International Journal of Multimedia and Ubiquitous Engineering
Vol.10, No.10 (2015), pp.119-126
http://dx.doi.org/10.14257/ijmue.2015.10.10.13

MADARS: A Method of Multi-Attributes Generalized Randomization Privacy Preserving

[1]Guo Xiaoli, [2]Zhang Jiajia, [3]Qu Zhaoyang, [4]Wang Yongwen and [5]Guo Ping

[1,2,3,4] College of Information Engineering, Northeast Dianli University, Jilin Province

[5]liaoning Jianzhu vocational university, Liaoning Province

243589657@qq.com,414899263@qq.com,qzywww@mail.nedu.edu.cn,995802183@qq.com,490750519@qq.com

Abstract

With the extending in the domain of data mining application, the research of the privacy preserving in data mining based on k-anonymity becomes more and more concerned. The key point of the research is to protect data while it's still effective in data mining. However, these methods result in too much information loss, and Multi-dimension bucketization does not generalize quasi-identifiers, which make the anonymized data easy to suffer from linking attack. To overcome these drawbacks, we put forward the thought of data classification processing, named MADARS. MADARS first deal with identifier attributes based on MA-Datafly and then randomized multi-sensitive attributes. Experiment results show that compared with the widely used algorithm Incognito the method can greatly improve the efficiency of the privacy protection while reduce the less personal information and the effectiveness of data is greatly increased. This work is supported by National Natural Science Foundation of China (No.51277023).

Keywords: *classify; k-anonymity; generalization; stochastic disturbance; multi-sensitive attributes*

1. Introduction

With the rapid development of data mining and database, privacy and security of personal information is gradually becoming more and more concerned. The data are further used by the data miner for the analysis purpose which helps the organizations for gaining useful knowledge. These data may contain sensitive or valuable information of any individuals [1]. At this stage, to maintain the validity of the data and at the same time ensure data privacy and security in the case of mining is the purpose of privacy protection [2]. In recent years, many researchers have done much research on k-anonymity algorithm and it's improved algorithms [3-4, 5, 10-11].At present, many privacy protection methods remain in dealing with a single sensitive attribute. And in multidimensional sensitive attributes possibilities should be taken into account before the release of their attackers' attacking when data is reasoning disclosure. So the above methods cannot guarantee the security of good information in multi-sensitive attributes dataset. For this reason, many researchers have proposed privacy protection methods multidimensional sensitive properties.

2. Related Work

Firstly, Samarati and Sweeney proposed the idea of k-anonymity [3].It used the basic idea of Datafly in the algorithm and projected the publish data on the quasi-identifiers as a sub-table. Loop the sub-table as follow: When the record number is greater than k (which does not meet the constraint of k-anonymity), Select the largest number of all

ISSN: 1975-0080 IJMUE

attributes. Generalized the attribute. After the end of the loop, recording the attributes which does not satisfy the remaining k- anonymity constraints. Literature [4] algorithm processing quasi-identifier attribute to avoid excessive loss of the foregoing information processing algorithm caused excessive anonymous, so that more property generalization grades k-anonymity meet more requirements, but does not consider the sensitive attribute of anonymity.

To improve the effectiveness, literature [5-6] alignment identifier attribute randomized algorithm processing, and sensitive attribute partial k-anonymity processing. But they did not consider the situation of multi-sensitive. Literature [7] proposed the idea of multi-sensitive bucketization thought, combined with the idea of decomposing a multidimensional space priority algorithm proposed maximum privacy-sensitive data distribution method for multi-dimensional attributes, data mining also have some losses.

3. Proposed Method

Based on previous work this paper proposed MADARS: support multi-attributes generalization of randomized multi-sensitive attributes privacy protection method. According to the character of data attributes, classified the attributes as Ei, Qi and Sn. Processing separately, so that the validity of the data set much higher than the traditional anonymous algorithms, dataset has also been efficiently protected.

3.1. Formal Definition

Each tuple represents an individual information. This article contains the display identifier attribute, and the quasi-identifier attribute properties of three types of sensitive properties.

Definition 1. Display identifier (referred to as Ei) attributes uniquely identify the attributes of identity information. Such as name, identity card number.

