International Journal of Security and Its Applications
Vol. 10, No. 11 (2016), pp.229-248
http://dx.doi.org/10.14257/ijsia.2016.10.11.20

FSSPCM: Fuzzy Publication of Data for Privacy Preserving

Yan Yan[1, 2], Xiaohong Hao[2] and Wanjun Wang[3]

[1.] *School of Electrical and Information Engineering, Lanzhou University of Technology, Lanzhou, China*
[2.] *School of Computer and Communication, Lanzhou University of Technology, Lanzhou, China*
[3.] *School of Digital Media, Lanzhou University of Arts and Science, Lanzhou, China*
yanyan@lut.cn, haoxh@163.com, wangwanjun1@163.com

Abstract

The rapid development of information technology makes it convenient to release, collect, store and analyze various types of data. At the same time, how to protect the privacy of individual and prevent disclosure of sensitive information during data publication has become a major challenge. K-anonymity method is the most widely used privacy protection model and has been well researched. However, generalization and suppression operations used in K-anonymity methods require high computational effort and cause excessive loss of original information, which will greatly reduce the availability of data after publishing. The paper proposed a transformation algorithm for privacy preserving data publishing based on fuzzy semantic set pair cloud model (FSSPCM). It transforms the sensitive attributes into the form of fuzzy semantic values, and privacy of individual has been maintained because exact values cannot be predicted after data publishing. In order to enhance the availability of data after publishing, semantic distinction (SD) and reserve degree (RD) are designed to reflect relationships between original data and fuzzy semantic information after transformation according to different characteristics of numerical sensitive attributes and categorical sensitive attributes. Experiments and analysis demonstrate the effectiveness of the proposed method both on numerical and categorical sensitive attributes. Classification performed on original and transformed information proves the proposed method maintains higher clustering similarity after fuzzy transformation, which will provide better availability for data mining and other processing.

Keywords*: Privacy Preserving Data Publishing, Fuzzy Semantic, Set Pair Analysis, Cloud Model*

1. Introduction

With the rapid development of mobile communication technology and the widespread of intelligent terminals, more and more data have been generated and collected, which brings about great conveniences to people. Meanwhile, a great deal of private information (such as bank deposits, medical records, credit records, shopping habits, travel information and other sensitive information or records) has been collected, processed, shared or used without control. People's private information has been got by illicit collecting, analyzing and reasoning, which brings about great threat on their normal life and personal safety [1-3]. Therefore, how to prevent the leakage of privacy information during data publication is the key problem of information security.

Traditional privacy preserving methods usually adopt anonymous publication after deleting identity attributes (such as name, identity card number, *etc*.) of data

ISSN: 1738-9976 IJSIA

which can uniquely identify an individual. Sweeney and Samarati pointed out that sensitive information can still be disclosed by "linking attack". Therefore, they proposed K-anonymity method [4-5] to handle the privacy issues during data publication. After that, many scholars have carried out lots of researches based on K-anonymity rules and put forward some improved strategies and anonymous algorithms [6-12]. The main approach to achieve K-anonymity privacy preserving is generalization, which defines attributes can be associated with others as quasi identifiers. Some more generalized values are used instead of the exact values, and tuples in original table can be divided into some equivalence classes. Each equivalence class contains at least K ($K \geq 2$) tuples, and they have the same value on quasi identifiers. While the published data are linked with some external information, each tuple cannot be distinguished from other K-1 tuples within a equivalence class, so that the privacy of user can be protected in a certain degree.

However, there are still some shortcomings in K-anonymity rules. Firstly, although it cuts off the contact between individual and a certain record, it did not destroys the relationship between individual and sensitive information. Therefore, when attackers get some knowledge about quasi identifiers of certain individual, they will directly obtain his sensitive information through the published data. If all the records within the same equivalence class have the same value of sensitive attribute, attackers can know exactly the sensitive information of all individuals. That is to say, K-anonymity rules cannot resist the background knowledge attack and homogeneous attack. Secondly, calculation of an optimal K-anonymous data is an NP-Hard problem [11]. When dimensions of quasi identifiers are higher, more computational efforts have to be paid to get anonymous data in line with K-anonymity rules by generalization, which results in excessive loss of information. In extreme case, all the tuples are generalized into one equivalence class, and the availability of anonymous data will be seriously affected. Finally, K-anonymity rules provide the same level of privacy protection for all the tuples in the same table, which cannot satisfy personalized privacy needs for different users with different data in different situations. In view of the above problems, it is necessary to study other simpler and more personalized privacy preserving method for data publishing.

The paper addresses the problem of privacy preserving data publishing by combining the theory of set pair analysis and cloud model. A privacy preserving fuzzy publishing method is proposed, which can not only realize the protection of sensitive information but also reduce the cost of computation greatly. In view of the different characteristics between numerical sensitive attributes and categorical sensitive attributes, different parameters are designed to reflect the differences and connections between original data and fuzzy semantic information after transformation, in order to provide good availability of sensitive attributes after fuzzy publishing.

The rest of this paper is organized as follows. Section 2 reviews some previous research work related to ours. Section 3 formalizes the underlying concepts of set pair analysis theory and the cloud model. Section 4 proposes the fuzzy semantic set pair cloud model (FSSPCM) for the fuzzy publication of numerical sensitive attributes. Section 5 presents how to use the proposed algorithm for the fuzzy publication of categorical sensitive attributes. Section 6 experimentally demonstrates the efficiency and availability of the proposed method. Section 7 is the conclusion of the paper.

2. Related Work

In order to avoid homogeneous equivalence classes in anonymous data, Machanavajjhala *et al* proposed *l-diversity* model [6], in which at least l ($l \geq 2$) "well-represented" sensitive values are required in each equivalence class. Therefore, attackers have at most $1/l$ confidence to infer the sensitive information of target individual, and the ability of malicious attacker to infer individuals' sensitive information has been reduced. *l-diversity* model has well solved homogeneous attack and has also reduced the risk of privacy leakage caused by background knowledge attack to some extent. However, under certain circumstances, *l-diversity* still cannot meet the needs of privacy preserving data publishing. For example, when the distribution of sensitive attribute in the dataset is skewed (occurrence frequency of sensitive values are quite different from each other), though the released data are subject to *l-diversity* model, it cannot completely prevent the leakage of sensitive information. To solve this kind of problems, Li *et al* considered the relationships between global privacy and individual privacy and proposed *t-closeness* model [7]. In this model, approximation degree of two distributions have been measured by the function of earth mover's distance (EMD), and the differences of distribution between sensitive attributes and the whole published table are required to be no more than *t*. Reference [8] proposed a complete anonymous algorithm framework based on *t-closeness* model, which is called SABRE. All of the above methods use generalization technology to meet the privacy requirement of anonymous data. The fundamental disadvantage of this technology is that a great deal of information within the original data would be lost during generalization, thus the availability of published data will be greatly reduced.

