An Approach to Achieving K-partition for Preserving Privacy by Using Multi-constraint Anonymous Parameter Based on Rough Sets

Tao Liu
Department of Computer, Nanchang University, Nanchang,330031,China
liutao@ncu.edu.cn

Taorong Qiu, Wenying Duan, Xiaoming Bai
Department of Computer, Nanchang University, Nanchang,330031,China
ncq1303201@163.com

***Abstract*—We propose an approach to achieving different *K*-partition for preserving privacy by using the multi-constraint anonymous parameter design method based on the attribute significance of rough set, in order to reduce the imbalance phenomenon between the privacy protection and data availability caused by adopting the same anonymous intensity. In this approach, taking into account the significance of quasi-identifier attributes, we carry out the dimension division automatically and obtain multi-constraint anonymous parameters. After that an anonymous algorithm is executed on the separate partition. Experimental results show that the proposed method can obtain a better balance between the privacy protection degree and data availability.**

***Index Terms*—Privacy Preservation; Anonymous Parameters; Rough Set; Attribute Significance**

I. INTRODUCTION

With the rapid development of information technology, it is available to collect and release relevant data for every walk of life, which may involve some confidential personal privacy information. An emerging problem is that it makes invasions of personal privacy possible if we issue the crude data directly. Consequently, it is of great significant to deliberate how to release actual and valid data without prejudice to the personal privacy. The original method, which is called anonymous, is to delete the unique attribute identified the specify tuple in order to protect personal privacy. In literature [1], Sweeney proposed a K-anonymization rule to solve privacy leakage caused by link attack for the first time. The shortcoming is that for the sensitive data K-anonymization does not offer the constraint treatment. For this reason, literature [2] advanced a new rule l-diversity to increase the diversity of sensitive attribute, and thus reducing the risk of privacy leakage. However, it does not solve the problem of losing vast quantities of information. On the other hand, the l-diversity model shows a little weak in similar attack. Literature [3] introduced a t-closeness anonymous rule to resist the similar attack. Literature [4] and literature [5] discussed the rule of (a,k)-anonymization and p-sensitive respectively, both of which having the similar idea with l-diversity. All above can be implemented by generalization, and the list extended to (c,k)-safety [6], privacy skyline[7] and so on. However employing generalization methods may largely reduce the data precision and availability. Literature [8] argued an exchange method to realize anonymous, a rule of high accuracy of data release which uses a lossy link method. In the meantime, (k,e)-anoymity[9] is also a typical anonymous model using the exchanging method.

The traditional algorithms are presented by a single constraint to process K-partition with supposing all the quasi-identifier attributes having the same significance. But when they are applied to the high dimensional data set, which make a lot of useful information loss. Literature [10] advanced a multi-constraint rule to fit various constraint conditions. Though it's well to balance the privacy protection with data availability, how to set the constraint parameters for every constraint set is still not specified. The multi-constraint anonymous parameter design method based on the attribute significance of rough set proposed in the present paper is more concerned with the distinct influences of separate quasi-identifier attributes on sensitive attributes, i.e. the importance degree of attributes are different. It tries to divide the attribute dimension of quasi-identifiers automatically according to the characters of the data set to be released (the distinction of different quasi-identifiers significance) without the prior knowledge, and designs the corresponding multi-constraint anonymous parameters. The results suggest that the approach proposed in this paper can achieve a better balance between the privacy protection and data availability.

Corresponding Author: Tao Liu, Tel.:+86 0791 83969684

© 2014 ACADEMY PUBLISHER
doi:10.4304/jcp.9.10.2493-2499

II. Relevant Concepts

A. Rough Set Preliminaries

Rough set is a mathematical approach to study imprecise, uncertain information[11-12,14],which is mainly used to mine patterns and regulations from an incomplete data set. It has been widely applied in medical diagnosis, processing control, information retrieval, industry manufacture, commercial economy, and so on, in twenty years. Because of existing data redundancy and missing values, the data to be protected can be viewed as an incomplete information system in the privacy protection of data issue. Instead, there is extraordinary superiority for rough set in attributes reduction. So, rough set can be used to deal with the imbalance problem in the privacy protection.

