Applied Soft Computing 28 (2015) 160–166

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Proposal of a statistical test rule induction method by use of the decision table

Yuichi Kato [a,*], Tetsuro Saeki [b], Shoutaro Mizuno [a]

[a] Interdisciplinary Faculty of Science and Engineering, Shimane University, 1060 Nishikawatsu-cho, Matsue, Shimane 690-8504, Japan
[b] Faculty of Engineering, Yamaguchi University, 2-16-1 Tokiwadai, Ube, Yamaguchi 755-8611, Japan

ARTICLE INFO

Article history:
Received 20 July 2012
Received in revised form
10 November 2014
Accepted 30 November 2014
Available online 9 December 2014

Keywords:
Rough sets
Rule induction
Statistical test
Rule box
p-Value

ABSTRACT

Rough sets theory is widely used as a method for estimating and/or inducing the knowledge structure of if-then rules from various decision tables. This paper presents the results of a retest of rough set rule induction ability by the use of simulation data sets. The conventional method has two main problems: firstly the diversification of the estimated rules, and secondly the strong dependence of the estimated rules on the data set sampling from the population. We here propose a new rule induction method based on the view that the rules existing in their population cause partiality of the distribution of the decision attribute values. This partiality can be utilized to detect the rules by use of a statistical test. The proposed new method is applied to the simulation data sets. The results show the method is valid and has clear advantages, as it overcomes the above problems inherent in the conventional method.

1. Introduction

Rough sets theory introduced by Pawlak [1] has been studied as an object of mathematical interest, and has provided many useful analytical methods in areas such as machine learning and data mining. One of the important tasks of rough sets is to compute the attribute reduct, which provides the construction of the decision table of interest by the minimum number of attributes, maintaining knowledge contained in it. Another task is to extract if-then rules from the decision table, by the use of approximations with indexes such as accuracy and coverage. This is useful for diagnosis systems for diseases, discrimination problems, and other aspects. Theoretical studies and practical algorithms for the above two tasks are available in the literature [2–7].

This paper presents the results of a retest of the ability of the conventional basic rough set [4,6,7] by the use of simulation data sets, and notes that the conventional method has two main problems. The first problem is that the estimated rules are composed of a large number of subsets of the true rule set and/or indifferent rules. That is, they cannot estimate rule sets specified in advance. The second is the strong dependence of the estimated rules on the data set sampling from the population. The results from one sample will differ from those from another. Attempts have been made to solve these two problems in the conventional method, by proposing the variable precision rough set (VPRS) [5] and by studying the effects of sampling from statistical viewpoints [8–11]. However, these trials seem do not to be essential and /or direct studies to resolve these problems, and they only partially succeed in improvement of them.

Consequently, we here propose a new method that directly estimates the if-then rules by the use of a statistical hypothesis test. Specifically, we regard the rule estimation problem by use of the decision table as the problem of identifying a black box containing rules. The inputs and outputs of the box are the condition part of a if–then rule and the decision part of the rule, respectively. We conduct a preliminary experiment in a white box which specifies if-then rules in advance, generates the input of the condition part randomly, and decides the output by use of the specified rules. From this preliminary experiment, we found that the decision table basically contains two types of data set. The first is the data set controlled by the rules in the rule box, and the second is the data set not to be applied to the rules, and the output is obtained by chance. Accordingly, we propose the following rule estimation strategy: (1) We set up the null hypothesis that a trying rule as a test does not exist in the black box. (2) We test the hypothesis using the data set of the decision table sampling the population, and decide whether the hypothesis is true or not. (3) We repeat procedures (1) and (2) changing the trying rule systematically and efficiently, and estimate all the rules in the black box. The validity and usefulness are

* Corresponding author. Tel.: +81 852 32 6470.
E-mail address: ykato@cis.shimane-u.ac.jp (Y. Kato).

http://dx.doi.org/10.1016/j.asoc.2014.11.041
1568-4946/

Table 1
An example of a decision table.

U	$C(1)$	$C(2)$	$C(3)$	$C(4)$	$C(5)$	$C(6)$	D
1	1	1	4	3	6	1	1
2	4	1	1	5	4	4	2
3	6	4	6	1	6	1	1
4	2	2	4	3	1	1	2
...	...	...	...	...	...	...	...
$N-1$	2	2	1	1	6	4	2
N	4	2	3	6	3	5	1

confirmed by comparing the results from our proposed method with those by the conventional methods in a simulation experiment.

2. Model of data generation

The decision table shown in Table 1 is first obtained. This is conventionally denoted with $S = (U, A = C \cup D, V, \rho)$. Here, $U = \{u(i) | i = 1, \ldots, N = |U|\}$ is a sample set. A is an attribute set, $C = \{C(j) | j = 1, \ldots, |C|\}$ is a condition attribute set, $C(j)$ ($j = 1, \ldots, |C|$) is the member of C and a condition attribute, and D is a decision attribute. V is a set of attribute values and denoted with $V = \bigcup_{a \in A} V_a$ and is characterized by an information function $\rho : U \times A \rightarrow V$. In this example, if $a = C(j) \in A$ ($j = 1, \ldots, |C| = 6$) then $V_a = \{1, 2, \ldots, 6\}$ and if $a = D$ then $V_a = \{1, 2\}$, and $\rho(u(1), C(1)) = 1$, $\rho(u(2), C(1)) = 4$ and so on.

The decision table in Table 1 is obtained by generating the condition part of $u(i)$ denoted by $u^C(i)$ with the uniform distribution with regard to $V_{a=C(j)}$ ($j = 1, \ldots, |C| = 6$) and by deciding the decision part of $u(i)$ denoted by $u^D(i)$ by use of the following specified rules $R(1)$ and $R(2)$, and decision assumptions (As1), (As2), and (As3):

Specified rules:

$R(1)$: if $C(1) = 1 \wedge C(2) = 1 \vee C(3) = 1 \wedge C(4) = 1$ then $D = 1$.
$R(2)$: if $C(1) = 2 \wedge C(2) = 2 \vee C(3) = 2 \wedge C(4) = 2$ then $D = 2$.