Besides generalization, other methods are also be used to achieve privacy protection of data publishing. Xiao *et al* proposed Anatomy method [13], in which quasi identifiers and sensitive attributes are released in two separate tables. There has no generalization process on quasi identifiers, so the method captures a large amount of correlation in original data and has greatly improved data availability. Besides, it weakens the link between quasi identifiers and sensitive attributes by using a grouping mechanism. Rastogi *et al* [14] use perturbation technique to achieve anonymity. They firstly set a retention probability "p" and generate a random number $x \in [0, 1]$ for all the records in the table waiting to be published. If $x \leq p$, the sensitive value of record will be preserved, otherwise it will be replaced by another sensitive value within the range of sensitive attributes. In reference [15] and [16], some geometric methods are used to achieve equidistant transformation for data point in hyperspace, such as translation, scaling, rotation and hybrid. The disadvantage is that the similarity between data will be destroyed and result in some error in clustering. Wightman *et al* [17] put forward another method based on matrix obfuscation and applied it in the protection of location information in LBS. It needs lower computational effort, but the geometric and morphological transformation on numerical sensitive information is likely to destroy their original meaning and reduce the availability of data.

Reference [18] introduced the theory of fuzzy mathematics into the research of privacy preserving data publishing, and proposed a new way to publish sensitive data by using their fuzzy semantic forms. This new method reserved the entire utility of quasi identifiers and has minimum overheads. Defect of this method is that the value of membership can only reflect the degree that sensitive attribute is subordinated to the fuzzy semantic, but cannot distinguish the size of data after fuzzy semantics transformation, which restricts the availability of sensitive attributes. In reference [19] fuzzy offset degree was proposed to distinguish different sensitive attributes mapped into the same fuzzy semantics. Vague set

theory adopted in this paper describes objectives both from the positive and the negative aspects by using true membership function and false membership function, which overcomes the shortcomings of single membership degree in fuzzy mathematical method to some extent. However, vague set theory only discusses objectives from two extreme aspects without considering the transitional change between true and false membership function. Therefore, it cannot describe the changing process of fuzzy states.

Some other references [20-24] also discussed privacy preserving methods based on fuzzy theory from different aspects. Reference [20] addressed the problem of privacy preserving in data mining by transforming the sensitive attributes to fuzzy attributes (for example: very low, low, medium, high, very high). The method used fuzziness to symbolize improbable, prospect and approximation, therefore, exact value cannot be predicted and privacy of individual has been maintained. In order to explain the slow assessment of the associated elements in relation to a set, the method defined a membership function $\mu \to [0,1]$, and used different kinds of membership functions (Linear membership function, Gaussian membership function and Triangular membership function) on different set of fuzzy attributes. Experimental results demonstrated the effectiveness of fuzzy anonymization and achieved better accuracy of mining results. Reference [23] protected the confidential attributes by using fuzzy logic. Unlike the method used in reference [20], the proposed algorithm chose an S-shaped fuzzy membership function and transformed the sensitive attributes into some distorted data. Accuracy of the proposed algorithm was measured by using classification and clustering techniques, and classification performed on original and perturbed data were relatively same. Reference [24] proposed fuzzy based data transformation methods for privacy preserving clustering in database environment, in which various experiments are conducted by varying the fuzzy membership functions such as Z-shaped fuzzy membership function, Triangular fuzzy membership function and Gaussian fuzzy membership function. Furthermore, it proposed a hybrid method as a combination of fuzzy data transformation approach and random rotation perturbation. Experimental results proved that the hybrid method provides better clustering quality than fuzzy membership functions.

3. Fundamental Definitions

Current processing methods for large-scale information have the features of complexity, randomness, diversity, ambiguity and uncertainty. Although the technology of computer hardware and software are able to solve many practical problems in reality, the things in real world are quite fuzzy, uncertain and random. In addition with complexity and ambiguity of the thinking of decision makers, it is really difficult for people to use some strict and accurate way to express objective things. More often, some fuzzy and uncertain expressions (such as "may", "might", "probably", "almost", "good", “bad”, "medium", *etc*.) are much closer to the understanding of human.

Set pair analysis is one of the mathematical tools for the processing of certain and uncertain information, which is developed on the basis of fuzzy mathematics, fuzzy set theory and intuition fuzzy theory. It sets up a mathematical description of the certain and uncertain relationships for attributes of objectives by establishing positive, uncertain and negative connection number functions on a pair of sets. The theory of set pair analysis overcomes deficiencies of fuzzy set theory in characterizing transitional change between true membership function and false membership function, and uses connection number function to reflect mutual variation between fuzzy, certain and uncertain concepts. So it is more scientific,

reasonable and effective to realize privacy preserving transformation by using fuzzy semantic set pair analysis theory.

In view of the deficiencies of probability theory and fuzzy mathematics in dealing with uncertainty issues, Professor Li Deyi from China Academy of Engineering proposed cloud model [25] to study the relevance between fuzziness and randomness. Cloud model uses three digital features to describe the overall characteristics of concepts in natural language, which are "expectation", "entropy" and "hyper entropy". The determinate degree of cloud droplet reflects the ambiguity, while the cloud droplet itself is a random value. Therefore, the cloud model of a certain concept represents not only the randomness of this concept, but also the fuzziness as well as the relevance between randomness and fuzziness, and it sets up the mapping relationship between qualitative and quantitative.

Definition 1 (Connection Number Function [26] [27]). Suppose X is a non-empty set, for $A=\{\langle x,a_A(x),b_A(x),c_A(x)\rangle \mid x\in X\}$ on the definition domain, $a_A(x)$ represents the positive degree of the element x in X belongs to A, $b_A(x)$ is the uncertain degree and $c_A(x)$ is the negative degree. Connection number function can be defined as:

$$\mu_A(x)=a_A(x)+b_A(x)i+c_A(x)j \tag{1}$$

In which $a_A(x):x\to[0,1]$, $b_A(x):x\to[0,1]$, $c_A(x):x\to[0,1]$, correspond with the normalization condition $a_A(x)+b_A(x)+c_A(x)=1$. $i\in[-1,1]$ is the coefficient of uncertain degree. j is the coefficient of negative degree, normally can be set $j=-1$.

Definition 2 (Potential function). Potential function of set pair analysis is defined to reflect the connection tendency between certain and uncertain information, which can be expressed as:

$$shi(\mu)=a_A(x)/c_A(x) \tag{2}$$

If $shi(u)>1$, it is a positive potential; if $shi(u)=1$, it is an average potential; if $shi(u)<1$, it is called a negative potential. Actually, potential function solves the problem by carrying out a simple "clustering" on the positive potential (feasible plan), average potential (general plan), and negative potential (infeasible plan).