Definition 2. Quasi-identifiers (referred to as Qi). Refers to the information which on the joint to the obtain additional information can identify the user's identity attribute uniquely. And all data holders may be able to identify with external information connected to private information set.

Definition 3. Sensitive-attributes (referred to as Sn). Refers to personal sensitive information, such as religion, wages, health and so on.

Definition 4. k-anonymity [10].The basic idea is: each record on the released data set cannot be distinguished from other k-1 records. That is satisfied with the k-anonymity table, the possibility of a record re-identified risk is no more than 1/k.

Definition 5. Generalization [11]. Generalization process follows the process of elimination, each attribute value is to determine the interval between the size of the range.

3.2. Precondition

Using the bottom-up support for multi-attribute k- anonymity algorithm generalization in quasi-identifier attribute data sets, here we use f0 for the generalized functions of first layer, f1 for the second layer, and so on.

3.3. Generalization

Algorithmic process:

Input：Dataset T, positive integer k;

Output: k-anonymity Table ;

Steps:

1.Determine the character of identifier attributes as display identifier Ei, quasi-identifier attribute Qi, and sensitive attribute table Sn;

2. Remove or suppress Ei;

3. Summarized domain and divided the hierarchical trees for each types of quasi-identifier attribute;

4. Generalize the maximum value attribute by domain generalization grade tree, if the tree has reached the highest layer (Suppose there are N layer), data sets is still not satisfied with k-anonymous, then replace the attribute value in the original table T with the function f0,form the new table T1;

5. Then treat T1 as the original table, find the maximum value attribute in the table;

6. If the currently found attribute is still the one that the maximum value attribute, then replace the attribute value in the original table T1 with the function f1, form the new table T2;

7. Repeat this procedure until you find another attribute g;

8. Generalize g, followed by a recursive judgment, if it meet the k-anonymity, then break out;

9. Else, recursively for the next layer until the data set to meet the k-anonymity;

10. Stop.

In the algorithm, we must delete or suppress the display attribute first, and used to uniquely identify the individual character attributes, and then using the algorithm for the data set.

3.4. Randomization

Randomize the sensitive attribute dataset and use the properties of transition probability matrix.

Algorithmic process:

Input: *Sensitive dataset S, the transition probability matrix M, and the mapping matrix P in the size of i*j (P is a mapping matrix between T and M);*

Output: Conversion Table C;

Steps:

*1. Generated the size of j * j transition probability matrix of M randomly;*

2. Generated the mapping matrix P randomly;

3. Distribute each transition probability matrix M (Ml, M2,⋯, Mj) and the matrix S (S1, S2, ⋯, Sj) one to one mapping by mapping matrix P;

4. Rearrange elements from the top down follow the position in matrix M, if the position on top have been used, then select the second high position of elements in matrix M, and so on. If there are two or more positions of elements in matrix M mapping in matrix S, then the selecte the front position of elements;

5. Recombinant matrix S;

6. Reconstruction Table C;

7. Stop.

First, generate probability matrix M and mapping matrix P randomly in the algorithm. Rearrangement the multi-sensitive attribute in set S (S1, S2,⋯, Sn) according to the mapping between the elements. For the probability matrix M, take the largest position of the first column. Take the first position, if there are as the maximum two or more same value in the column; if the column position is already occupied, then choose the position of the next big value. Rearrange substitution table S according to this method.

4. Example

Generated probability matrix M randomly, the mapping matrix P = [3,1,2] mean that M3 applied to S1, M1 is applied to S2, M2 applied to S3. According to the multi-sensitive attributes randomized algorithm rules, choose the position of the maximum value in M, if the column position of the maximum is already occupied, then select the position of the

second largest value, if there are the same value two or more column then take position of the previous value, and so on.

