ISBN 978-952-5726-00-8 (Print), 978-952-5726-01-5 (CD-ROM)
Proceedings of the 2009 International Symposium on Web Information Systems and Applications (WISA'09)
Nanchang, P. R. China, May 22-24, 2009, pp. 136-139

A Novel Risk Recognition Method in Web Services

E Xu[1,2], Shao Liangshan[2], Sun Fuming[1], Cao Guanghui[1], Wu Hao[1]
[1] School of Electronic and Information Engineering, Liaoning University of Technology, Jinzhou 121001, P.R.China
[2] School of Resources and Environment Engineering, Liaoning Technical University, Fuxin 123000, P.R.China
Exu21@ieee.org

***Abstract*—To deal with the risk recognition from the complex Web environment , a novel risk recognition method was studied through attribute reduction and rule extration based on rough set. According to the indiscernible relation in rough set, discernible vector and its addition rule were defined. And meanwhile the core attribute set and the attribute reduction were obtained by scanning the information table just only one time depending on the discernible vector addition rule. Attribute value reduction was realized through gradually deleting the redundant attribute value for every rule in the information table by the correlation of condition attributes and decision attributes. Finally, a concise rule set for risk recognition was obtained. The illustration and experiment results indicate that the method is effective and efficient.**

***Index Terms*—web services; risk recognition; Rough Set; data mining**

I. INTRODUCTION

Risk management is activity directed towards the assessing, mitigating (to an acceptable level) and monitoring of risks. Risks can come from accidents, natural causes and disasters as well as deliberate attacks from an adversary. In some cases the acceptable risk may be near zero. The main ISO standards on risk management include. In businesses, risk management entails organized activity to manage uncertainty and threats and involves people following procedures and using tools in order to ensure conformance with risk-management policies. The strategies include transferring the risk to another party, avoiding the risk, reducing the negative effect of the risk, and accepting some or all of the consequences of a particular risk. Some traditional risk management programs are focused on risks stemming from physical or legal causes (e.g. natural disasters or fires, accidents, ergonomics, death and lawsuits). But the risk recognition is the first step of risk management, so there are a lot of construction methods of decision at present[1-4], especially ID3[5]proposed by Quinlan is the best-represented algorithm.The scholars put forward a number of optimization algorithms against id3 algorithm 's shortcoming which based on information gain, these optimization algorithms 's difference lies mainly in attribute choosing criterion, such as using gain rate, Gini index , statistical norm,CM standard, attribute importance and Decision-making support and so on, These algorithms have their own strengths and weaknesses, must be based on the characteristics of data to choose the appropriate algorithm. However, every ways will introduce some new problem, so a new risk recognition method based on rough set rule extraction was proposed in this paper.

II. BASIC CONCEPTION

Definition 1 In the decision system $S=(U,A,V,f)$, A can be divided into condition attributes C and decision-making D, In which $A=C\cup D$, $C\cap D=\emptyset$. If $X\in U$,the sample size |X|,and U|IND(D)={Y_1 ,Y_2 ,⋯ , Y_m }, and if xi is the the number of samples condition that element in X belonging Y_i,then information entropy is $H(X)= -\Sigma p_i\log_2(p_i)$, where $p_i= x_i / |X|$.

Definition 2 Condition information entropy: In the information system S,condition attribute set C,U|IND(C)= {X_1 , X_2 ,⋯ , X_n } the stratified information entropy is: E (C)=Σ－I(Xi), in fact, condition information entropy is the stratified information entropy $H(D\,|\,C) =-\sum p(X_i)\sum p(Y_j\,|X_i)\log_2(p(Y_j|X_i))$ where $p(X_i)=|X_i| / |U|$, $p(Y_j\,|X_i)=|X_i\cap Y_j\,| / |X_i|$, $i=1,2,\Box,n$, $j=1,2,\Box,m$.

Definition 3 Granularity entropy $H(C\mid D)$: $H(C\mid D)= -\sum p(Y_j)\sum p(X_i\,|Y_j)\log_2(p(X_i\mid Y_j))$, in which , $p(X_j)=|Y_j| / |U|$, $p(X_i\,|\,Y_j)=|X_i\cap Y_j\,| / |Y_j|$, $i=1,2,\Box,n$, $j=1,2,\Box,m$.

