Personalized Privacy-Preserving Granular Computing Model

Yanguang Shen¹, Yonghong Liu¹, and Meiye Zhang²
¹ School of Information Science and Electrical Engineering
Hebei University of Engineering, Handan, Hebei, 056038, China
Email: shenyanguang@yahoo.com.cn, liulyh2003@sohu.com
² School of Humanities Hebei University of Engineering, Handan, China
Email: zmy_7420987@126.com

Abstract—As a new computing model, Granular computing provides a new efficient way for solving complicated problems, massive data mining, and fuzzy information processing. Privacy is becoming an increasingly important issue in many data mining applications. In this paper, we combined the existing model of granular computing with personalized privacy-preserving demand, and proposed a new granular computing model, which is called personalized privacy-preserving granular computing model. We also proved that the new model can make individual privacy preserving more rational, improve the accuracy of the individual privacy preserving.

Index Terms—granular computing; data mining; privacy-preserving

I. INTRODUCTION

In 1979, Zadeh, a famous mathematician of America, firstly put forward the fuzzy information granulating based on fuzzy set theory. It promoted fuzzy logic theory and its application [1].

In 1996, professor T.T.Lin and Zadeh proposed granular computing, and they thought it was the best general description for solving problems with granular computing method. Nowadays Granular computing becomes a hot researching area, it brings so much convenience for us to observe and analyze problems.

It is the rapid development of information technology and the massive growth of information storage, which not only promoted the rapid development of data mining and knowledge discovery, but also caused some new problems of privacy protection. In order to reduce private data leakages in the process of data distribution, people usually use data perturbation, adding random noise, aggregation, and anonymity methods to deal with raw data. The anonymity method became more important to protect the privacy leakage in current days. Traditional anonymity methods usually only use system parameters or expert decision-making parameters, not those decided by personalized decision-making.

In this paper, we analyzed the basic granular computing models, applied the basic idea of granular computing in privacy preserving data mining. We used the personalized decision-making as the precondition of granularity, proposed personalized privacy preserving granular computing model. It has been theoretically proved that we can get more rational anonymous protection and improved the accuracy of privacy protection by this model.

II. GRANULAR COMPUTING

There are three components in Granular computing: granule, granular hierarchy, and granular structure. Granule is the basic element of the Granular computing. Granularity is the conception in measuring the size of the granule. Granular hierarchy is a hierarchical structure, which can follow granular rules based on practical requirements. It is the most important target for us to solve problems in different granular hierarchies, and the different granular hierarchies can be transformed each other. Granular structure is a net structure with all granular hierarchies connected with each other. To some extent, the complexity of granular hierarchy decides the complex degree of solving problems.

Granular computing has two basic problems: granulating and granular computing. Granulating is a constructive process of solving problems. It is mainly involved in granular rules, granular algorithm (method), granule and granular structure (description), as well as the problems of the qualitative (quantitative) description about the granule and the granular structure. It directly decides the success of granular computing.

III. GRANULAR COMPUTING MODEL

A. Computing With Words

Computing with words is a computing and reasoning model or method through using words instead of numbers. It solves the problems such as the limitation of the numerical representation of membership function in fuzzy set theory, the lack of relationship between the front and the rear in the expression of concept, the complexity of logical expression and implementation in computing, and etc. it can make them more suitable for human thoughts. Fuzzy logic plays a central role in the computing with words model. A granule is defined as: \( G = (X | X \text{ isr } R) \), \( X \) is the variable for restraining relation in domain \( U \), isr is a variable relational operator, the \( r \) in it is a discrete variable, \( r \) decides a way that \( R \) restrains \( X \). The restrain methods are equivalence, probability, fuzzy, and etc.
B. Rough Set

Rough Set Model, which was put forward by Pawlak of Poland [2] in 1982, is a tool for characterizing the incompleteness and the uncertainty. Its main usage is to solve approximate problems. In the knowledge of the domain U, we give an equivalent relation R, x ∈ U, as = (U, R) is an approximate space, [x]_R = \{y | x R y\} is the equivalent class containing x, all equivalent classes of R constitute a division of U, X/R is all the equivalent classes of R derived from U. Each equivalent class, known as equivalent granule, is the granule made up by indistinguishable element. The two objects are equivalent if their set-values will be equal. Then we discuss how to use the basic knowledge to express a general concept X. For those sets unable expressed by the union set of sets in (X, R), we use the lower approximation and the upper approximation to express: R^-(X)={x∈U | [x]_R⊆X} and R^+(X)={x∈U | [x]_R∩X≠Φ}.

C. Quotient Space Theory

Professor Bo Zhang and Ling Zhang put forward the theory of quotient space. The main content of quotient space theory Model includes the description of the quotient space problem, the hierarchical structure, the decomposition and synthesis of the quotient space, granular computing of the quotient space, the relative reasoning of granular space and the heuristic searching of problems [5].