Table 1. Data Set T

Qi	Ei		Sn		
Name	**Age**	**Sex**	**Race**	**salary/week**	**illness**
Lili	48	F	Black	6100	C
Xiao Ming	39	M	White	8000	T
Leaves	42	F	White	2200	H
blossoming	27	f	Yellow	4500	H
Peas	24	m	Yellow	5300	G
Dragon	33	m	White	3300	C

$$M1=\begin{bmatrix} 0.6 & 0.4 & 0.2 & 0.8 & 0.3 & 0.2 \\ 0.9 & 0.7 & 0.1 & 0.4 & 0.9 & 0.2 \\ 0.6 & 0.5 & 0.8 & 0.5 & 0.4 & 0.7 \\ 0.5 & 0.1 & 0.4 & 0.3 & 0.1 & 0.5 \\ 0.2 & 0.5 & 0.3 & 0.4 & 0.6 & 0.3 \\ 0.1 & 0.5 & 0.1 & 0.2 & 0.2 & 0.4 \end{bmatrix} \quad M2=\begin{bmatrix} 0.83 & 0.22 & 0.74 & 0.63 & 0.57 & 0.34 \\ 0.25 & 0.17 & 0.44 & 0.32 & 0.76 & 0.33 \\ 0.65 & 0.36 & 0.77 & 0.87 & 0.64 & 0.77 \\ 0.28 & 0.47 & 0.47 & 0.25 & 0.16 & 0.33 \\ 0.46 & 0.47 & 0.64 & 0.75 & 0.27 & 0.47 \\ 0.15 & 0.36 & 0.35 & 0.57 & 0.55 & 0.24 \end{bmatrix}$$

$$M3=\begin{bmatrix} 0.4 & 0.8 & 0.6 & 0.4 & 0.1 & 0.3 \\ 0.7 & 0.9 & 0.9 & 0.1 & 0.6 & 0.2 \\ 0.3 & 0.8 & 0.6 & 0.4 & 0.7 & 0.5 \\ 0.3 & 0.3 & 0.5 & 0.8 & 0.2 & 0.1 \\ 0.5 & 0.6 & 0.6 & 0.2 & 0.4 & 0.1 \\ 0.6 & 0.5 & 0.3 & 0.1 & 0.2 & 0.3 \end{bmatrix}$$

As shown in this example, the position of the maximum M3 of the first line is 2; it is the maximum position 2 and 3 in the second row, the location should be selected the prev value of 2, but the first row has selected this position, so select the second largest value of the position of 1; The maximum position in the third row of 2 has been selected, then choose the second largest value of the position which is the position of 5; Similarly, choose the location for each line, and ultimately determine the order in M3 of each value: the position should be 2,1,5,4,6,3.Mapping M3 to dataset S1 (= race),the element position should be: 2 (white), 1(black), 5 (yellow), 4 (yellow), 6 (white), 3 (white),form a new column C1. Similarly, we can also get C2, C3 columns, and construction a new Table C. Shown as Table 2.

Table 2. New Dataset C

Qi	Ei		Sn		
Name	Age	Sex	race	salary/week	illness
Lili	48	f	white	4500	C
Xiao Ming	39	m	black	6100	G
Leaves	42	f	yellow	2200	H
blossoming	27	f	Yellow	8000	T
Peas	24	m	white	5300	H
Dragon	33	m	white	3300	C

5. Steps of Scoring

Definition 6. Information loss [11]. Some information would loss in the data anonymization process.

In this paper, Weight loss information and to measure the pros and cons of anonymity. In this paper, Information loss based on the weight sum of different types generalization. And direct its strategy.

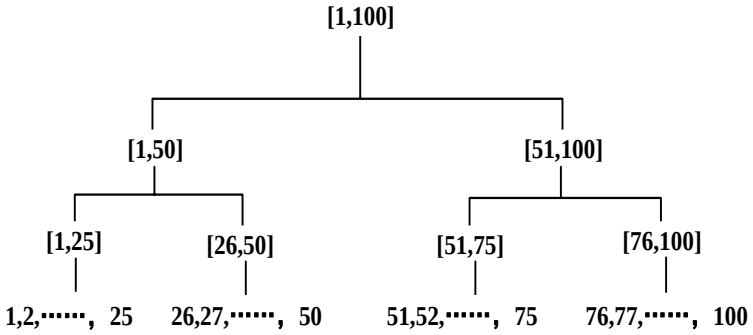

Figure 1. Numeric Type Generalization Tree

When the attribute is numeric, the information loss recorded as:

$$\text{Infoloss}(r_i) = \frac{r_{ib} - r_{ia}}{|R|} \tag{2-1}$$

Where r_i represents any attribute of a generalization of the elements in r, r_{ib} is the upper limit of the generalization range of r_i, r_{ia} is the lower limit. R is the value range, and $|R|$ is its value.