Definition 3 (Hesitating potential function). Hesitating potential function describes the variation tendency of certain and uncertain information, and it reflects the deviation degree of hesitation to certainty. It can be defined as:

$$shid(\mu)=a_A(x)/b_A(x) \tag{3}$$

If $shid(u)>1/3$, it is a strong hesitating potential function; if $shid(u)=1/3$, it is an average hesitating potential function; if $shid(u)<1/3$, it is a weak hesitating potential function.

Definition 4 (Set pair cloud connection number function). Based on the concept of cloud model [25], set pair cloud contact number function can be defined as:

$$\mu_g(x)=E_x(x)+E_n(x)i+H_e(x)j \tag{4}$$

In which $E_x(x)$, $E_n(x)$, $H_e(x)$ are the "expectation", "entropy", and "hyper entropy" of Gauss cloud model. $E_x(x)$ is the expectation value of spatial distribution of cloud droplets, which best represents the qualitative concept. Therefore, it is used to describe deterministic information in the Gauss set pair cloud connection number function. $E_n(x)$ is the entropy value of cloud model, which represents the uncertain measurement of qualitative concept. The bigger the value is, the more uncertain information it contains. $H_e(x)$ is the uncertain measurement of the entropy $E_n(x)$ (that is the entropy of entropy), which is determined by randomness and fuzziness, and reflects the cohesiveness of a numerical value belonging to a certain semantic value. The higher the hyper entropy value is, the greater dispersion degree the cloud has.

Definition 5 (Gauss set pair potential function and Gauss set pair hesitating potential function). Similar as definition 2 and definition 3, Gauss set pair potential function $shi_g(x)$ and Gauss set pair hesitating potential function $shid_g(x)$ can be defined as:

$$shi_g(x) = \frac{E_x(x)}{H_e(x)} \tag{5}$$

$$shid_g(x) = \frac{E_x(x)}{E_n(x)} \tag{6}$$

Definition 6 (Semantic distinction function). Suppose $\mu(x)$ is the membership value of the sample x belonging to Gauss cloud model, a Gauss set pair semantic distinction function can be defined as:

$$SD_g(x) = \begin{cases} \dfrac{E_x(x)\mu(x) + E_n(x)(1-\mu(x))}{E_x(x) + E_n(x)} & x > E_x \\ 1 & x = E_x \\ -\dfrac{E_x(x)\mu(x) + E_n(x)(1-\mu(x))}{E_x(x) + E_n(x)} & x < E_x \end{cases} \tag{7}$$

Semantic distinction function establishes a relationship between the determinate membership value $E_x(x)\mu(x)$ and the indeterminate membership value $E_n(x)(1-\mu(x))$ within a certain semantic, and highly aggregates the certain and uncertain information from integral aspect. Therefore, semantic distinction function is more reasonable and effective than membership function of fuzzy mathematics and the true/false membership of vague set theory in describing fuzzy semantics.

4. Privacy Preserving Data Publication based on Fuzzy Semantic Set Pair Cloud Model

Numerical data is the most common and important form in data publishing. The values of numerical attributes are usually continuous or discrete numbers (for example, “height”, “weight”, “age” and “income” *etc*.). If the publication of data does not consider privacy protection of this kind of sensitive attributes, attackers may get the accurate privacy information of individuals by “linking attacks” or “background-knowledge attacks”. In order to protect the privacy of numerical sensitive attributes and avoid reduction of data availability caused by excessive generalization on quasi identifiers, a fuzzy semantic set pair cloud model (FSSPCM) is proposed in this section.

FSSPCM method will transform numerical sensitive attributes into some corresponding fuzzy semantic forms. For example, attribute “income” can be described as {low, medium, high}; attribute “weight” can be represented by {light, medium, heavy}; attribute “age” can be divided into {juvenile, youth, middle-aged, the old} *etc*. Firstly, the set of fuzzy semantics X_i $(i = 1,2...n)$ can be determined according to the needs of practical situations, subjective experiences of data publishers or clustering analysis carried out on publishing data. Secondly, thresholds for different fuzzy semantics can be determined according to specific properties of numerical sensitive attributes, and combined with the time, place, and other factors related with data publication. The value of threshold is a reference point which can be adjusted according to actual needs of privacy. It reflects the randomness and uncertainty of information, and is beneficial to protect the privacy of user information. Then, Gauss cloud model [25] can be constructed for different semantic intervals, and the values of sensitive attributes can be transformed into some fuzzy semantic values according to certain principles (for example, the

principle of maximum membership degree). If user or data publisher have special needs of privacy on (parts of) sensitive attributes, the transformed fuzzy semantic value can be adjusted to adjacent or opposite semantic so as to realize personalized privacy protection. Figure 1 is the distribution diagram of the Gauss cloud model for attribute "age", where the range of attribute "age" is divided into three semantic intervals: {young people, middle-aged people, old people}. The user who is 53 years old (marked with circle) can either be published in "middle-aged" interval or be released into "old" interval according to the maximum membership degree or special need of privacy. Finally, in order to improve the availability of sensitive attributes after fuzzy semantic transformation and distinguish different data within the same fuzzy semantic interval, semantic distinctions for all the tuples can be calculated according to formula (7) and published together with fuzzy semantic values. When doing so, expectation value of the Gauss cloud model $E_X(x)$ can be replaced by the threshold $Thresh(X_i)$ of a certain fuzzy semantic interval; membership value $\mu(x)$ for a sample x can use the specific value of Gauss cloud model on corresponding interval; and entropy value $E_n(x)$ for Gauss cloud model can be obtained through the reverse Gauss cloud algorithm [25].

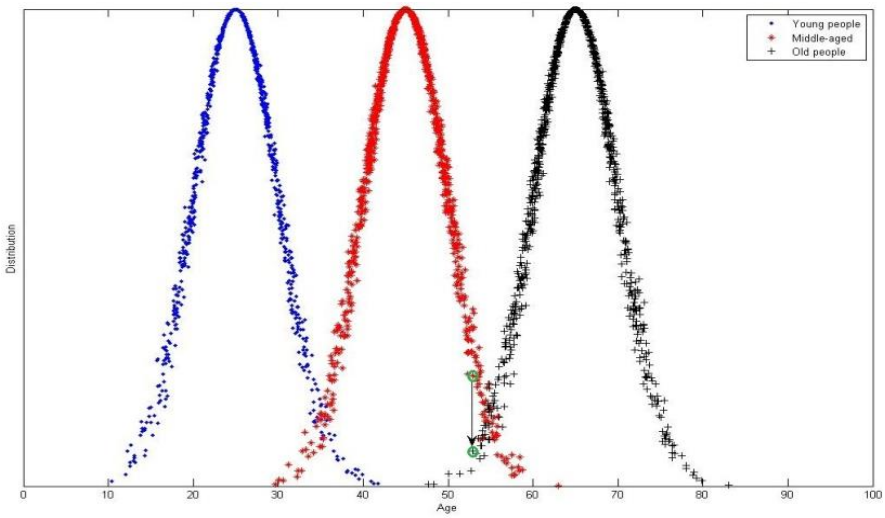


Figure 1. Gauss Cloud Model with Three Semantic Intervals

Algorithm 1. *FSSPCM—N*

1. *T = numerical sensitive attributes in original data*
2. *X = the set of fuzzy semantics*
3. *Thresh = threshold values for different fuzzy semantics*
4. *N = the number of tuples in T*
5. *for i = 1: N*
6. *if there is no special needs for privacy*
7. *T_i ← determine fuzzy semantic value for tuple i according to maximum membership principle*
8. *T_i ← calculate semantic distinction for tuple i*
9. *elsif there has some special needs for privacy*
10. *adjust T_i to adjacent or opposite fuzzy semantic*
11. *T_i ← calculate semantic distinction for tuple i*
12. *end*
13. *end*
14. *numerical sensitive attributes in the published data ← T**