Quotient space model can be expressed as three element set, as (X, F, T). X is the domain, F is the attribute set, T is a topology on X. When getting the rough granule, that is, give an equivalence relation R, then get R corresponds to the quotient set (as [X]). It corresponds to \{[X], [F], [T]\} that was called with the quotient space R. Professor Bo Zhang and Ling Zhang also gave the essential characters of this model: "the principle of false security," "the principle of fidelity," "weak principle of false security" and "fuzzy quotient space theory."

D. Other Granular Computing Model

Computing with words, rough set and the quotient space are the three main basic models of granular computing model. There are so many models based on these three models, such as, granular computing model based on partition [4], granular computing model based on covering [5], granular computing model based on tolerance relation [6], granular computing model based on the concept lattice [7] [8], and granular computing model based on biological genetics [9] etc.

IV. THE RELATIONSHIP OF THE GRANULAR COMPUTING MODEL

Computing with words based on fuzzy set focused on discussing the expression of granularity. It express granularity with language and words. Then use the fuzzy logic for reasoning and calculation of the granular computing with words. This method is closely related to people’s subjective factors, and very effectively to deal with complex human systems.

Rough set is an object set described by multi value attribute. It divides objects by the differences attributes. Once the attributes is made, the partition of granularity will be certain, so that the upper approximation and the lower approximation can be formed, and get the corresponding rules. Granularity partition is a static process.

The important of quotient space theory is to find the most suitable quotient space in all possible quotient space, to observe the same problem in different quotient spaces, to get different perspectives of the same problem, and ultimately integrate them into the overall problem solution. Granularity partition is a dynamic process in solving practical problems with quotient space theory.

Granular computing based on the coverage model is a special case of theoretical model. It can be divided into connotation and extension of the concept knowledge. In extension, it uses intersection and merging computing of the set theory to synthesize or decompose granule, while at connotation, it uses the number of attributions to control granular size.

Granular computing based on the division of rough set model is a generalized model of rough set theory. They have the same process, such as to get the neighborhood systems by using some binary relationship, to achieve the transformation among the different granular hierarchies by using two computing operators the zooming-in and zooming-out.

In fact, granular computing based on tolerance relation model and granular computing based on compatible relation are the same model. They are all based on the reflexive and symmetric binary relationship, but have different emphases.

Granular computing model based on the concept lattice is a hierarchical structure model. The granules form a poset relation, by including relationship, and constitute a complete lattice. In the knowledge system, the transformation of the concept granule of information and the concept information granule are the foundations of conception. It has also provided acquisition methods for getting the concept of knowledge.

Granular computing based on bio-genetics model, which combined the neural network and genetic theory of knowledge acquisition, must have much better prospects in the intelligence acquisition.

A common characteristic of every granular computing model is to find the better approximate solution according to the demand of reality. Nowadays, granular computing applications should not be limited to image processing and wavelet analysis. Here we applied granular computing in privacy preserving data mining, and gave the granular computing model of privacy preserving as follows.

V. PRIVACY PRESERVING GRANULAR COMPUTING MODEL

We focused our work on K-anonymity method, an important method for protecting privacy and avoiding the link attack. After studied the existing k-anonymity
methods [10-16], we found that all the methods were basically not taking into account the fact that the same sensitive attribute may be have different privacy degree for different individuals, because that different individuals have different opinions and requirements. Anonymity can conceal all data by using the same method will cause two defects: privacy information with higher protection degree may be disclosed, while those with lower protection degree may be excessively concealed.

We proposed a new granularity model on k-anonymity of privacy preserving. Considering that there will be one or more sensitive attributes in the raw data set, we first applied the privacy preserving anonymous granularity model on the data set with one sensitive attribute, then on the multiple one.

A. privacy preserving granular computing model with one sensitive attribute

We defined the original data set as a big granular domain $U$, divided $U$ with rough set theory, here sensitive attribute $P_i$ is decided by individual $i$ in $U$, $P_i = +$ means the attribute value with high sensitive; $P_i = *$ means the attribute value with general or uncertain sensitive; and $P_i = !$ means the attribute value with no sensitivity. The distribution of the sensitive attribute values are shown in figure 1 and clustered into three data sets as a result and these values are clustered into three data sets as a result.

We define the set $POSP(U)$ as privacy preserving upper approximation domain that all the $X \in POSP(U)$ satisfy $P_i = +$; the set $BNDP(U)$ as privacy preserving boundary domain that all the $X \in BNDP(U)$ satisfy $P_i = *$; the set $NEGP(U)$ as privacy preserving negative domain that all the $X \in NEGP(U)$ satisfy $P_i = !$. After the granularity, we can get three granules: $POSP(U)$, $BNDP(U)$ and $NEGP(U)$, each granule is a data set. The data in $POSP(U)$ have the highest privacy preserving degree, in $BNDP(U)$ have the normal privacy preserving degree, and in $NEGP(U)$ have the lowest privacy preserving degree. Thus we can focus privacy preserving on $POSP(U)$ by using higher anonymity method, and use normal anonymity method on $BNDP(U)$. If the data in $POSP(U)$ don’t satisfy the highest anonymity method, we can add some data in $BNDP(U)$ to $POSP(U)$, so that all data in $POSP(U)$ can satisfy the highest anonymity. If the data in $BNDP(U)$ don’t satisfy the normal anonymity method, we can add some data in $NEGP(U)$ to $BNDP(U)$, so that all data in $BNDP(U)$ can satisfy the normal anonymity.

