Randomization Test for Importance Degree of Variables in Rough Set Theory

Dan Hu Xianchuan Yu
College of Information Science and Technology
Beijing Normal University
Beijing, China, 100875
hufengdd@163.com

Yuanfu Feng
Basic Courses Department
Beijing Union University
Beijing, China, 100101
longlonghd@163.com

Abstract

The appraisement of variable importance and contribution is the central problem for variable selection and relevance analysis, particularly in the domains of ecological and medical science. Except for statistical modelling, more interesting methods, such as rough set and artificial neural network, are used to analyze the variable contribution in systems. But the results derived from rough set and statistical theory can not be compared with each other because the lack of common descriptions. In this paper, we propose and demonstrate a randomization test for statistically assessing the variable importance degree in rough set theory. The randomization approach can identify variables that significantly contribute to the predictions of the system and reach a more objective result which can be compared with statistical analysis. Thus, the bridge of rough set theory and statistical approach is constructed. Furthermore, by the randomization test, the interaction of variables can be easily appraised and the variables which have the same importance degrees can be distinguished. At last, an experiment shows the function of randomization test for importance degree of variables in rough set theory.

1. Introduction

One of the primary goals of ecological and medical sciences is to establish and appraise correlative links between variables and predications[1]. Statistical analysis is the traditional approach to determine the input variable importance and significance. But as we all know, there are some shortcomings of statistical approaches, especially the suppositions of model and variables. Many results are holden based on the strict suppositions and will be suspected when some of the suppositions are not satisfied. On the other hand, there are lots of new ways have been put forward for intelligent data analysis, such as rough set theory(RST) and artificial neural network(ANN). The great successes of these new methods in prediction, variable analysis and model construction encourage us to supplement them with traditional statistical methods.

RST was originally proposed by Pawlak[2] as a mathematical approach to handle imprecision, vagueness and uncertainty in data analysis. The theory has been demonstrated to have its usefulness in successfully solving a variety of problems. In most cases, rough set are used as a tool of decision making,rule mining and attribute selection. To the appraisement of input variable importance and significance, there is a special measure named importance degree existed in RST. By this measure, a value included in $[0, 1]$ will be assign to an variable as the importance degree and scales the contribute of the input variables on output variable. The importance degree of all input values or variable sets can be used to sort the order of them[5]. An variable is said to have no contribution to the predication of the system if the importance degree of it is bigger than zero. And attribute C_1 is said to be more important than C_2 if the importance degree of C_1 is bigger than C_2.

In traditional statistical analysis, liner regression and logistical regression can be used to appraise the correlative links between variables and predications[3,4]. Compared with the importance degree in RST, statistical methods can find the attributes significantly contribute to the predictions of the system for a given significance level, while the concept, significance degree, is never discussed in RST.

For most of the users who are accustomed to statistical analysis of data, it is not easy to accept rough set theory or other intelligent approaches as a new tool for data analysis because the lack of "P-value" and "significance level". A lot of work have been done to construct the bridge between RST and probabilistic theory[6,7], and the bridge between ANN with statistical theory[8]. But almost no work is done to construct the relation of RST and statistical theory.

In this paper, a randomization test for statistically assessing the importance degree of variables in RST is proposed. Based on this nonparametric approach, the variables that significantly contribute to the predictions of the system can

978-0-7695-3490-9/08 $25.00 © 2008 IEEE
DOI 10.1109/PACIIA.2008.113

be identified based on the given significance level and the bridge of rough set theory and statistical approach is constructed. Different randomization approaches are demonstrated based on the type of the data, continuous or discrete. At last, an experiment is done to show the effect of randomization test in RST.

2. The importance degree in rough set theory[10,11]

Let $S =< U,\ A = \ C \cup D,\ V,\ f >$ be the database. U is the universe. $C = \{C_1,\ C_2, \cdots,\ C_n\}$ is the set of input variables(condition-attributes), and D is the set of prediction variables(decision-attributes). Suppose there is only one attribute included in D. $V = (\bigcup_{i=1}^{n} V_{C_i}) \cup V_D$, $f : U \times A \rightarrow V$ is the evaluation function and $\forall a \in A$, $f_a : U \times A \rightarrow V_a$.