Definition 1 (Information System) Let $S=(U,A)$ be an information system, where U is a non-empty set of finite objects(the universe) and A is a non-empty finite set of attributes. Such that for every $a\in A$, V_a is the set of values that attribute a may take values. A can also be denoted as $A=C\cup D$, where C is the set of conditional attributes and D is the set of decision attributes. A decision table is an information system including both the condition attributes and decision attributes.

Definition 2 Let R be an equivalent relation, and $X\subseteq U$ be a certain subset of U, the lower approximation can be defined as:

$$\underline{R}(X)=\bigcup\{Y\in U/R:Y\subseteq X\} \tag{1}$$

And the upper approximation is :

$$\overline{R}(X)=\bigcup\{Y\in U/R:Y\cap X\neq\Phi\} \tag{2}$$

Definition 3 Supposing $X\subseteq U$, the *R-positive region* can be denoted as:

$$POS_R(X)=\underline{R}(X) \tag{3}$$

Correspondingly *R-negative region* is*:*

$$NEG_R(X)=U-\overline{R}(X) \tag{4}$$

Definition 4 Let $T=(U,A)$ be a decision table, C is the set of condition attributes and D is the set of decision attributes, $C\rightarrow_k D$ indicates that D depends on C in a degree $k(0<=k<=1)$, where

$$k=card(POS_c(D))/card(U) \tag{5}$$

And $POS_C(D)$ is called the *C-positive region* of D.

Definition 5 Suppose $T=(U,A)$ be a decision table, $C=(c_1,...,c_m)$ be the set of condition attributes and $D=(d_1,...,d_k)$ be the set of decision attributes, $U/(C\cup D)=\{Y_1,Y_2,\ldots,Y_k\}$ and $U/C=\{X_1,...,X_l\}$ denote the partition of U on the attribute set C and $C\cup D$, respectively. The conditional information entropy of the decision attribute set D on the set of condition attributes C can explicitly be written as

$$I(D|C)=\left|\sum_{i=1}^{k}\frac{|Y_i|}{|U|}\log_2|Y_i|-\sum_{i=1}^{l}\frac{|X_i|}{|U|}\log_2|X_i|\right| \tag{6}$$

Definition 6 Let C and D be sets of condition and decision attributes respectively of the decision table $T=(U,A)$, the significance of attribute $c\in C$ can be defined as follows:

$$NewSig(c)=I(D/(C-\{c\}))-I(D/C)+I(D/\{c\}) \tag{7}$$

B. Multi-constraint Rule

Let $T=(U,QI,SA)$ be an original data table, which is an essential decision table, among which, QI is the quasi-identifier attribute set, C, D and SA are the condition attribute set, decision attribute set and sensitive attribute set respectively. T is called the data set to be released. In $T=(U,QI,SA)$, C is the anonymous constraint, noted as $C=<QI,k>$ [10], where $QI=\{attr_1,...,attr_m\}$, QI is the quasi-identifier of C, composed of a group of attributes $attr_i(1\le i\le m)$. K is the anonymous parameter of C. This shows that there are K equivalent tuples on QI. The multi-constraint rule can be formulated as below:

Definition7[10](Constraint Set) Let $CSet=(C_1,C_2,...,C_n)$ denote a constraint set, in which $C_i=\langle QI_i,K_i\rangle,1\le i\le n$ is the ith constraint and n is the number of constraints in $CSet$, noted as $|CSet|$.

III. Intelligent Selection Of Multi-Constraint Anonymous Parameter Based on Attributes Significance

A. A Model of K-value Selection Based on The Attribute Significance

For a data table T to be released, K is the anonymous parameter. T' is the anonymous data table. After the equivalence partitioning, if the cardinality of every equivalence class is no less than K, then we say that T' is satisfied with K - anonymous.

K takes different values associated with various attribute significances. The relationship with the definition is as follows:

Definition 8 (Relation between attribute significance and anonymous parameters) Given two arbitrary attributes of a data table T, $attr_i(1\le i\le m)$ and $attr_j(1\le j\le m)$, $i\neq j$. $kattr_i$ and $kattr_j$ denote the attribute significances, respectively. Suppose K_i and K_j be the anonymous parameters of $attr_i$ and $attr_j$, then the

© 2014 ACADEMY PUBLISHER

relation between K_i and K_j can be defined as follows: $K_i / K_j = kattr_j / kattr_i$.