Example 1. Suppose the original data to be published are shown in Table 1, in which attribute "income" is the numerical sensitive attribute related to user's personal privacy. Select a set of fuzzy semantics with three values $\{low,\ medium,\ high\}$ and set threshold values to be $\{2000,4000,6000\}$ according to the actual situation of the region. According to *FSSPCM-N* algorithm, original data can be transformed into fuzzy semantic values by using the principle of maximum membership degree and published together with their semantic distinctions (shown in Table 2).

In the published results (shown in Table 2), identity attribute “name” has been replaced by some numbers in order to protect privacy and values of sensitive attribute "income" have been published into some fuzzy semantic values. What’s more, different tuples within the same fuzzy semantic interval have been distinguished by their semantic distinctions. For example, user NO.102, NO.103, NO.104 and NO.108 have the same fuzzy semantic value “medium”, but there are big differences between their semantic distinctions. The distance from the value of real income to the threshold of interval for user NO.102 and user NO.103 are equal, so semantic distinctions for the two users are of the same. But the semantic distinction for user NO.102 is positive, indicating that the actual value is greater than the threshold of this interval, while the semantic distinction for user NO.103 is negative, meaning that the actual value is less than the threshold. The real value of income for user NO.102 is closer to the threshold value than that of the user NO.104, so the value of semantic distinction for user NO.102 is greater than that of the user NO.104. Theoretically, the above situations are consistent with the realistic logic. Published data after such transformation process are able to well reflect the nature meaning and variation trends of the original information, and maintain good availability of sensitive data without perverse phenomenon, which improves the effectiveness, availability and portability of data largely.

Table 1. Original Data (partial)

Name	Age	Sex	Zip	Income
Alice	46	M	110030	5500
David	62	M	130010	4200
Bob	30	F	621020	3800
Susan	32	M	540000	4700
Kara	48	M	110050	7000
Jane	73	F	130030	2800
Leo	36	M	240050	6200
Linda	53	F	621030	3200
Mole	66	F	540010	2500

Table 2. Published Result

NO.	Age	Sex	Zip	Income	
				Fuzzy result	Distinctions
101	46	M	110030	high	-0.4631
102	62	M	130010	medium	0.8177
103	30	F	621020	medium	-0.8177
104	32	M	540000	medium	0.2643
105	48	M	110050	high	0.1010
106	73	F	130030	low	0.1761
107	36	M	240050	high	0.8346
108	53	F	621030	medium	-0.1971
109	66	F	540010	low	0.4625

5. Fuzzy Semantic Transformation Method for Categorical Sensitive Attributes

Fuzzy semantic transformation of sensitive attribute is also one kind of generalization in some certain sense, which also uses broad range of values to instead of the accurate values. Therefore, this kind of fuzzy semantic transformation can not only be applied on numerical sensitive attributes, but also appropriate for categorical sensitive attributes. The value of categorical sensitive attributes (for example: gender, marital status, education background, religious faith, *etc*.) are usually finite and discrete. In order to achieve fuzzy semantic transformation for categorical sensitive attributes, semantic generalization tree has been constructed based on fuzzy semantic theory and combined with the concept of attributes generalization in K-anonymity method [5].

Definition 7 (Semantic generalization tree). Suppose the finite domain for categorical sensitive attribute $A_i \in A^{SA}$ is D_i, if there are some sequential (non-overlapping) semantic divisions $U_1, U_2 ... U_n$ on D_i, the mapping relationship from D_i to its semantic division U_j forms a tree, called semantic generalization tree SGT_{A_i}.

Within a semantic generalization tree SGT_{A_i}, leaf nodes are composed by the value D_i of categorical sensitive attribute A_i; non-leaf nodes are formed by semantic values summarized from its sub-tree nodes. The entire semantic generalization tree has the following characteristics: (1) The value of any intermediate node in the semantic generalization tree is a semantic generalization of a sub-tree with itself as the root node; (2) The root node of the semantic generalization tree is a semantic generalization of all the nodes within it.

Definition 8 (Generalization degree). In a semantic generalization tree SGT_{A_i}, the number of generalization layers from any of the node v_i to any of its ancestor node v_j (including its father node) is called the generalization degree from v_i to v_j, referred as $GD(v_i, v_j)$.

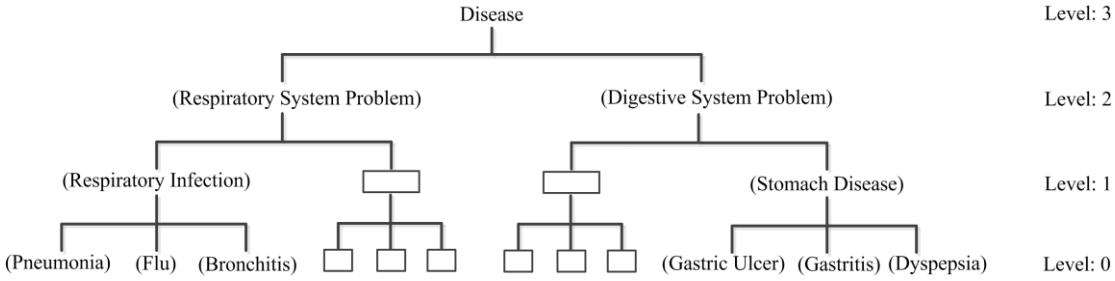


Figure 2. An Example of Semantic Generalization Tree (Fragment)

Figure 2 is an example of semantic generalization tree for the categorical sensitive attribute "disease", in which the generalization degree from node "gastritis" to "stomach disease" is 1, and the generalization degree from node "gastric" to "digestive system problem" is 2. The fuzzy semantic transformation mentioned above is able to achieve personalized privacy preserving according to user's different needs of privacy. For example, when user has a low requirement of privacy protection, generalization degree of the sensitive attributes can be smaller; while larger generalization degree should be selected when the user needs a stronger privacy protection.