Definition 4 Assume $S=(U,A,V,f)$is a Decision Making System,in which $A=C\cup D$, C is condition attribute set , $D=\{d\}$is decision attribute set and $P\Box C$, then the importance of any attribute $a\in C/P$ can be defined as $SGF(a,P,D)$, $SGF(a,P,D)=H(D\mid P)-H(D\mid P\cup\{a\})$;especially if $P=\emptyset$, then $SGF(a,P,D)=H(D)-H(D\,|\,\{a\})$, is named as mutual information about attribute a and decision D noted as $I(a,D)$.

Definition 5 The rules of confidence, coverage and support: definite rule can be get by consistent decision tables, however the rule's uncertainty produced by inconsistent decision tables can be measured by degree of confidence, degree of coverage degree of support . Confidence level can also named as identity, credibility and uncertainty factor, and so on, which represented as $con(A\rightarrow B)=|X_i\cap Y_j\,| / |X_i|$, coverage degree represented as $cov(A\rightarrow B)=|X_i\cap Y_j\,| / |Y_j|$, support degree as $sup(A\rightarrow B)=|X_i\cap Y_j\,| / |U|$.

Confidence is used to describe the ownership problem that conditions equivalent class against decision equivalent class, that is the degree of decision rule's inconsistent. Coverage is used to evaluate the diverse situation of decision rules and escribe the degree of conditions equivalent class meeting decision equivalent class. Support degree describes representive extent of a rule in the whole decision set.

AP-PROC-CS-09CN001

III. DESCRIPTION OF THE ATTRIBUTE IMPORTANCE

A. Problem descirption

There are a lot of construction methods of decision tree at present, especially ID3 proposed by Quinlan is the best-represented algorithm. The scholars put forward a number of solution against id3 algorithm 's shortcoming of using more attribute in the process of classifying which based on information gain, C4.5 adjust this bias by using gain rate ,but it self tend to unbalanced split. Konenko et al believe it can make attribute with more values and attribute with less values fair treatment. However, the random combination of property value, to a large extent changed the information contained in the original knowledge system, leading to different result from the real value. Kaishe introduced a conception of user interesting degree which size can be decide by A priori knowledge or expertise in the field. Although this method can reduce the dependence on multi-value attribute and lower the selecting of no importance attribute to a certain extent, it would increase artificial factor and have a effect on the authenticity of the source of information systems. For instance, imortance attribute based on approximate classifying quality only take into account main doamin not border domain, hat is not taking into uncertainty situation. An example will be illustrated as below.

Suppose there is a two partition in the information system : $U|IND(A)=\{\{x_1, x_2, x_3, x_4\},\{x_5, x_6, x_7\},\{x_8, x_9, x_{10}\}\}$, $U|IND(B)=\{\{x_1, x_2, x_9\},\{x_3, x_4, x_6, x_8\},\{x_5, x_7, x_{10}\}\}$. And decision partition $U|IND(D)=\{\{x_1, x_2, x_5, x_8\},\{x_3, x_4, x_6, x_7, x_9, x_{10}\}\}$. Here the granularity size of partition A and B is the same ,that means the same thickness, so is approximate classification quality ,both of them equal zero. However the effect on classification is different between two. Therefore, merely examing attribute importance of main domain rule's definition has some defects which can't examine the overall situation of the rule.for this purpose, importance of attribute can be measured with the aid of information entropy. Because information entropy can describe uncertainty not only the consistent rule but the inconsistent rule, furthermore examine them together. Here is $H(D \mid A)$=0.951; $H(D \mid B)$ =0.875. This data's contrast just reflect overall ownership situation of partition A and B, which can't realize by approximate classification quality.This is because approximate classification quality only emphasis knowledge main domain but condition information entropy emphasis the overall, that is main domain and border domain. Entropy has the characteristics of symmetry; furthermore condition entropy should have narrow symmetry. For instance , information system S have different partition $U|IND(A)$ and $U|IND(B)$, and can get $H(D \mid A)=H(D \mid B)$. All these indicates that entropy and condition entropy illustrate certainty and uncertainty of the overall allocation situation. The definition of importance of attribute in paper [11] has the similarity to the definition 5, which has some defects in the process of construction decision .the main shortcoming take place in the constructing root node: $P=\emptyset$, $SGF(a,P,D)=H(D)-H(D \mid \{a\})$. In such a case, single attribute's effect on the decision rule should be emphasis, not the idea of getting nuclear by reducing. For attribute set C, finding attribute a which has the most influence on the rule set, constructing root node ,then delete a get $B=C / \{a\}$, then filter on the B, repeat this process untill completing constructing tree