Definition 9(Average Attributes Signification of Quasi-identifier) Let T be a data table with m quasi-identifier attributes, and $k_{set} = \{k_{attr_1},...,k_{attr_m}\}$ be the set of attribute significances, the average attribute significance of quasi-identifiers can be calculated by:

$$\overline{KSet} = \sum_{i=1}^{m} K_{attr_i} \Big/ m$$

Proposition 1: Let K be the anonymous parameter given to be carried on K-anonymous on data table T, and K' be the corresponding anonymous parameter computed by $\overline{KSet}$, then $K' = K$.

The proof is as below:

Each attribute in Quasi-identifier(QI) is deemed the same significance in the general *K*-anonymous algorithm. On that condition, suppose T be a data table with m attributes, the weight of arbitrary attribute $attr_i (1 \le i \le m)$ is

$$\omega_i = \frac{K_{attr_i}}{\sum_{k=1}^{m} K_{attr_k}} = \frac{1}{m}$$

Whereas, after obtaining all quasi-identifiers' attribute significances using rough set theory, the weight to the average significance can be figured out by

$$\overline{\omega} = \frac{\frac{\sum_{k=1}^{m} K_{attr_k}}{m}}{\sum_{k=1}^{m} K_{attr_k}} = \frac{1}{m}.$$

Obviously, $\omega_i = \overline{\omega}$ and according to Definition 8, $K' = K$ holds.

That ends.

B. Intelligent Partition Algorithm of Constraint Set

Algorithm Description

Algorithm 1: Intelligent partition algorithm of a constraint set

Suppose T is a data table to be released, $QISet = \{Attr_1,...,Attr_m\}$ denotes the set of quasi-identifiers whose significance is $K_{attr_i} (1 \le i \le m)$.

Input: A data table T with n items and m quasi-identifiers.

Output: the constraint set $CSet = \{C_1, C_2,..., C_h\}$.

Step 1: Compute the significance of every quasi-identifier in line with Definition 6 denoted as $KSet = \{K_{attr_1}, K_{attr_2},..., K_{attr_m}\}$.

Step 2: Set the threshold ε and divide $KSet$

$kSet = \{subkSet_1, subkSet_2,...subkSet_h\}$,

and get the partition set of the quasi-identifier

$QISet' = \{subQISet_1, subQISet_2,..., subQISet_h\}$.

Step 3: Calculate the average attribute significance of every subset and put them into set

$$\overline{kSet} = \{\overline{Subkset_1}, \overline{Subkset_2},..., \overline{Subkset_h}\}$$

Step 4: Obtain the set of anonymous parameters $\{K_1, K_2,..., K_h\}$.

Step 5: Output $CSet$.

C. Case Study

Take Table 1 for instance, the average attribute significance of quasi-identifier is $\overline{k_{set}} = \frac{\sum_{i-1}^{5} k_{attr_i}}{5} = 0.452$ and in the first subset $\overline{k_{set1}} = \frac{k_{attr_1} + k_{attr_2}}{2} = 0.09$.

$\overline{k_{set2}} = 0.5$ and $\overline{k_{set3}} = 0.79$ can be similarly obtained. On the basis of Proposition 1, there are $K_1 = (\overline{k_{set}} / \overline{k_{set1}})K'$, $K_2 = (\overline{k_{set}} / \overline{k_{set2}})K'$ and $K_3 = (\overline{k_{set}} / \overline{k_{set3}})K'$ obtained, respectively (K_1, K_2 and K_3 are rounded)

With those steps, we get the partition of the constraint set:

$$c1 = \left\langle \{attr_1, attr_2\}, \left(\overline{k_{set}} / \overline{k_{set1}}\right)K' \right\rangle,$$

$$c2 = \left\langle \{attr_3\}, \left(\overline{k_{set}} / \overline{k_{set2}}\right)K' \right\rangle$$

and $c3 = \left\langle \{attr_4, attr_5\}, \left(\overline{k_{set}} / \overline{k_{set3}}\right)K' \right\rangle$.

It is clearly depicted in Fig.1.

The anonymous parameter of each subset is subjected to that corresponding to the average significance of QI.