In order to provide some standards to evaluate availability of categorical sensitive attributes after fuzzy semantic transformation, loss of information and reserve degree of information are defined in the following part. Intuitively, the reserved degree of information about categorical sensitive attribute after fuzzy semantic transformation is not only related with the generalization degree but also correlated with the number of sub-tree nodes where it located in. It is obviously, the greater the generalization degree is, the broader the described semantic is, the more the detailed description information lost. Generalization of a sub-tree uses a small number of semantic nodes to represent all, therefore, the more the children nodes, the larger the information lost after generalization.

Definition 9 (Loss of information). Use $N\_SGT_{A_i}$ to represent the number of leaf nodes of semantic generalization tree SGT_{A_i}, $N\_subT(v^*)$ is the number of leaf nodes of a sub-tree with the root node v^*, then the amount of information that has been lost during the generalization from v to v^* can be expressed as $ILoss(v^*)$.

$$ILoss(v^*) = \frac{N\_subT(v^*) - 1}{N\_SGT_{A_i}} \tag{8}$$

Definition 10 (Reserve degree). For a semantic generalization tree SGT_{A_i}, the total layers of semantic generalization for sensitive attribute $A_i \in A^{SA}$ is L, $L(v^*)$ represents the generalization layer of node v^* and $ILoss(v^*)$ is the amount of information that has been lost during the generalization from v to v^*. Reserved information after generalization can be expressed as $RD(v^*)$.

$$RD(v^*) = (1 - \frac{L(v^*)}{L})(1 - ILoss(v^*)) \tag{9}$$

In the semantic generalization tree shown in Figure 2, if the node "gastritis" has been generalized into "stomach disease", information loss after generalization will be $ILoss(v^*) = \frac{3-1}{12} = \frac{1}{6}$, and the reserved information after generalization is $RD(v^*) = (1 - \frac{1}{3})(1 - \frac{1}{6}) = \frac{5}{9}$.

Algorithm 2. *FSSPCM—C*

1. *T = categorical sensitive attributes in the original data*
2. *SGA = the semantic generalization tree*
3. *GD = the generalization degree*
4. *N = the number of tuples in T*
5. *for i = 1 to N*
6. *if GDi = 1 (there is no special needs for privacy)*
7. *Ti ← generalize tuple i for 1 level according to semantic generalization tree*
8. *Ti ← calculate reserved information for tuple i*
9. *elsif GDi > 1 (there has some special needs for privacy)*
10. *Ti ← generalize tuple i for GDi levels according to semantic generalization tree*
11. *Ti ←calculate reserved information for tuple i*
12. *end*
13. *end*
14. *categorical sensitive attributes in the published data ← T**

Table 3. Original Data (partial)

Name	Age	Sex	Zip	Disease
Alice	46	M	110030	Flu
David	62	M	130010	Bronchitis
Bob	30	F	621020	Gastritis
Susan	32	M	540000	Dyspepsia
Kara	48	M	110050	Pneumonia
Jane	73	F	130030	Gastric Ulcer
Leo	36	M	240050	Flu
Linda	53	F	621030	Pneumonia
Mole	66	F	540010	Bronchitis

Example 2. Suppose the data to be published is shown in Table 3, in which "disease" is a categorical sensitive attribute related to user's personal privacy. Use the semantic generalization tree shown in Figure 2 and carry out the fuzzy semantic transformation algorithm *FSSPCM-C* on the sensitive attribute "disease", the final released data are shown in Table 4 (according to actual privacy needs of users, different generalization degrees have been applied on different users). Although user NO.101 and user NO.107 have the same value of sensitive attribute, their generalization degree is different, therefore, the reserved information after generalization is quite different (the same situation happens on user NO.102 and user NO.109). Sensitive degree of "disease" for user NO.105 is higher than others and it has been totally protected, so it has the lowest reserve degree. The information of "disease" for user NO.106 has been completely released without generalization; therefore, reserve degree for this user has the maximum value.

Table 4. Published Results

NO.	Age	Sex	Zip	Disease		
				Fuzzy semantic values	Reserve degree (generalization degree)	
101	46	M	110030	Respiratory Infection	5/9	(1)
102	62	M	130010	Respiratory system Problem	7/36	(2)
103	30	F	621020	Stomach Disease	5/9	(1)
104	32	M	540000	Digestive System Problem	7/36	(2)
105	48	M	110050	Disease	0	(3)
106	73	F	130030	Gastric Ulcer	1	(0)
107	36	M	240050	Respiratory system Problem	7/36	(2)
108	53	F	621030	Respiratory system Problem	7/36	(2)
109	66	F	540010	Respiratory Infection	5/9	(1)

Example 3. A semantic generalization tree for a categorical sensitive attribute is shown in Figure 3, which contains 25 leaf nodes and 4 levels of generalization operation. In Figure 3, attribute x and z are located in different levels of semantic generalization tree, if both of them has been generalized for one layer, the amount of information that has been lost during the generalization from x to A_1 can be expressed as $ILoss(A_1) = \frac{3-1}{25} = \frac{2}{25}$, the reserved information after generalization can be expressed as $RD(A_1) = (1-\frac{1}{4})(1-\frac{2}{25}) = \frac{69}{100}$. While for the generalization from z t A_3, the values will be $ILoss(A_3) = \frac{3-1}{25} = \frac{2}{25}$ and $RD(A_3) = (1-\frac{2}{4})(1-\frac{2}{25}) = \frac{46}{100}$. Comparatively speaking, A_3 has a higher level of generalization than A_1, and the fuzzy semantic value expressed by A_3 is also more broadly, so the amount of information retained after generalization is fewer. Attribute x and y located in the same level of semantic generalization tree, if attribute x has been generalized for one layer, the amount of information that has been lost during the generalization and the reserved information after generalization are just like the former case. While for attribute y, it has been generalized for two layers into A_2, so the amount of information that has been lost during the generalization from y to A_2 can be expressed as $ILoss(A_2) = \frac{5-1}{25} = \frac{4}{25}$, and the reserved information after generalization can be expressed as $RD(A_2) = (1-\frac{2}{4})(1-\frac{4}{25}) = \frac{42}{100}$. Comparatively speaking, A_2 has a higher level of generalization than A_1, and the fuzzy semantic value expressed by A_2 is also more broadly, so the amount of information retained after generalization is smaller.

