, so as to strengthen the protection of highly sensitive attributes. The idea of the algorithm is: generalize all quasi identifier attributes to a unique equivalent class root node; Then, the same team inverse generalizes the attribute values of each quasi identifier to deal with iterative splitting, and finally forms several equivalent class nodes. When the distance between the equivalent node and the whole does not meet the t-similarity technology, the leaf node and its sibling nodes are generalized upward to the upper node; Repeat this process until all leaf equivalent nodes meet the t-similarity technical requirements.

1. Algorithm input: to publish data table D, divided sensitive attribute set D (SEN), given distance t
2. Algorithm output: release table conforming to t-similarity Technology
3. Algorithm steps:
4. Generalize all records in D to the root node, and its collection is r
5. \( R'=R,O=\emptyset \) // Assign all records in R to R ', and O is an empty set
6. \( S < - \) split (R) // completely split r to generate an equivalent class node set s
7. While (d (\{s\} > t)) // ends when the t-similarity distance of all equivalent classes in S is less than t
8. Max_ Nodes -> getdmax (\{s\}) // get the node with the largest D value in the child node set
9. generalization(children nodes) //Generalize the equivalence class node and its sibling node to the parent node
10. Until D (\{s\}) < T // until all sub equivalent class nodes in s meet the t-similarity distance requirements
11. return S

### 3.1 Sensitive gradient Division

Through the semantic consideration of sensitive attribute values, we flexibly divide the sensitivity of sensitive attribute values according to different environments and requirements, and classify and arrange them from low to high. We divide the disease sensitivity attribute values into four categories, and their sensitivity increases from top to bottom; With the current global epidemic spreading, we can specifically define fever as the value with the highest sensitivity requirements for gradient arrangement.

Suppose that the classification set of K sensitive attribute values is arranged in a gradient from low to high \( S=\{s_1,s_2…s_k\} \), The classification sensitivity on the k-th gradient is the highest, \( s_k \) is the number of sensitive attribute values in the k-th sensitive classification. The overall distribution of S is \( W=\{W_1,W_2…W_K\} \), We give the gradient value \( \omega_\_Calculation formula of I:\n\)

\[
\omega_i = \frac{(1-w_i)}{K} \quad (1)
\]

According to the formula, we can get the gradient value set \( w = \{w_1,w_2…w_k\} \) of sensitive attributes corresponding to S.

### 3.2 Distance Measurement

The methods used to calculate the distribution distance of t-similarity technology mainly include KL divergence, EMD distance, Hellinger distance. KL divergence is the most commonly used, but its distance calculation method can not calculate the distribution probability of 0. In this paper, we use js divergence distance mentioned in reference to measure the distance of t-similarity technology. JS divergence is further calculated based on the results of KL divergence calculation, and makes up for the defect when the distribution probability in KL divergence calculation is zero.
This paper only considers the case that the value of sensitive attribute is in discrete state. Let the state of two discrete distributions be \( p \) and \( Q \), then the formula for JS divergence to calculate the distance between the two distributions is:

\[
JS[P, Q] = \frac{1}{2} (KL[P, \text{avg}(P, Q)] + KL[Q, \text{avg}(P, Q)])
\]  

(2)

Where \( KL[x, y] \) is the KL divergence calculation of \( X \) and \( y \), and \( \text{AVG}() \) is the average of the two distributions.

The gradient set \( w = \{w_1, w_2, \ldots, w_k\} \) is divided according to the sensitive gradient. By increasing the influence degree of attribute value distribution difference with higher sensitive gradient value on the overall distance, the calculation method of JS divergence distance between distribution \( P \) and distribution \( q \) is given:

\[
D[P, Q] = \frac{1}{k} \sum_{i=1}^{k} w_i JS[p_i, q_i]
\]

(3)

4. Conclusion

T-similarity technology based on sensitive attribute value gradient division can provide better protection for sensitive attributes, and when the sensitive gradient increases, the ability to resist attacks will be stronger. Through the analysis and comparison of \( k \)-anonymity, \( l \)-diversity and t-similarity technologies, this paper adds the gradient division of sensitive attribute values on the basis of t-similarity technology, and easily takes the semantics of sensitive attribute values into account in t-similarity algorithm. Next, we will explore how to improve desensitization efficiency on the basis of considering semantics, so as to cope with the current scene of explosive growth of data volume.

Reference

[1] Huang Xuezhen. (2015) Research on models and methods of privacy protection data publishing. Beijing Jiaotong University, Beijing.

[2] Amin S U, Alsulaiman M, Muhammad G, et al. (2019) Multilevel weighted feature fusion using convolutional neural networks for EEG motor imagery classification in: IEEE Access, 7:18940-18950.

[3] Soria-comas J, Domingo-ferrer J. (2012) Probabilistic k-anonymity through micro aggregation and data swapping. In: Proceedings of the 2012 IEEE International Conference on Fuzzy Systems. Piscataway. pp. 1-8.

[4] Bhattacharyya D K. (2012) Decomposition: improving l-diversity for multiple sensitive attributes. In: Proceedings of the 2012 International Conference on Computer Science and Information Technology. Berlin. pp. 403-412.

[5] Wang Qiuuye, Ge Lina, Geng Bo, Wang Lijuan. (2018) Classification based on multi sensitive attributes \((a_i, k, m)\) - anonymous privacy protection method. Computer Applications, 38 (1). pp. 67-72