$\forall B \subseteq A$, the indiscernibility relation on U based on B is denoted by $I(B)$ and defined as follows:

$$I(B) = \{(x, y) \in U \times U : f_a(x) = f_a(y), \forall a \in B\}. \quad (1)$$

It's easily seen that the indiscernibility relation defined in this way is an equivalence relation. The family of all equivalence classes of the relation $I(B)$ is denoted by $U/I(B)$, in short U/B, and an equivalence class containing an element x will be denoted as $I(B)(x)$, in short $B(x)$.

Let X be a subset(a concept) of the universe, $\forall B \subseteq A$, the B-lower and B-upper approximation of X are denoted by $B_*(X)$ and $B^*(X)$ respectively, and defined as follows:

$$B_*(X) = \{x \in U : B(x) \subseteq X\}, \quad (2)$$

$$B^*(X) = \{x \in U : B(x) \cap X \neq \emptyset\}. \quad (3)$$

$\forall B, E \subseteq A$, B depends in degree $k(0 \leq k \leq 1)$ on E, denoted by $E \Rightarrow_k B$, if

$$k = r_E(B) = \frac{|POS_E(B)|}{|U|}, \quad (4)$$

where $|\cdot|$ is the cardinality of a set and

$$POS_E(B) = \cup_{X \in U/B} E_*(X). \quad (5)$$

The expression $POS_E(B)$, called a positive region of the partition U/B with respect to E, is a set of all elements of U that can be uniquely classified to blocks of the partition U/B, by means of E.

$\forall B \subseteq C$, the importance degree of B related to the decision attribute D(in short, the importance degree of B) is denoted by $RI_C^D(B)$ and defined as:

$$RI_C^D(B) = r_C(D) - r_{C \setminus B}(D). \quad (6)$$

when $B = \{a\}$, $RI_C^D(a)$ is the importance degree of attribute a respected to D.

3. The randomization test of importance degree in RST

3.1. The algorithm of randomization test of importance degree

In this section, we propose and demonstrate a randomization test for the importance degree of input variables in RST. This objective is similar to statistical pruning techniques(e.g. asymptotic t-test), yet does not have to conform to the assumptions of parametric and non-parametric methods because the randomization approach empirically constructs the distribution of expected values under the null hypothesis for the test statistic(i.e. importance degree) from the data at hand. This approach can be used as a quantitative tool for relativity analysis and selecting statistically significant input variables for system.

The following is the randomization approach for testing the statistical significance of input variables in C:

Step 1: Compute the importance degree of all subsets of input variables set C using the original data;($\forall B \subseteq C$, compute $RI_C^D(B)$)

Step 2: $\forall B \in \mathcal{P}(C)$;

Step 3: randomization of the original data based on the objects concerned, and a randomized data S_{random} is obtained;(This step will be detailedly discussed in the following part)

Step 4: Compute the importance degree of B of input variables set C using the randomized data S_{random}, and the value is just the randomized value of $RI_C^D(B)$;

Step 5: Repeat steps 3 and 4 a large number of times(i.e.1000 times or more);

Step 6: Compute the P-value of the importance degree of B, which can be calculated as the proportion of randomized values whose value is more extreme than or equals to the original observed values, the P-value of the importance degree of B is denoted by $Pvalue(B)$;

Step 7: For a given significance level α, study the statistical significance of all of the subsets of condition variable set according to their P-value.

The randomization of the original data is one of the main steps in the process of randomization test.

If there is not continuous variable existed in the attribute set A, $\forall B \in \mathcal{P}(C)$, and we want to estimate the significance level of the importance degree of B, the randomization of data can be performed as follows:

Method 1: Randomly permute the original response values of decision attribute D.

Method 2: Randomly permute the original values of variables in B.

This two methods show us there are two way of randomization, one deals with the input(condition) variable set, the other deals with the response(decision) variable set. If there

is a continuous variable existed in the variable set which is waiting for randomization, there are two way of randomization for the continuous variable as well.

Method 3: There are two steps of this method. They are shown as follows:
step1: discretize the continuous variable by some fixed discretization method (such as Eqw, Eqf, Chimerge and Zeta, etc.), and a discrete variable is obtained;
step2: randomly permute the values of the discrete variable.