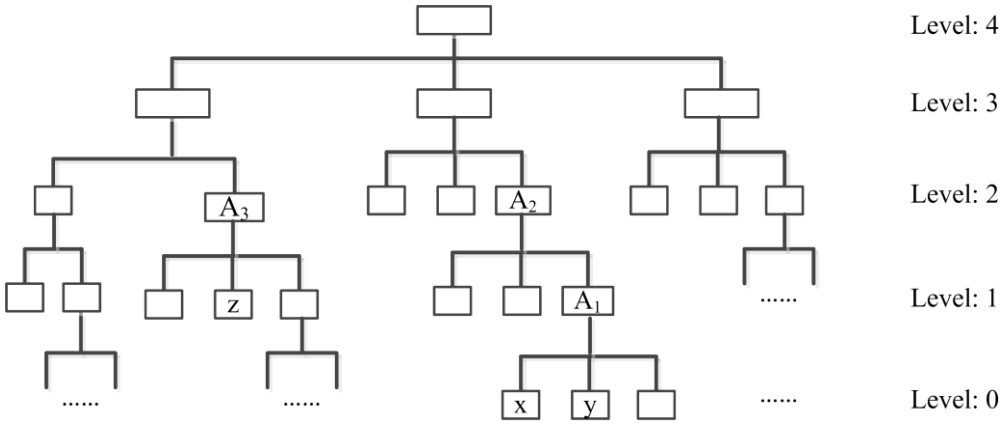


Figure 3. Semantic Generalization Tree for Example 3

6. Experimental Results

In this section, we will carry out some testing and analysis on the proposed fuzzy semantic transformation method and compared it with other previous fuzzy methods from the aspects of privacy preserving effect, availability of data after privacy protection and execution time of the algorithm. Experiments in this paper have been carried out on Intel (R) Core (TM) i3-2120, 3.3GHz, 4.00GB hardware platforms, and programed by MATLAB software. Testing data has been selected from datasets Adult (16808 samples), Page Blocks Classification Data Set (5473samples), Pima Indians Diabetes (768 samples) from UCI (University of California at Irvine) machine learning database [28].

6.1. Privacy Preserving Effect

As stated in the beginning, K-anonymity is the most popular approach for privacy preserving data publishing, in which data is classified into different equivalent classes and each class has a set of k-records indistinguishable from each other on sensitive attributes. Therefore, individual's privacy has been protected in some extent. However, K-anonymity method cannot resist background knowledge attack and homogeneous attack and may amplify the computational effort to some infeasible levels.

In this paper, the problem of privacy preserving data publishing was addressed by transforming the exact value of sensitive attributes into fuzzy semantic values. It can effectively prevent "linking attack" in the following aspects: (1) Division of the fuzzy semantic intervals depends on the subjective intuition fuzziness of data publishers and the needs of privacy protection, and can realize personal privacy protection according to different privacy protection needs of different users. If the user has lower needs of privacy protection, sensitive information can be transformed into appropriate semantic intervals accurately; while if the user has higher needs of privacy protection, fuzzy semantic value can be expanded to a larger degree or even transformed into different semantic intervals. (2) Threshold values for different fuzzy semantic intervals are determined by actual conditions and privacy needs with fuzziness and randomness. Changing of the threshold values will cause the variation of semantic distinctions, so as to dynamically adjust the degree of data privacy protection. (3) Values of sensitive attributes before and after transformation had good correlations, and this kind of correlation was not set up based on numerical mapping functions. Therefore, attackers cannot get more privacy information through reverse deduction. When attackers get some part of the quasi identifier

information of certain individual, they can only get a qualitative description of the sensitive attributes, but not exact value. This kind of qualitative description can be fuzzier or even wrong by strengthening the level of privacy protection. Therefore, it is an effective way to deal with the background knowledge attack.

6.2. Data Availability

Compared with K-anonymity privacy protection methods based on generalization of QI attributes, fuzzy semantic transformation method cuts off the connection between a certain individual and the sensitive information. To some extent, fuzzy semantic transformation is also a kind of K-anonymity, which contains a lot of users with the same values on the sensitive attributes. Therefore, users cannot be distinguished from each other. In order to achieve this kind of effects, generalization-based K-anonymity methods need to carry out a lot of generalization operations. There has a problem that the information of original data has been lost because the selected set of QI attributes may be too large. In some extreme cases, all tuples are generalized into one QI-group, which seriously affects the availability of anonymous data. While the fuzzy semantic transformation method proposed in this paper has fully retained the data besides the sensitive attributes, which not only achieved privacy protection but also maintained the original information of data at the maximum degree in the meanwhile.

Table 5. Availability Comparison of Sensitive Attributes

Sensitive Attribute	Methods		
	Method in [18]	Method in [19]	FSSPCM
Numerical SA	$\mu(x)$	*offset*	$SD_g(x)$
Categorical SA	PL、DL	not applicable	$RD(v^*)$

Table 6. Comparison of Distinction Effect on Sensitive Attribute

NO.	Age	Sex	Zip	Income					
				Method in [18]		Method in [19]		FSSPCM-N	
				Fuzzy result	$\mu(x)$	Fuzzy result	*offset*	Fuzzy result	$F_g(x)$
101	46	M	110030	high	0.5000	high	-0.5000	high	-0.4631
102	62	M	130010	medium	0.6333	medium	-0.3667	medium	0.8177
103	30	F	621020	medium	0.4222	medium	-0.5778	medium	-0.8177
104	32	M	540000	medium	0.9667	medium	-0.7837	medium	0.2643
105	48	M	110050	high	1.0000	high	0.0000	high	0.1010
106	73	F	130030	low	0.8667	low	0.1333	low	0.1761
107	36	M	240050	high	0.6444	high	-0.3556	high	0.8346
108	53	F	621030	medium	0.6889	medium	-0.8647	medium	-0.1971
109	66	F	540010	low	1.0000	low	0.0000	low	0.4625

Availability comparison of sensitive attributes between the proposed method and some other fuzzy methods is shown in Table 5. Reference [18] uses triangular membership function to divide semantic intervals for numerical sensitive attributes, and the value of membership function $\mu(x)$ is directly used to measure the approximate extent between the original data and transformed result. The value of membership function only reflects the degree that a sensitive attribute is

subordinated to a fuzzy semantic interval, but cannot distinguish the size of data after fuzzy semantic transformation. For categorical sensitive attributes, reference [18] uses additional parameters privacy level (PL) and disclosure level (DL) to tell whether the data is to be released or not. If the user does not mind revealing the data, DL is set to "T" and the ancestor of the sensitive attribute will be returned as a response to query on the tuple. While the user does not want others to associate the data with himself, DL is set to "F" and the attribute value itself is returned in response to any query on that tuple. But there has no quantitative indicators to describe the publication result for categorical sensitive information in reference [18]. Reference [19] proposed the concept of fuzzy offset ($offset = \pm(1-\mu(x))$) on the basis of reference [18], which can further distinguish the relationship between original information and the transformed results. However, it can only be applied on numerical sensitive data and there is no consideration of the categorical sensitive information. In order to improve the availability of sensitive information after fuzzy semantic transformation, semantic distinction (SD) and reserve degree (RD) are designed for numerical and categorical sensitive attributes in this paper. Table 6 shows the transformation results of different methods on numerical sensitive attributes "income" shown in Table 1. To be fair, all the selected methods use the same fuzzy semantic intervals. Compared with other fuzzy-based methods, the proposed algorithm has a better distinguish effect for different information with the same semantic values. What's more, it can not only be used to protect numerical sensitive attributes but also suitable for the work of categorical sensitive attributes.