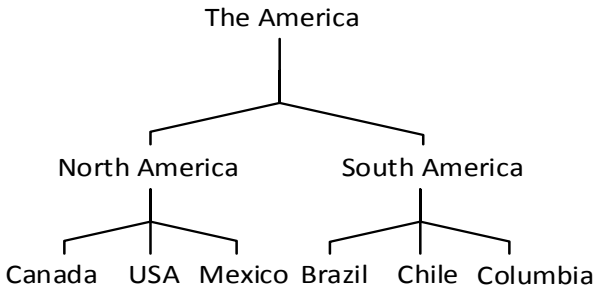

Figure 2. Categorical Type Generalization Tree

When the attribute is categorical, the information loss recorded as:

$$\text{Infoloss}(r_i) = \frac{|H_i| - 1}{|D_i| - 1} \tag{2-2}$$

Where H_i represents the covered leaf nodes after the corresponding nodes generalized in the set; D_i represents the set of all the leaf nodes. $|H_i|$ and $|D_i|$ represent node numbers. the covered leaf nodes D_i of node H_i is the leaves of the tree node H_i.

Then, take the weight sum as the final Infoloss:

$$\text{Infoloss}(r) = \sum_{i=1}^{n} \lambda_i \text{Infoloss}(r_i) \tag{2-3}$$

λ_i is the weight of each attribute($\sum_{i=1}^{n} \lambda_i = 1$)

6. Experiment

Experiment hardware configuration: Inter (R) Core (TM) i5-4200U 1.60GHzCPU, 4GB RAM, software configuration: Microsoft Windows8 SP2, JDK6.0, MATLAB R2013a. And the experimental operate five times averaging value as the final result.

6.1. Experimental Data Sets

In this experiment, the test dataset come from Adult dataset of UCI machine learning repository in Irvine University, this dataset has been widely used in study of data anonymity, and it is about 5.5M. First, remove the missing attributes in data records, after preprocessing，left 30,163 elements. Choose 9 attributes in the dataset:{Age，Gender，Race，Education，Native Country，Work Class，Salary Class，Occupation，Marital Status．}(Occupation，Marital Status，Salary Class are Sensitive attributes, Sn).The rest of the attributes are quasi-identifier attributes: Qi, where age is a numeric attribute, the rest five are categorical attributes.

Table 3. Experimental Data Table

No.	attribute	Type	attribute value Number	hierarchy tree height
1	Age	Numerical	74	4
2	Gender	Classification	2	2
3	Race	Classification	5	2
4	Education	Classification	16	4
5	Native country	Classification	41	3
6	Work class	Classification	8	3

6.2. Analysis of Experimental Results

IL1 represents information loss of Incognito algorithm, ET1 represents it's time consuming; IL2 represents information loss of MA-Datafly algorithm, ET2 represents it's time consuming; IL3 represents information loss of MADARS.

Table 4. IL and ET Varies with the Value of QID

QID	IL1	ET1	IL2	ET2	IL3
1	0.333	0.480	0.27	0.141	0.13
2	0.45	2.827	0.391	0.517	0.18
3	0.667	7.263	0.5	1.692	0.25
4	0.733	13.438	0.648	2.435	0.37
5	0.8	21.515	0.7	3.610	0.46
6	0.83	32.316	0.77	4.362	0.52

The MADARS method processing Qi with MA-Datafly, and Sn with randomize algorithm, but as the amount of data increases exponentially in randomize algorithms, the number of iterations of the matrix increases, and consume a lot of the stacks and Ram, so that the current environment can be achieved under about the 1500 line data. We intended to build a Hadoop platform to parallel processing when data amounts are huge, the processing time is about 40ms.

Table 5. Running Time of Randomized Algorithms

The amount of data	time（ms）
10	0
50	1
100	5
500	10
1000	20
1500	40
1510	41

Figure 3 is MADARS method, MA-Datafly algorithm and Incognito algorithm, compared information loss with the change of QID.