B. Analysis of attribute

As can be seen from the definition 5, the indexes which use to measure decision rule set have such as: confidence of rule, Coverage and support. So we should examine importance of attribute from three aspects.

Definition 2 shows that $H(D \mid C)$ mainly examine the confidence of rule set , the process of it's quality .yet the variety of $H(D \mid C)$ value is caused by importable parts which include borde or introducing write conflict, however, the effect on the $H(D \mid C)$ is the combined action of compatible and incompatible as a whole .If the information granularity of A is smaller than B(A is more fine than B),then $H(D \mid A) < H(D \mid B)$。This shows that the more coarse of knowledge ,the more difficult sample set definition. Thus,confidence is examined by $H(D \mid C / \{a\}) - H(D \mid C)$: Attribute set C after removing attribute a caused the change of condition information entropy, namely the change of overall ascription situation of rule (the uncertainty of consistence rule and inconsistence rule). Removing different attribute "a" will produce various change, such change is the modification of main domain and border domain.the larger of the difference, the more importance of attribute "a" in the process of rule extraction.We can know from definition 3, $H(C \mid D)$ mainly examine the coverage of rule set, is its quality, is equivalence class of decision-making' measure for precondition of decision-making of decentralization situation, present rule's certain local randomness for equivalence class of decision-making.

The greater the coverage degree, the more the common with better concentration and fewer decentralization about all rules of Y_j. The smaller single condition granularity, the same to the coverage degree,all these will lead to poor decision-making effect of decision-making information system under decision-making equalivence unchange condition .It can be known by the condition information entropy property that $H(\{a\} \mid D)$ reflects condition attribute' relation to the decision-making attribute. The smaller the Granularity entropy $H(\{a\} \mid D)$, the more closer of the dependence relation between a and D,which will indicate that attribute 'a'is more important.About support degree measure, there exists such definition as below:

Definition 6 Decision-making entropy $H(C \to D)$: In the information system S,if decision-making attribute is divided as $U|IND(D)=\{Y_1, Y_2, \cdots, Y_m\}$, condition attribute as $U|IND(C)=\{X_1, X_2, \cdots, X_n\}$, then the decision-entropy of information system is : $H(C \to D)=-\sum\sum p_{ij}\log_2(p_{ij})$, in which $p_{ij}=|X_i \cap Y_j| / |U|$, $i=1,2,\cdots,n$, $j=1,2,\cdots,m$. Decision-making entropy is a quatization to the support degree , describing an randomness of rules to the whole information system, showing the extent of the representitive of the rule .The larger the support degree ,the stronger the rule's utility,the more important to the future prediction .Since $H(C\rightarrow D)=-\sum\sum p_{ij}\log_2(p_{ij})$, $\sum\sum p_{ij}=1$, then assumed $p_k=|X_i\cap Y_j| / |U|$, then $\sum p_k=1$, where $k=n\times m$. That means $H(C\rightarrow D)=-\sum p_k\log_2(p_k)$, $k=n\times m$ 。 therefore $H(C\rightarrow D)$ turns into information entropy (formular change), possessing the overall property of informtio entropy.If information granularity A is smaller than B (a is finer than b),then $H(A)>H(B)$, just reverses with condition entropy. If A and B represent decision-making attribute,then indicate the problem of sample ascription difficult .however ,here refer to condition attribute,the more coarse the knowledge,the more concetrate the rule set,the stronger the representitive. Hence , to measure support degree through examing $H(C\rightarrow D)-H(C/\{a\}\rightarrow D)$.Attribute set C except attribute a bring the channge of decision entropy ,bring whole rule set randomness change. the larger of the difference indicate the more importance of the process of the rule set extraction, the more critical of the rule set integration classification,in fact ,this also indirectly reflect knowledge coarse or fine together with the ascription problem of decision-making equivalence.