Method 4: There are two steps of this method. They are shown as follows:
step1: randomly permute the original values of the continuous variable;
step2: discretize the continuous variable by some fixed discretization method.

3.2. Some remarks

There are some remarks related to the randomization test of importance degree, which should be noticed.

Remark 1: If the importance degree of a variable set equals to zero, the P-value of the variable set determinately equals to 1. This result is guaranteed by proposition 1.

Proposition 1 For the given data $S =< U,\ A = \ C \cup D,\ V,\ f >, B \in \mathcal{P}(C)$,

$$RI_C^D(B) = 0 \Leftrightarrow Pvalue(B) = 1. \tag{7}$$

proof. Suppose S_{random} is the data after randomization, $RI_C^D(B)_{S_{random}}$ and $RI_C^D(B)_S$ are the importance degrees of B based on the data set S_{random} and S respectively.
($\Rightarrow$)$\forall S_{random}$,

$$RI_C^D(B)_{S_{random}} \geq 0 = RI_C^D(B)_S. \tag{8}$$

if there are n randomization times, then

$$Pvalue(B) = \frac{n}{n} = 1. \tag{9}$$

($\Leftarrow$)If $Pvalue(B) = 1$, then $\forall S_{random}$,

$$RI_C^D(B)_{S_{random}} \geq RI_C^D(B)_S. \tag{10}$$

If $RI_C^D(B)_S \neq 0$, there exists a randomization such that $RI_C^D(B)_{S_{random}} = 0$, this contradicts with equation (10).

Remark 2: There are more than one method of randomization in our algorithm. But in the process of randomization test, once the randomization method is chosen, it must be fixed.

Remark 3: In the algorithm of randomization test of importance degree, the randomization time should be rightly chosen. Although 1000 times is enough in a general way, another way should be proposed to choose the right number. The new method is shown as follows:

– – – – – – – – – – – – – – – – – –
Input: Original data S, Threshold λ, $\forall B \in \mathcal{P}(C)$
Output: P-value(B)
  Randomize S for m times.
  $\theta = 1$; $n = m + 1$
  do while $\theta < \lambda$
    new data S^n_{random} is obtained by n^{th} randomization;
    $RI(B)^n = RI_C^D(B)_{S^n_{random}}$
    compute P-value$(B)^n$(based on $\{RI(B)^1, \cdots, RI(B)^n\}$)
    $\theta = \sum_{i=n-k}^{n-1} \|P - value(B)^i - P - value(B)^{i+1}\|$
    $n = n + 1$;
  end
  P-value$(B) = RI(B)^n$
– – – – – – – – – – – – – – – – – –

In this algorithm, m and k are chosen by practical data. In general way, $m = 400$, $k = 20$.

3.3. The merits of the randomization test of importance degree

There are some important merits of the randomization test of importance degree proposed in this paper. They are shown as follows:

(1).Easily appraise the interaction of the variables.

In traditional variable importance appraisement,such as statistical method and ANN, it's very difficult to discuss the interaction of variables. In statistical method, there are some suppositions must be satisfied, and the type of the interaction must be determined before the measurement. In ANN, only two-way interaction[13]is discussed and it not easy to generalize the algorithm to more than two way's interaction. RST has the merit of discuss the importance degree of variable set. Thus, the randomization test of importance degree in RST can be easily used in interaction appraisement of multiple variables.

(2).Distinguish the variables which have the same importance degrees.

Since the importance degree analysis in RST depends on the indiscernibility relations and inclusion operator, some variables will have the same importance degrees whereas the importance of them are actually different. How to distinguish the variables which have the same importance degree? We will find the fact that if the variables actually have different importance degree, the P-values of them will be different.

(3).Realize the comparison of RST and traditional statistical method.

Although there are lots of merits of RST, for most of the users who are accustomed to statistical analysis of data, it is not easy to accept rough set theory or other intelligent approach as a new tool for data analysis because the lack of "P-value" and "significance level". In this paper, we introduce randomization test to RST, and "P-value" and "sig-

nificance level" are also easily obtained. Now, there is a bridge between RST and traditional statistical method, and their results are changed to be comparable. Thus, the introduction of the randomization test of importance degree in RST enhances the usability of RST in practical domains, especially in epidemiology.