Theorem 1: Let $GNum$ as the number of all anonymous data in domain $U$, $GNum_{POSP(U)}$ as the number of anonymous data in $POSP(U)$, thus the inequality (1) is always true.

$$\frac{GNum_{POSP(U)}}{Num} \leq \frac{GNum}{Num} \leq 1.$$  (1)

Proof: Let the total of all elements in $U$ as $Num$, the total of all anonymous elements in $U \ GNum$, $GNum_{POSP(U)}, GNum_{BNDP(U)}, GNum_{NEGP(U)}$ as the total of all anonymous elements in $POSP(U), BNDP(U)$ and $NEGP(U)$, therefore, we have the following equality:

$$GNum = GNum_{POSP(U)} + GNum_{BNDP(U)} + GNum_{NEGP(U)}.$$  

So the result $GNum \leq Num$ must be true, that is to say (1) always be true.

B. privacy preserving granular computing model with multiple sensitive attribute

Most of the real-world data sets contain multiple sensitive attributes. Today more and more researchers carried out anonymity methods on multiple sensitive attributes [17] [18]. As the same reason, the anonymous methods are always not considered the different individual privacy degree of the same attribute.

We defined the original data set, which has m-sensitive attributes, as a multi-dimensional granular space $U'$. The dimension number is m, which is the total number of sensitive attributes. We used $S_i$ as the j-sensitive attribute, j-dimension as the corresponding dimension space of $U'$.

Sensitive attribute $QS_j$ is decided by individual $i$ in the j-sensitive attribute of $U'$, $QS_j = +$ means the attribute value with high sensitive; $QS_j = *$ means the attribute value with general or uncertain sensitive; and $QS_j = !$ means the attribute value with no sensitivity. The distribution of the multiple sensitive attributes values are shown in figure 2.

We will get a set of character strings after every individual gave his privacy degree of the corresponding sensitive attribute. For example, there are four sensitive attributes in the domain $U'$, the string “+$!*+” means that the individual, who chose it, considered that the value of the first sensitive attribution has the highest privacy degree, the value of the second sensitive attribution has normal privacy degree, the value of the third one has no privacy degree, and the value of the forth one also has the highest privacy degree, etc.

Then based on the granularity methods, we can approximately cluster the character strings into three granules: $POSQ(U')$, $BNDQ(U')$ and $NEGQ(U')$, each granule is a approximate data set. The data in $POSQ(U')$ have the highest multiple privacy preserving degree, in $BNDQ(U')$ have the normal multiple privacy preserving degree, and in $NEGQ(U')$ have the lowest multiple privacy preserving degree. Similarly, we focus privacy preserving on $POSQ(U')$ by using higher multiple anonymity method, and use normal multiple anonymity method on $BNDQ(U')$. If the data in $POSQ(U')$ don’t satisfy the highest multiple anonymity method, we can add some data in $BNDQ(U')$ to $POSQ(U')$, so that all data in $POSQ(U')$ can satisfy the highest multi-degree anonymity. If the data in $BNDQ(U')$ don’t satisfy the
normal multiple anonymity method, we can add some data in $NEGQ(U')$ to $BNDQ(U')$, so that all data in $BNDQ(U')$ can satisfy the normal multiple anonymity.

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we have analyzed the existing granular computing models, studied the k-anonymity method of privacy preserving, gave the personalized privacy preserving granular computing model. Because of the fact that every data set for publishing has not only one sensitive attribute but also multiple ones, we described the definitions and used the sample figures for the new model in the two aspects. This model made the individual privacy preserving more rational, improved the accuracy of the individual privacy preserving, and reduced the complexity of the individual privacy preserving. In the future we will apply it in more areas.

ACKNOWLEDGEMENT

This work was supported by the Science Research Plan Program “Research on the Privacy Preserving Classification Data Mining (2009421)” from the Hebei Education Department.

Figure 1. Granularity $U$ with one sensitive attribute clustered into $POSP(U)$, $BNDP(U)$ and $NEG(P(U)$ based on the personalized privacy preserving decision.

Figure 2. Granularity $U$ with multiple sensitive attribute clustered into $POSQ(U')$, $BNDQ(U')$ and $NEGQ(U')$ based on the personalized privacy preserving decision.
REFERENCES