© 2014 ACADEMY PUBLISHER

TABLE I.
DATA TABLE T TO BE RELEASED

ID	Name	Physical Quality	Velocity	Passing the ball	Dribbling	Shooting	Social Status
1	Meysey	General	Very Fast	Admirable	Admirable	Admirable	High
2	Cristiano Ronaldo	Good	Very Fast	Good	Admirable	Admirable	High
3	Pato	General	Very Fast	Poor	General	Good	Medium
4	Heskey	Good	Slow	Very Poor	Poor	Poor	Low
5	Barlow Terley	Good	General	Good	Good	Good	High
6	Van Persie	General	General	Good	Good	Good	High
7	Ibra	Admirable	General	Admirable	Good	Good	High
8	Falcao	Good	Fast	General	General	Admirable	High
9	Mesut Ozil	General	Fast	Admirable	Good	General	High
10	Gao Lin	General	General	Very Poor	Poor	Very Poor	Low
11	Beretta	Good	General	Poor	Poor	Poor	Low
12	Pazzini	General	General	Poor	General	Good	Medium
13	Carole	Admirable	Slow	Poor	Poor	General	Low
14	Aguero	General	Fast	Good	Good	General	High
15	Crouch	General	Slow	Poor	Poor	General	Low
16	Huntelaar	General	General	Poor	General	Admirable	Medium

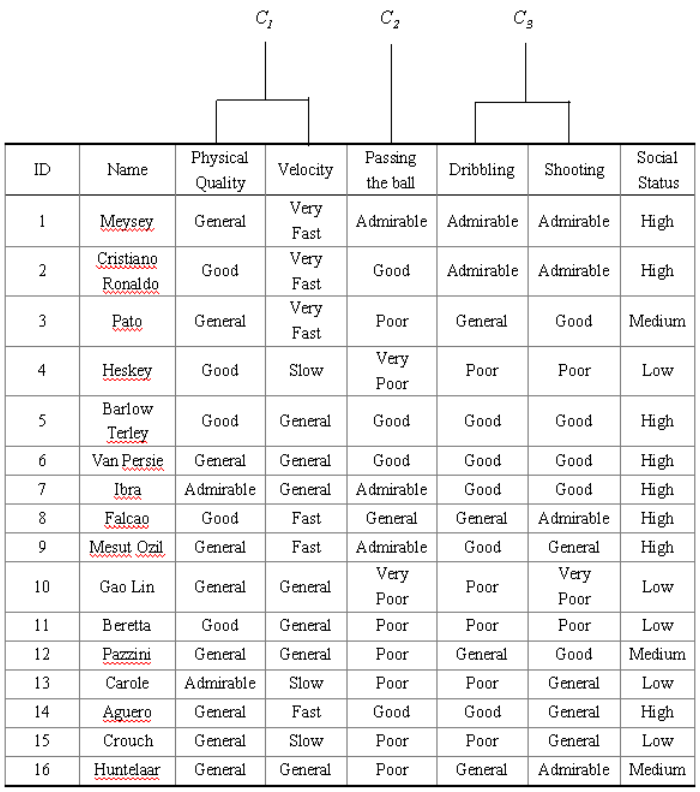

Figure 1. Constraint set partition of Table1

© 2014 ACADEMY PUBLISHER

D. Anonymous Method Based on Multi-constraint

On the basis of the above analysis, an anonymous method Based on multi-constraint is advanced, the process is described as following:

Algorithm 2: Multi-constraint anonymous algorithm.

Input: A data table T to be released

Output: A deliverable table T'

Step1: Obtain the constraint set in line with Algorithm 1.

Step2: Divide table T according to the constraint set and obtain the subsets $T_1, T_2, ... T_h$.

Step3: Perform the anonymization referring to $CSet = \{C_1, C_2, ..., C_h\}$.

Step4: Combine the subset in accordance with ID to generate all tuples, and output T'.

IV. Experiments And Results Comparison

The experiments are performed on a loaded Intel Core i3M380 2.53Ghz, 2G of DDR3 Memory, Windows 7(32 bits), and Matlab 7.0. With the intention of testing our algorithm, an adult data set, the information from U.S. census, generated by UCI is employed. We randomly take 1000 samples from the total 45222 items as test data.

Risk Evaluation Parameters: A tuple-linked method based on distance is adopted to evaluate the anonymity degree:

$$DLD = \frac{linked\_record\_num}{total\_record\_num}.$$

where, $linked\_record\_num$ is the number of records linking successful as well as $total\_record\_num$ is the number of total items in the anonymous table.