6.3. Clustering Effects

FSSPCM method proposed in this paper has fully retained the availability of data except for sensitive attributes. Tuples have similar (or even different) values on sensitive attributes are published into the same fuzzy semantic interval, which achieved the same effect of privacy protection compared with K-anonymity methods from another way. Besides, FSSPCM method generated the values of semantic distinction (SD) or reserve degree (RD) during the process of data publishing, which can effectively maintain the availability of transformed information for data mining and other subsequent applications.

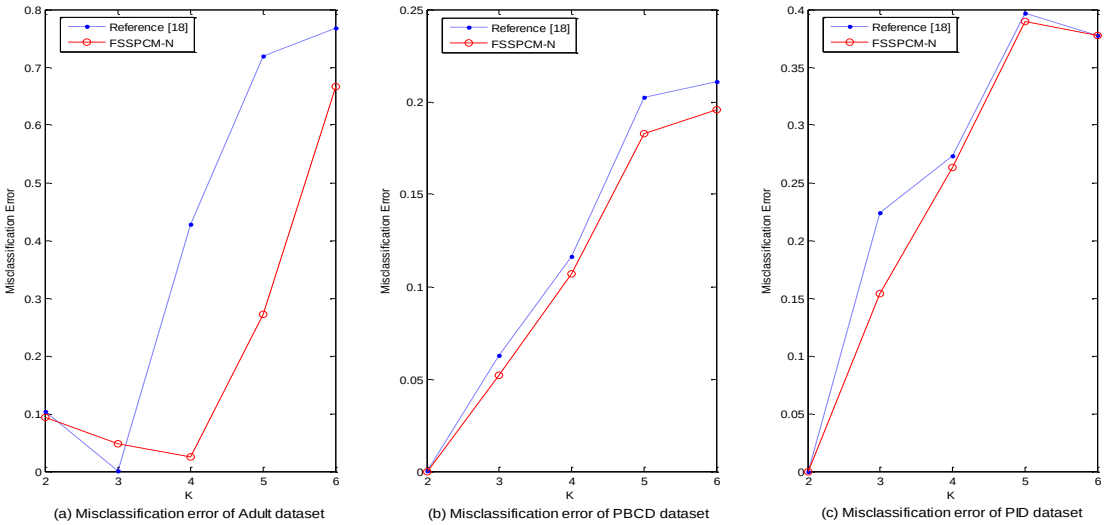


Figure 4. Misclassification Error After Fuzzy Semantic Transformation

To prove the above conclusions, we select attribute "age" from Adult dataset, attribute "area" from PBCD dataset and attribute “diastolic blood pressure” from PID dataset to be the sensitive attributes, and use k-mean clustering method to carry out clustering analysis on the sensitive information before and after fuzzy transformation. Figure 4 shows the misclassification error results on different datasets. Misclassification error is defined as:

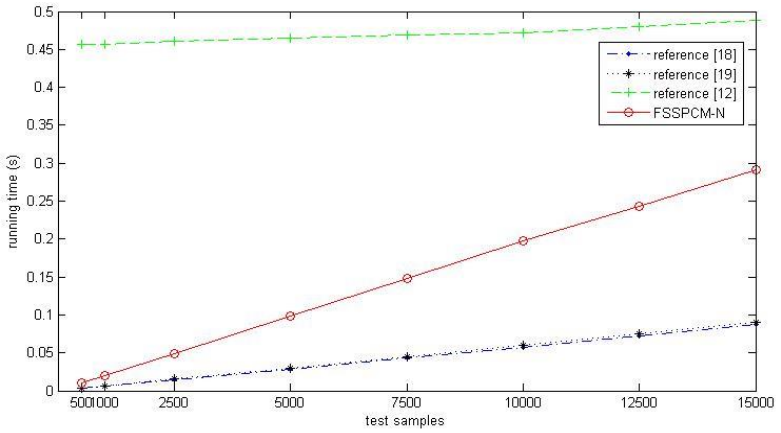
$$ME_i = \frac{1}{N} \times \sum_{j=1}^{k} (\left| Cluster_j(T_i) \right| - \left| Cluster_j(T_1) \right|) \quad (10)$$

In formula (10), N is the total number of the testing sample, cluster number k equal to the number of fuzzy semantic intervals. Use k-mean clustering results of the original attributes to be the standard (T1), and compare the misclassification error of the proposed method with the method proposed in reference [18]. (Actually, reference [19] uses the same membership function as reference [18], their clustering results are of the same).

As it can be seen from Figure 4, the proposed method has lower misclassification error than the method used in reference [18] on both three testing datasets, which means the data mining results calculated from the data after fuzzy semantic transformation will be closer to the results carried out on original data. Therefore, the proposed fuzzy semantic transformation method preserved better availability for sensitive attributes.

6.4. Execution Time

Take the transformation process of numerical sensitive attributes for example, execution time of the proposed method has been compared with the generalization based K-anonymous algorithm [12], fuzzy conversion method based on the membership function [18] and fuzzy conversion method based on offset [19]. Choose attribute “age” in Adult dataset to be the sensitive information, randomly and non-redundant select 500, 1000, 2500, 5000, 7500, 10000, 12500 and 15000 samples to carry out 8 groups of experiments. For each group, the experiment was repeated 10 times to get a final average result. Comparison results of execution time carried out on different algorithms is shown in Figure 5. To be fair, the K-anonymous algorithm only generalized one quasi identifier, all of the fuzzy methods use the same fuzzy semantic intervals, and the method in reference [18] and [19] use the same membership function.



It is obviously in Figure 5, the fuzzy method in reference [18] and [19] have lower operation cost. That is mainly because the transformation methods used in these references are linear operations and have lower computational complexity. Execution time of the fuzzy semantic transformation method proposed in this paper

is slightly larger than the above two methods, but far superior than the generalization based K-anonymity algorithm. What's more, with the increasing amount of data and number of quasi identifiers, the advantage in execution time of the proposed method will appear gradually.

7. Conclusion

Data publishing without proper protection method will directly lead to the disclosure of personal privacy. Meanwhile, analysis and mining technology carried out on the released data also bring some threats to the privacy of data. Therefore, how to publish and analyze data without disclosing private information has become the main purpose of privacy protection technology. The paper addresses the problem of privacy preserving data publishing by transforming sensitive attributes into fuzzy semantic values and put forward different transformation algorithm for both numerical and categorical sensitive attributes based on fuzzy semantic set pair cloud model. Construction methods of the proposed fuzzy semantic intervals and semantic generalization tree are accorded with objective facts and subjective logic, and will avoid excessive loss of original information caused by generalization operations in K-anonymity methods. Semantic distinction (SD) and reserve degree (RD) designed for numerical sensitive attribute and categorical sensitive attribute achieved good privacy protection effects and reflected consistency and availability of the transformed information and the original data. Experiments and analysis showed that the proposed method had good availability and operational efficiency, and there was no anonymous failure problem during data updating and was also suitable for data publishing with multiple sensitive attributes.