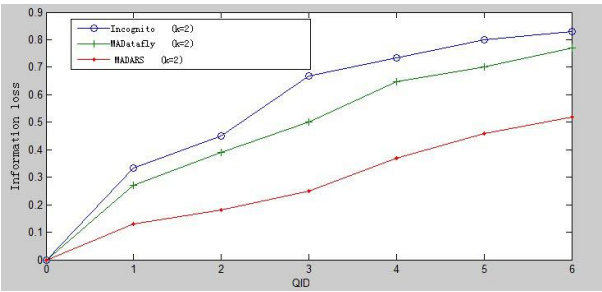

Figure 3. Loss Information Varies with the Value of QID

Figure 4 is a comparison of their time loss. And Incognito algorithm takes a lot of time in searching attributes from a domain induction level tree to another.

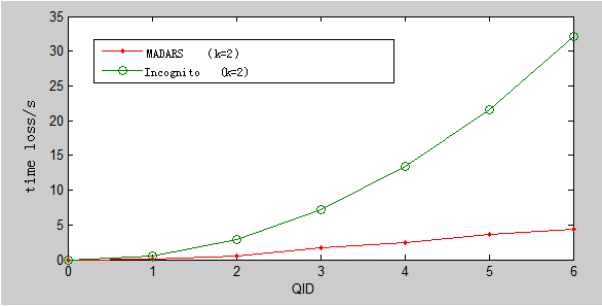

Figure 4. Loss Time Varies with the Value of QID

7. Conclusion

This paper studies the problem of multi-sensitive attributes of privacy protection in data mining, And propose MADARS method, processing Qi with MA-Datafly, and Sn with randomize algorithms to overcome shortcomings: large information loss, low utility data in existing algorithms. This method protect the relationship between the high interdependence attributes. Experimental results show that the proposed method can effectively protect sensitive information from disclosure, while maintaining high data utility and low information loss, and the efficiency high.

References

[1] M. Dhanalakshmi and E. S. Sankari, "Privacy preserving data mining techniques-survey", Information Communication and Embedded Systems (ICICES), February 27-28, **(2014)**.
[2] Vaidya, J., Shafiq B., W. Fan, D. Mehmood and D. Lorenzi, "A Random Decision Tree Framework for Privacy-Preserving Data Mining", IEEE Transactions on Dependable and Secure Computing, vol. 11, no. 5, September/October **(2014)**.
[3] Sweeney L., "K-anonymity: A model for protecting privacy", International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 5, **(2002)**.
[4] L. Pin, Y. W. Bing and C. N. Sheng, "A Unified Metric Method of Information Loss in Privacy Preserving Data Publishing", Second International Conference on Networks Security Wireless Communications and Trusted Computing (NSWCTC), **(2010)**.
[5] X. Guo and X. Han, "Research on Grid Knowledge Collaborative Discovery Strategies", Journal of Northeast Dianli University, **(2014)**.
[6] Q. Zhaoyang, C. Shuai, Y. Fan and Z. Li, "An Attribute Reducing Method for Electric Power Big Data Preprocessing Based on Cloud Computing Technology", Automation of Electric Power Systems, **(2014)**.
[7] Manish S., Atul C. and Manish M,. "An efficient approach for privacy preserving in data mining", 2014 International Conference on Signal Propagation and Computer Technology (ICSPCT), August **(2014)**.

[8] Y. A. Rahim and S. Sahib, "Randomization techniques in privacy studies", Computing and Convergence Technology (ICCCT), **(2012)**.
[9] Q. Liu, H. Shen and Y. Sang, "A Privacy-Preserving Data Publishing Method for Multiple Numerical Sensitive Attributes via Clustering and Multi-sensitive Bucketization", Parallel Architectures, Sixth International Symposium on Algorithms and Programming (PAAP), **(2014)**.
[10] Sweeney L., "Achieving k-anonymity privacy protection using generalization and suppression", International Journal on Uncertainty，Fuzziness and Knowledge-based Systems, **(2002)**.
[11] Agrawal R. and Srikant R., "Privacy preserving data mining", Proceedings of the ACM SIGMOD Conference on Management of Data (SIGMOD), Dallas, Texas, **(2000)**.
[12] J. Wang, Y. C. Luo, Y. Zhao and J. Le, "A Survey on Privacy Preserving Data Mining", Database Technology and Applications, April 25-26, **(2009)**.
[13] M. B. Malik, M. A. Ghazi and R. Ali, "Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects", Third International Conference on Computer and Communication Technology, **(2012)**.