Data Availability Evaluation:For the data table with the classified variables,

$$\beta = (\sum_{i=1}^{n}\sum_{j=1}^{m}\alpha_{b_j}(x_i))(-\sum_{k=1}^{g}\frac{|X_k|}{n}\log_2^{\frac{|X_k|}{n}})/(m\times n)$$

is used to evaluate the data availability. Where $\alpha_b(x) = \frac{|\underline{x_b}|}{|\overline{x_b}|}, b \in QI, x \in U$ denotes the approximate precision of the item $x \in U$ on the attribute $b \in QI$, and $\underline{x_b}$ expresses the low-approximation. And correspondingly, $\overline{x_b}$ is the upper-approximation. $X_k \in U/QI = \{X_1, X_2, ...X_g\}$ forms a partition of U. Meanwhile n is the record number of T' and m denotes the sum of attributes.

For the table with the continuous variables, we can calculate the data availability according to literature[13].

In this chapter, concentrating on the Adult set we compare the anonymous method realized by the privacy protection rule of the intelligent constraint set partition (namely, the proposed algorithm) with the micro-aggregation algorithm based on traditional *K*-anonymous rules(denoted by MDAV algorithm).

In view of the features of the Adult set, the following constraint sets are obtained based upon the attribute significance of rough set:

$$C_1 = \{ < education, marital-status >, K_1 \},$$

$$C_2 = \{ < occupation, relationship >, K_2 \},$$

$$C_3 = \{ < native-country, work\ class >, K_3 \},$$

where $K_1 : K_2 : K_3 = 5:3:1$.(Rounding approximation ratio)

A. Leakage Analysis

In our experiment, evaluations on the tuple-linked(DLD) is used for leakage analysis. The risk evaluations with regarding to the proposed and MDVA algorithms are summarized in Table2.

TABLE II.
RISK EVALUATIONS OF ADULT ANONYMOUS TABLE

K	MDVA	The Proposed Algorithm
3	0.9920	0.9990
4	0.9620	0.9780
5	0.9410	0.9590
6	0.9140	0.9310
7	0.9380	0.9320
8	0.9090	0.9230
9	0.9150	0.9040
10	0.8850	0.8950

As it is demonstrated in Table 2, though the leakages are some more on the same value of K, it is not by much. The MDVA method divides the tuples into several equivalence classes while our approach is to partition on separate subset. Therefore, the deliverable table does not satisfy *K*-anonymous condition and at the same time the distortion is to a less degree. Because of the closer distance of every tuple, it increases the probability of linking to the original tuple.

B. Analysis of Data Availability

Table 3 reveals the data availabilities of Adult data set after adopting the two anonymous methods respectively.

© 2014 ACADEMY PUBLISHER

TABLE III.
THE LOSS OF THE ADULT SET

K	MDVA	The Proposed Algorithm
3	0.9509	0.7925
4	1.3323	0.8533
5	1.8666	1.3201
6	2.5334	1.8467
7	2.7182	2.0223
8	2.9507	2.5602
9	3.1022	2.6681
10	3.2308	2.9377

Experimental results show that the proposed method can reduce the dimension of primary data set, and makes an equivalence partitioning towards every subset in a lower dimension. So then, it provides a better protection for the data veracity. By contrast, in traditional MDAV algorithm, anonymous costs increase along with the number of quasi-identifier(dimension raised). Selecting a larger *K* value is an action taken to lower the chance of data leakage. For instance, when K takes 5, the data availability of the proposed method is better than that of MDAV and the leakage risk is lower as such.

V. CONCLUSION

Different attribute in quasi-identifier attributes has different influence on the sensitive attribute, therefore, it will generate unwanted information loss if we take the same anonymous parameters during the partition process. This paper aims to research and design a fresh anonymous rule for dimension division based on rough set. Owning to sum up the data on different levels, the method presented in this paper can produce diverse *K*-partitions on the different constraint subsets rather than on the whole table. The results indicate that the advanced method can keep a better balance between privacy protection and data availability. In future studies, it is necessary to perfect and optimize the algorithm, in the meantime comes up with the model of appraising data availability aiming at various attribute types.

ACKNOWLEDGMENT

Supported by National Natural Science Foundation of China(No. 61070139), Natural Science Foundation of Jiangxi Province(No. 20114BAB201039)and the Science and Technology Support Planning Project of JiangXi Province(No. 20112BBG70087, GJJ12148).