$H(D|C)$ is used to describe the sufficiency of the rule, $H(C|D)$is used to describe the necessity of the rule,but $H(C\rightarrow D)$ is used to describe the representitivity of the rule. The rule set with high support ,confidence and coverage is our need. In which , support degree and confidence degree are the main examing goals, together with different measure situation of their three,the definition of attribute as follow.however, the measure of coverage is used as subsidiary condition in the algorithm.

Definition 7: let $S=(U,A,V,f)$ be a decision-making system, in which $A=C\cup D$, C is the condition attribute set , $D=\{d\}$ is decision-making attribute set .Then $\forall a\in C$, the importance of set is : $\triangle H=H(D|C/\{a\})-H(D|C)$; $\triangle E=H(C\rightarrow D)-H(C/\{a\}\rightarrow D)$; in order to avoid sacrificing anothter measure for emphasis some importance of a measure ,the importance need to find a balance between them.that is , $SGF(a)=2\triangle H\times\triangle E/\triangle H+\triangle E$.

IV. ALGORITHM DESCRIPTION AND INSTANCE ANALYSIS

A. Algorithm Describtion

Preprocessing to decision-making table (all the attribute is discrete , for continute attribute value which has been discreted beforehand) is as below:

Input :decision-making table $S=(U, C\cup D,V,f)$.

Output:decision-making tree T.

Step1:initialize selected attribution $B=\emptyset$, $A=C$, $U^*=U$, T is null.

Step2: create node N.

Step3: if U belong to V,then return N as leaf node, marked with V.

Step4: if C is null, then return N as leaf node, marked N with the most equivalence in T.

Step5: for $a\in C$, a is attribute ,calculate $SGF(a)$。

Step6:select attribute a to make $SGF(a)$ maximum as root node or branch node.

(1) if many attribute make $SGF(a)$ maximum ,then select the attribute with the smallest granularity $H(\{a\}|D)$.

(2)if many attribute make granularity $H(\{a\}|D)$ smallest, then select the attribute with the smallest attribute value.

(3) if there exsist many attribute 'a' with the same number of attribute value, select the attribute with smallest sequence number.

(4) $B=B\cup\{a\}$.

Step7: if $B=A$ or U^* is classified overall, then go to Step10; Else go to Step8。

Step8: calculate $U|IND(a)=\{U_1,U_2,\square,U_t\}$, set up sub-table of decision-making. Set $S_i=(U_i,C^*\cup D,V,f)$, in which $C^*=C/\{a\}$, $i=1,2,\square,t$.

Step9: FOR each Si

{Let $U=U_i$, $C=C^*$;

Go to Step2, and then continue to split.

}

Step10: end the alogorithm, output decision tree T.

B. Illustrate and Analysis

Suppose there is a consistence discrete decision-making system S, where $C=\{a,b,c,e\}$, $D=\{d\}$, as Table 1.

TABLE I. DECISION-MAKING TABLE

U	a	b	c	e	d
x1	0	0	0	0	0
x2	0	0	0	1	0
x3	1	0	0	0	1
x4	2	1	0	0	1
x5	2	2	1	0	1
x6	2	2	1	1	0
x7	1	2	1	1	1
x8	0	1	0	0	0
x9	0	2	1	0	1
x10	2	1	1	0	1
x11	0	1	1	1	1
x12	1	1	0	1	1
x13	1	0	1	0	1
x14	2	1	0	1	0

Using the proposed algorithm to classify the above table:

(1) Create root node:

At this point , $SGF(a)=0.229$; $SGF(b)=SGF(c)=SGF(e)=0$。Thus, using a as root node.