4. Experiment

There is a medical record of the eyeglass type of different patients[12]. The data is shown in table 1. There are four input variables related to the type of the glasses:
C_1:Age. $V_{C_1} = \{1.young;\ 2.midlife;\ 3.old.\}$
C_2:Kind $V_{C_2} = \{1.myopia;\ 2.hyperopia.\}$
C_3:Astigmatism $V_{C_3} = \{1.no;\ 2.yes.\}$
C_4:Tear $V_{C_4} = \{1.decrease;\ 2.normal.\}$
The decision variable is :
D:eyeglass type $V_D = \{1.type\ 1;\ 2.type\ 2;\ 3.type\ 3\}$.

Table 1. The medical record of the eyeglass type

U	C_1	C_2	C_3	C_4	D
1	1	1	2	2	1
2	1	2	2	2	1
3	2	1	2	2	1
4	3	1	2	2	1
5	1	1	1	2	2
6	1	2	1	2	2
7	2	1	1	2	2
8	2	2	1	2	2
9	3	2	1	2	2
10	1	1	1	1	3
11	1	1	2	1	3
12	1	2	1	1	3
13	1	2	2	1	3
14	2	1	1	1	3
15	2	1	2	1	3
16	2	2	1	1	3
17	2	2	2	1	3
18	2	2	2	2	3
19	3	1	1	1	3
20	3	1	1	2	3
21	3	1	2	1	3
22	3	2	1	1	3
23	3	2	2	1	3
24	3	2	2	2	3

In the model of eyeglass type, we use equation (6) to compute the importance degrees of variable sets and use the randomization test put forward in section 2 to obtain the P-values of variable sets. The results of one variable and two variables are shown in table 2. In the process of randomization test, $m = 400$, $k = 20$. The distribution of random importance degrees of $\{C_3, C_4\}$ is shown in figure.1, where arrow represents observed importance degrees by original data. The P-values of $\{C_3, C_4\}$ in the process of randomization are shown in figure.2. By increase of the randomization times, the P-values of all variable sets are going to changeless.

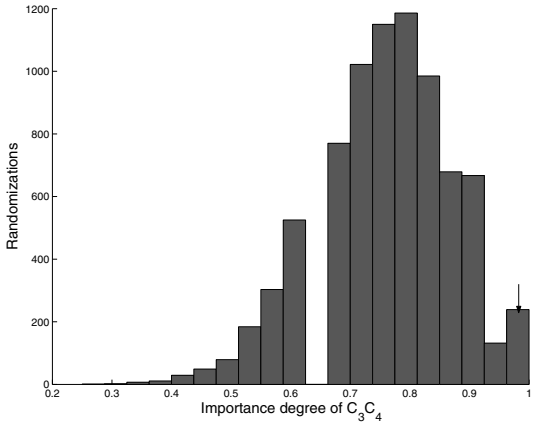

Figure 1. The distributions of random importance degrees of $\{C_3, C_4\}$

Table 2. The importance degrees and the P-values of variables in the model of eyeglass type

i	Variables	Importance Degree	P-value
1	$C_1 : Age$	0.25	0.293
2	$C_2 : Kind$	0.25	0.318
3	$C_3 : Astigmatism$	0.5	0.102
4	$C_4 : Tear$	0.75	**0.068**
5	$C_1 \wedge C_2$	0.5	0.159
6	$C_1 \wedge C_3$	0.5	0.158
7	$C_1 \wedge C_4$	1	0.108
8	$C_2 \wedge C_3$	0.5	0.110
9	$C_2 \wedge C_4$	1	**0.033**
10	$C_3 \wedge C_4$	1	**0.029**

In this experiment, if we discuss the importance degrees of variables by equation (6), three groups of variable sets have the same importance degrees and can not be distinguished. For example, the variable sets $\{C_1, C_4\}$,$\{C_2, C_4\}$ and $\{C_3, C_4\}$ have the same values of importance degree,

$$RI(\{C_1, C_4\}) = RI(\{C_2, C_4\}) = RI(\{C_3, C_4\}). \quad (11)$$

At this time, if randomization test is used to analyze these three variable sets, they can easily be distinguished as

$$\text{P-value}(\{C_1, C_4\}) < \text{P-value}(\{C_2, C_4\}) < \text{P-value}(\{C_3, C_4\}). \quad (12)$$