Decision assumptions:

(As1): $u^C(i)$ can be applied to $R(k)$ and $u^D(i)$ is uniquely determined as $D = d(k)$.
(As2): $u^C(i)$ cannot be applied to any $R(k)$, and $u^D(i)$ can only be determined randomly.
(As3): $u^C(i)$ can be applied to several $R(k)$ ($k = k1, k2, \ldots$) and their outputs of $u^C(i)$ conflict with each other. Accordingly, the output of $u^C(i)$ must be randomly determined from the conflicted outputs.

Here, $\wedge$ is conjunction and $\vee$ is disjunction. The case of (As1) is $u(1)$ and $u(4)$, (As2) is $u(2)$, $u(3)$ and $u(N)$ and (As3) is $u(N-1)$ in Table 1. The case of (As1) is a uniquely determined case, (As2) is what is known as an indifferent case, and (As3) is an inconsistent case.

3. Results of retest by the conventional method and its consideration

We generated three cases of the decision table using the model in Section 2 with $N = 10{,}000$, applied two representative conventional methods to them, and retested their abilities of rule induction. The first method was the LEM2 algorithm [4], for which software could be downloaded [12], and the second was the FDMM algorithm [7] which was developed from the decision matrix method [2,6]. Table 2 shows the results from the two methods using the lower approximation. N_{actual} is the net number of N deleting the same data. The row of the table denoted STRIM will be explained in Section 6. The experiment was conducted on a PC with a Celeron(R) CPU with 2.67 GHz clock speed and 992 MB of RAM memory. Reducing time in the table is the execution time of the rule induction for reference. Number of rule length is the number of the conjunction in the condition part of the estimated rules. For example, if the condition part is $C(1) = 1$ then the length is 1 (hereafter this part is denoted with (100000) for convenience). If $C(1) = 1 \wedge C(2) = 1$ ((110000)) then the length is 2. The length of $C(1) = 1 \wedge C(2) = 1 \wedge C(3) = 1$ ((111000)) is 3, the length of $C(1) = 1 \wedge C(2) = 1 \wedge C(3) = 1$ $C(6) = 4$ ((111004)) is 4, and so on. The number 48 at the rule length 3 of FDMM in Case 1 shows that FDMM inducted 48 rules of the rule length 3. Total is the sum of the numbers of all the estimated rules.

Table 2
Three cases of comparisons of the time and the rule length on reducing rules between LEM2, FDMM and STRIM (N_{actual}: actual data number).

Case no.	Method	Reducing time [s]	Number of rule length						
Data number (N_{actual})			1	2	3	4	5	6	Total
Case 1	FDMM	100	0	0	48	631	3650	35	4364
10000	LEM2	7061	0	0	47	523	3819	35	4424
(9482)	STRIM	5	0	4	1	0	0	0	5
Case 2	FDMM	101	0	0	51	609	3751	37	4448
10000	LEM2	8504	0	0	47	527	3920	37	4531
(9417)	STRIM	5	0	4	5	0	0	0	9
Case 3	FDMM	101	0	0	45	602	3770	42	4459
10000	LEM2	7059	0	0	43	501	3945	42	4531
(9432)	STRIM	5	0	4	2	0	0	0	6

Table 3 incidentally shows one of the rules with the highest coverage index by every rule length (RL) inducted by LEM2 for three cases in Table 2. Here, $(n1, n2)$ is the frequency of samples $\{u^{D=1}(i)\}$ and $\{u^{D=2}(i)\}$ respectively which satisfy the corresponding condition part $\bigwedge_j (C(j_k) = V_{C(j_k)})$ of the inducted rules and is arranged them as ($|\{D = 1\}| = n1$, $|\{D = 2\}| = n2$), and accuracy and coverage indexes are defined as $\frac{\max(n1,n2)}{n1+n2}$ and $\frac{\max(n1,n2)}{|\{D=i\}|}$ and respectively. The column of kind of rule shows a category to which the corresponding inducted rule belongs. Sub denotes a sub-rule of true rules which are included in the true rules specified in advance, and Ind denotes an indifferent rule which is not included in the true rules, that is, not specified in the rule box in advance. For example, at the first row of Case 1, $RL = 3$, the inducted rule is if (220300) then $D = 2$, $(n1, n2)$ was (0, 61), the accuracy is 61/61 = 1.0 since the lower approximation was used, the coverages is $61/4773 \approx 0.013$, and kind of rule is Sub since (220300) is included in $R(2)$ specified in advance. From Tables 2 and 3, we can see that:

(1) Regarding $u(i)$ ($i = 1, \ldots, N_{\text{actual}}$) as a rule of $D = 1$ or $D = 2$, LEM2 and/or FDMM reduced the rule set, that is, the decision table to about its 45 [%] with more than the rule length 3. The rules with length 6 were not arranged at all.
(2) $R(1)$ and $R(2)$ specified in advance were rules with the length 2 and their total rules were 4. The rules induced by LEM2 and/or FDMM were confirmed to be included in either $R(1)$ or $R(2)$, or in different rules as shown in Table 3. Accordingly, LEM2 and/or FDMM estimated the sub-rules of the original rules and/or indifferent rules obtained by chance from the sample set, multiplying the number by several thousands. To say other words, the estimated results by both methods have scarce reliability since the coverage of those rules is very low though their accuracy is one.
(3) Both methods are thought to be highly dependent on the sample data set, since the estimated rule set of the three cases differed considerably.