REFERENCES

[1] L. Sweeney. K-anonymity, " a model for protecting privacy". *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, Vol.10, pp. 557-570, 2002

[2] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity", In Proc. 22nd International Conference on Data Engineering (ICDE),Atlanta, Georgia, USA,pp. 24-35, 2006

[3] N. Li and T. Li, t-closeness: Privacy beyond k-anonymity and l-diversity,. In Proc. Of the 23rd International Conference on Data Engineering (ICDE),Istanbul, Turkey, pp. 106-115, 2007

[4] R C. Wong, L. I. J, A. W. Fu ,et al, (a,k) -Anonymity: an enhanced k-anonymity model for privacy preserving data publishing,. Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. New York: ACM.Press, pp.754 – 759, 2006

[5] T. M. Truta, V. Inayb. Privacy protection: p-Sensitive k-Anonymity property, Proceedings of the 22nd International Conference on Data Engineering Workshop s. Washington DC: IEEE Computer Society, pp. 94, 2006

[6] N. Koudas, D. Srivastava and T.Yu, et. al., Aggregate query answering on anonymized tables, In the 23rd International Conference on Data Engineering(ICDE'7).Istanbul, Turkey: IEEE Computer Society Press, 2007.

[7] B. C. Chen, K. LeFevre and R. Ramkrishnan, Privacy skyline:Privacy with multidimensional adversarial knowledge, In the 33rd International Conference on Very large Data Bases(VLDB'07).Vienna, Austria. pp.770-781, 2007

[8] X. Xiao and Y. Tao. Anatomy: Simple and effective privacy preservation, In Proc. of the 32nd Very Large Data Bases (VLDB) Conference, Seoul, Korea, pp. 139-150, 2006

[9] D. Martin, D. Kifer, and A. Machanavajjhala et. al, Worst-case background knowledge in privacy. In: the 23rd International Conference on Data Engineering(ICDE'7).Istanbul, Turkey:IEEE Computer Society Press, pp.116~125, 2007.

[10] X. C. Yang, X. Y. Liu, B. Wang, et. al, "K-Anonymization approaches for supporting multiple constraints"(in Chinese). *Journal of Software*, Vol. 17, pp. 1222-1231, 2006

[11] Q. Liu, *Rough Set and Rough Reasoning*.Beijing:Science Press,2005,Edition 3, pp.20-67

[12] W. X., Zhang, W. Z.,Wu, J. Y. Liang, et. al, *Rough Set theory and method*.Beijing:Science Press,2001,pp.138-154

[13] J. C. Chen, M. Jian, J. M. Han, "Evaluation model for quality of k-anonymity data oriented to microaggregation", (in Chinese) *Application Research of Computers,* Vol. 27, pp. 2345-2347, 2010

[14] T. R. Qiu, Y. Y. Lin, X. M. Bai, "The difference degree of condition attributes and its application in the reduction of attributes", *Journal of Computer*, Vol. 7, pp. 1087-1091, 2012

[15] G. J. Huang, S. L. Wang and X. G. Zhang, "Query expansion based on associated semantic space", *Journal of Computers*, Vol. 6, pp. 172-177, 2011

Tao Liu was born in 1978 on July 13, received the Master of Science (Internet Computing) from University of Abertay Dundee, Dundee, UK, in 2006. Currently, he serves as an lecturer with the Department of Computer, Nanchang University. He has been worked on artificial intelligence and machine vision more than 5 years, and has published three articles about these fields of research. Currently, his research interest covers artificial intelligence, machine vision and data mining.

© 2014 ACADEMY PUBLISHER

Taorong Qiu received the Ph.D degree in computer application technology from Beijing Jiaotong University, Beijing, China, in 2009. Currently, he serves as a professor with the Department of Computer, Nanchang University. His research interest covers rough sets, granular computing and knowledge discovery.

Wenying Duan received the Diploma in computer engineering degree from the Department of Computer Engineering at Nanchang University, China in 2010. Currently, he is a graduate student with the Department of Computer, at Nanchang University. His research interest includes rough set and modeling for prediction.

Xiaoming Bai received the Diploma in Engineering degree from the Department of Electrical Engineering at the Nanchang University, China in 1986. Currently, he serves as an associate professor with the Department of Computer, Nanchang University. His research interest covers computer networks, data mining.

© 2014 ACADEMY PUBLISHER