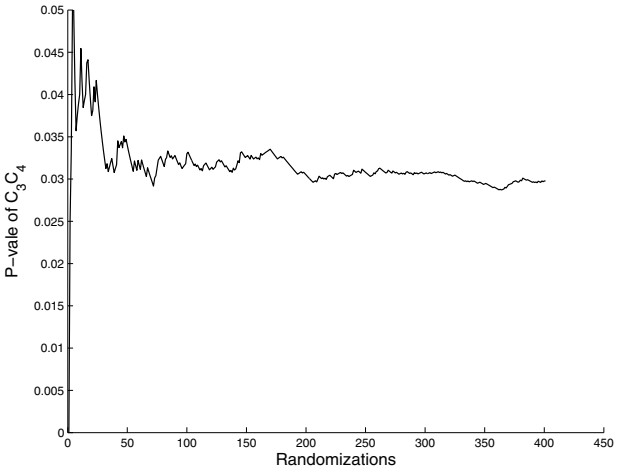

Figure 2. The P-values of $\{C_3, C_4\}$ in the process of randomization

Table 3. The linear regression about the model of eyeglass type

$Model$	B	$Std.Error$	t	$Sig.$
$Constant$	3.708	0.646	5.737	0.000
C_1	0.188	0.136	1.381	0.183
C_2	0.250	0.222	1.128	0.274
C_3	-0.250	0.222	-1.128	0.274
C_4	-1.083	0.222	-4.886	**0.000**

We use SPSS13 to do the linear regression about the model of eyeglass type, and the results are shown in table 3.In table 2,3, if 0.1 is chosen as the significance level, we will find the fact that the variable C_4 is significantly contribute to the decision. The result of randomization test is similar to the results of statistical method.

5 Conclusion

Randomization test for importance degree of variables in rough set theory is proposed and demonstrated in this paper. By the introduction of this algorithm, a bridge between RST and traditional statistical method is constructed, and their results are changed to be comparable, and this result enhances the usability of RST in practical domains, especially in epidemiology.

Acknowledgments.

This research is supported in part by Program for New Century Excellent Talents in University(NCET-06-0131); National Natural Science Foundation of China (No. 40672195); the National 973 Fundamental Research Project of China (Grant No. 2002CB312200)

References

[1] J.K.Stanley, Z.Patricia, H.Frank. An approach for determining relative input parameter importance and significance in artificial neural networks. Ecological Modelling,204(3-4):326-334,2007.

[2] Z.Pawlak.Rough sets.International Journal of computer and information science,11(5):341-356,1983.

[3] A.David.Statistical models:theory and practice.Cambridge,New York,Cambridge University Press,2005.

[4] A.C.Davison.Statistical models.Cambridge,New York,Cambridge University Press,2003.

[5] S.Jerzy,S.Krzysztof. Rough sets as a tool for studying attribute dependencies in the urinary stones treatment data set.Lecture Notes in Computer Science,1263:36-46,1997.

[6] Y.Y.Yao.Probabilistic rough set approximations.International Journal of Approximate Reasoning,doi:10.1016/j.ijar.2007.05.019,2007.

[7] W.Ziarko.Probabilistic approach to rough sets.International Journal of Approximate Reasoning,doi:10.1016/j.ijar.2007.05.014,2007.

[8] J.D.Olden,D.A.Jackson. Illuminating the "black box":a randomization approach for understanding variable ocntributions in artificial neural networks.Ecological Modelling,154:135-150,2002.

[9] W.J.Conover.Practical Nonparametric Statistics.New York,Wiley Publishing,1999.

[10] Z.Pwalak.Rough set approach to knowledge-based decision support.European Journal of Operational Research,99:48-57,1997.

[11] H.L.Zeng.Rough set theory and its application.Chongqing,China,Chongqing University Press,1996.

[12] M.Boryczka.Derivation of optimal decision algorithms from decision tables using rough sets.Bulletion of Polish Academy of Sciences,vol.36:252-260,1988.

[13] M.Gevreya,I.Dimopoulosb,S.Leka.Two-way interaction of input variables in the sensitivity analysis of neural network models.Ecological Modelling,1995:43-50,2006.