Acknowledgements

This work is supported by National Nature Science Foundation of China (NO.61363078).

References

[1] http://anquan.baidu.com/bbs/thread-189113-1-1.html
[2] Man accused of stalking ex-girlfriend with GPS. http://www.foxnews.com/story/2004/09/04
[3] P. Phillips and I. Lee, "Crime analysis through spatial areal aggregated density patterns", J. Geoinformatica, vol.15, no.1, **(2011)**, pp. 49-74.
[4] L. Sweeney, "K-Anonymity: A model for protecting privacy. International Journal of Uncertainty", J. Fuzziness and Knowledge-Based Systems, vol.10, no.5, **(2002)**, pp. 557-570.
[5] L. Sweeney L. "Achieving k-anonymity privacy protection using generalization and suppression", J. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol.10, no.5, **(2002)**, pp. 571-588.
[6] Machanavajjhala, D. Kifer and J. Gehrke,"L-diversity:Privacy beyond k-anonymity", Proceedings of 22nd International Conference on Data Engineering, Atlanta,Georgia,USA, **(2007)**.
[7] N. Li, T. Li and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity". Proceedings of IEEE 23rd International Conference on Data Engineering, Istanbul, Turkey, **(2007)**.
[8] D. Rebollo-Monedero, J. Forne and J. Domingo-Ferrer, "From t-Closeness-Like Privacy to Postrandomization via Information Theory", J. IEEE Transactions on Knowledge and Data Engineering, vol.22, no.11, **(2010)**, pp. 1623-1636.
[9] F. Mokbel, Y. Chow and G. Aref, "Casper*: query processing for location services without compromising privacy", J. ACM Trans on Database Systems, vol.34, no.4, **(2009)**, pp. 24-48.
[10] Y. Chow, F. Mokbel and X. Liu, "Spatial cloaking for anonymous location-based services in mobile peer-to-peer environments", J. Geoinformation, vol.15, no.2, **(2011)** ,pp. 351-380.
[11] Meyerson, R. Williams, "On the complexity of optimal k-anonymity", Proceedings of the 23rd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, Paris, France, **(2004)** June.
[12] G. Aggarwal, R. Panigrahy, T. Feder, D. Thomas and K. Kenthapadi, "Achieving Anonymity via Clustering", Proceedings of the 25th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, Chicago, Illinois, USA, **(2006)** June.

[13] X. Xiao, Y. Tao, "Anatomy: Sample and effective protect preservation", Proceedings of the 32nd International Conference on VLDB, **(2006) .**
[14] V. Rastogi, S. Hong, and D. Suciu, "The boundary between privacy and utility in data publishing", Proceedings of the 33rd International Conference on VLDB, Vienna, Austria, **(2007)**.
[15] S. Oliveira, O. Zaiane, "Privacy preserving clustering by data transformation", Proceedings of the 18th Brazilian Symposium on Databases, October, Manaus, Amazonas, Brazil, **(2003)**.
[16] S. Oliveira, O. Zaiane, "Achieving privacy preservation when sharing data for clustering", Proceedings of the International Workshop on Secure Data Management in a Connected World in Conjunction with VLDB, Toronto, Canada, **(2004)** August.
[17] Pedro M. Wightman, Miguel A. Jimeno and Daladier Jabba,"Matlock: A location obfuscation technique for accuracy-restricted applications", IEEE Wireless Communications and Networking Conference, **(2012),** April.
[18] VV Kumari, SS Rao, K. Raju, KV Ramana and B. Avadhani,"Fuzzy based approach for privacy preserving publication of data", J. International Journal of Computer Science and Network Security, vol.8, no.1, **(2008)** ,pp. 115-121.
[19] G. Zhang, J. Yin, "Privacy preserving using fuzzy sets. Computer Engineering and Applications", J. Computer Engineering & Applications, vol.46, no.28, **(2010)** ,pp. 118-121.
[20] M. Sridhar, B. Raveendra Babu, "A Fuzzy Approach for Privacy Preserving in Data Mining", J. International Journal of Computer Applications, vol.57, no.18, **(2012)** ,pp. 1-5.
[21] Wang, Eric Ke, Ye, Yunming, "A Fuzzy-Based Context-Aware Privacy Preserving Scheme for Mobile Computing Services", Proceedings of international conference on soft computing techniques and engineering application, Kunming, China, **(2013)** September.
[22] Bhuyan, HK, Kamila, NK, "Privacy preserving sub-feature selection based on fuzzy probabilities",
[23] J. Cluster computing, vol.17, no.4, **(2014)** ,pp. 1383-1399.
[24] T. Jahan,G. Narasimha and CVG Rao, "A Comparative Study of Data Perturbation Using Fuzzy Logic to Preserve Privacy", J. Networks and Communications, Vol. 284 of the series Lecture Notes in Electrical Engineering, **(2014)**,pp. 161-170.
[25] Syed Md. Tarique Ahmad, Shameemul Haque and SM Faizanut Tauhid, "A fuzzy based approach for privacy preserving clustering", J. International Journal of Scientific & Engineering Research, vol.5, no.2, **(2014)** ,pp. 1067-1071.
[26] Li Deyi, Du Yi, "Artificial Intelligence with Uncertainty (second edition)", National Defense Industry Press, Beijing, **(2014).**
[27] Zhao Ke-qin, "Set pair analysis and its preliminary application", Zhejiang science & technology press, Hangzhou, **(2000).**
[28] Wang Wanjun, Yan Yan, "Research and application of set pair analysis on uncertain information processing", Lanzhou University press, Lanzhou, **(2015).**
[29] http://archive.ics.uci.edu/ml/datasets.html

Authors

Yan Yan, she is born in 1980. She has been an assistant professor of Lanzhou University of Technology of Computer and Communication since 2015. Now she is a Ph. D. candidate of Lanzhou University of Technology in electrical and information engineering. Her main research interests include information hiding and privacy preservation technology.

Xiaohong Hao, he is born in 1960. He has been a professor and Ph.D. supervisor of Lanzhou University of Technology since 2000. Director of Chinese Association for Artificial Intelligence. His main research interests include complex system theory, intelligent control, computer network and computer control technology.

Wanjun Wang, he is born in 1974. He has been an assistant professor of Lanzhou University of Arts and Science of Information Engineering since 2010. Member of China Computer Federation. His main research interests include privacy preservation and intelligent information processing technology