(2) Partition U with a

U|IND(a)={{x_1 ,x_2,x_8,x_9 ,x_{11}},{x_3 ,x_7 ,x_{12} ,x_{13}},{x_4 ,x_5 ,x_6 ,x_{10} ,x_{14}}}. Three sub-table of decision-making can be seen as below:

TABLE II. DECISION TABLE S_1

U	b	c	e	d
x_1	0	0	0	0
x_2	0	0	1	0
x_8	1	0	0	0
x_9	2	1	0	1
x_{11}	1	1	1	1

TABLE III. DECISION TABLE S_2

U	b	c	e	d
X_4	1	0	0	1
X_5	2	1	0	1
x_6	2	1	1	0
X_{10}	1	1	0	1
x_{14}	1	0	1	0

TABLE IV. DECISION TABLE S_3

U	b	c	e	d
X_3	0	0	0	1
X_7	2	1	1	1
X_{12}	1	0	1	1
X_{13}	0	1	0	1

(3) For sub-table S3, its sample attribute to one class, thus, this sample set as leaf node. For sub-table S_1, exist *SGF(b) = SGF(c) = SGF(e)* =0 。Further examine granularity entropy *H({a} | {d})*：

H({b} | {d})= H({e} | {d})=0.951；*H({c} | {d})*=0。Therefore, *c* as branch node.

(4) Gets the result *U|IND(a)={{x1,x2,x8},{x9 ,x11 }}* after partition *U={x1,x2,x8,x9,x11}* with *c*, at this time, U_i is the same class. thus, get two leaf nodes.

(5) Operating with sub-table S_2 same to S_1, therefore, get the decision tree T as Figure 1.

As can be seen from above ,the decision-making tree achieved by the proposed method is similar to the paper [11]. Every reasonable rule can be get from root node to leaf node ,the overall tree correspond to a group of expression rule set.Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

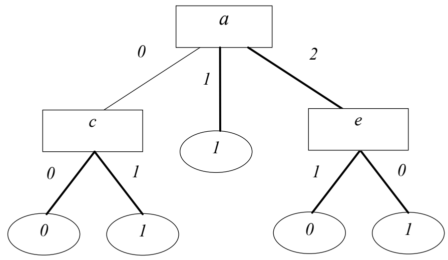

Figure 1. Decision Tree T

V. CONCLUSION

This paper proceed from rough set theory, combing with the character of information entropy ,propose a novel measure method of attribute importance ,which make up for the current system's shortcoming that other attribute importance and support degree can only examine main domain or only can measure certainty or uncertainty independently.

This paper also study rule measure standard and quantitative them and further propose a kind of rule extraction algorithm. This algorithm examine rule's every aspects, and make extraction rule reliable and effective. According the idea of rough set theory, this proposed algorithm avoid the step of attribute reduction ,and can deal with consistence decision-making table and inconsistence decision-making table. The example shows that this proposed algorithm is not only effective but easy to implementations. Yet, this algorithm is more sensitive to data noise .thus, it need further study about how to combine it with rough set of variable precision and probability

ACKNOWLEDGMENT

This work is supported by Chinese National Foundations (60674056，70771007), Liaoning Province Academic Talent Foundations(2005219001), Chinese National Foundations (70572070), Liaonig Creative Team Foundations (2008T090).

REFERENCES

[1] Krysikiewicz M, "Rough set approach to incomplete information system", Information Sciences, Vol.112, pp.39-49, 1998.

[2] Wang Guo-yin, Rough Set Theory and Knowledge Acquisition, Xi an communication university publisher, Xi an, China, 2000.

[3] E Xu, Gao Xudong. An algorithm for attributes reduction. In: Proceedings of the Fifth International Conference on Machine Leaning and Cybernetics. Dalian, China, August 2006:1438-1442.

[4] E Xu, Shao Liangshan, Ye Baiqing, Li Sheng. Algorithm for rule extraction based on rough set. Journal of Harbin Institute of Technology, Vol 14, SUP, January,2007, pp34-37.

[5] Wu Wei, Chen Weiqiang, Liu Bo.: Prediction of Ups and Downs of Stock Market by BP Neural Networks. Journal of Dalian University of Technology, Vol. 41, No.1. Dalian China (2001) pp9-15