Attribute Reduction Algorithm for Information System without Decision Attributes

Li Hong-Chan[1, a], Zhu Hao-Dong[1,b]

[1] School of Computer and Communication Engineering, Zhengzhou University of Light Industry Zhengzhou Henan, 450002, China

[a]ketizulunwen@163.com, [b]zhuhaodong80@163.com

Keywords: Attribute Reduction; Decision Attribute; Information System; Set Theory

Abstract. The classical attribute reduction algorithm and its extended algorithms base on information system with decision attributes and can not be applied to attribute reduction for information system without decision attributes. So, based on rough set theory, this paper studied attribute reduction for information system without decision attributes in domain division of set theory and presented a heuristic attribute reduction algorithm. To a certain extent, the algorithm can resolve the attribute reduction problem for information system without decision attributes and extend application of Rough Set Theory. The analysis of the realistic example shows that the algorithm is effective and feasible.

Introduction

Rough set theory is a mathematical tool which was proposed by Poland scientist Z.Pawlak in 1982 on the study of uncertain and inaccurate knowledge[1,2] and has been widely adopted in artificial intelligence, data mining, model identification and intelligent information process, thus attracting great attention from the international academics[3,4]. Because attribute reduction is one of the core concepts in Rough Set theory, many researchers are devoting themselves to the study of information system attribute reduction algorithm. Currently, the researchers have advanced many attribute reduction algorithms such as the one based on the positive region, the one based on discernibility matrix, the one based on the information entropy and so on. However, theses attribute reduction algorithms are all based on the information system with decision attributes. The information system without decision attribute is also a kind of important data set, for example, the information system in cluster analysis has no decision attributes. Without attribute reduction, the information system without decision attributes will greatly reduce the performance of its relevant algorithms due to its high dimension sparse character, which not only will take much more time but also generate unsatisfying result. The most effective method to solve the above problem is to reduce the dimension through attribute reduction. Currently, most of the researchers will adopt the unsupervised feature selection method of data mining on the attribute reduction for information system without decision attributes while very few of them adopt attribute reduction algorithms based on rough set theory. On such condition, therefore, this paper studied attribute reduction for information system without decision attributes in domain division of set theory and presented a heuristic attribute reduction algorithm.

Concepts Introduction

The Rough Set theory defines the knowledge as: knowledge is a cluster set of the indiscernibility relation. Thus, knowledge is endowed with a distinct mathematic meaning and could be dealt with mathematic methods.

Definition 1. Information system is expressed as $S=<U,R,V,f>$, U is the object set, $R=C \cup D$ is the attribute set, where C is the conditional attribute set, D is the decision attribute set, V is the attribute value set, V_r stands for the range of attribute r, f: $U \times R \rightarrow V$ is a mapping function which assigns the attribute of each object in U. Information system also could be expressed with the two dimension

Published by Atlantis Press, Paris, France.

which is known as decision table, in which the row stands for objective x_i and the column stands for attributer, thus $r(x_i)$stands for the value of attribute r on the ith objective[5-7]. If $D=\varnothing$ establishes , then the information system is called no decision attribute information system, which is presented as $S=<U,C,V,f>$.

Definition 2. For each attribute subset $B\subseteq C$,we define a discernibility binary relation (indistinguishability relation) Ind(B): Ind(B)=$\{(X,Y)|(X,Y)\in U\times U,\forall b\in B\ (f(b,X)=f(b,Y))\}$. Ind($B$) is the equivalence relation which deduces the divide of U/Ind(B).

Definition 3 The attribute reduction problem of the no decision attribute information system $S=<U,C,V,f>$ is to figure out $R\subseteq C$ to meet the following requirements:$J=\min|R|$ and U/ind(R)=U/ind(C).

Attribute Importance Based on Set Division

As for information system without decision attributes $S=<U,C,V,f>$, set the division of the selected attribute subset $R\in C$ and the conditional attribute $c\in C$-R that is to be selected as: $\pi_R=U/\text{ind}(R)=\{X_1,X_2,\ldots,X_r\}$, $\pi_c=U/c=\{Y_1,Y_2,\ldots,Y_t\}$ .Then, divideπ(including the empty set) with the divided product[8]:

$$\pi=\pi_R\bullet\pi_c=U/ind(R\cap c)=\begin{Bmatrix} E_{11} & E_{12} & \ldots & E_{1t} \\ E_{21} & E_{22} & \ldots & E_{2t} \\ \vdots & \vdots & \vdots & \\ E_{r1} & E_{r2} & \ldots & E_{rt} \end{Bmatrix} \quad (1)$$

Where, $E_{ij}=X_i\cap Y_j$,i=1,2,…r , j=1,2…,t meets the requirement of $X_i=\bigcup_{j=1}^{t}E_{ij}$, $Y_j=\bigcup_{i=1}^{r}E_{ij}$.

Increasing the equivalence relation of R set, namely, refine the division of R set, then the non-null elements in formula (1) will grow in number. The much the R set is refined, the more the *rank* (the divided block number) of R set will increase.

As for the conditional attribute of information system S, $c\in C$-core(C), firstly set R= core(C), U/ind(R)= $\{X_1,X_2,\ldots,X_r\}$, $U/c=\{Y_1,Y_2,\ldots,Y_t\}$,$E_{ij}=X_i\cap Y_j$,thus we can get the E matrix in the formula (1), then simplify and define the E matrix in such way as:

$$E1_{ij}=\begin{cases}0,\ E_{ij}=\varnothing\\ 1,\ E_{ij}\neq\varnothing\end{cases}\Rightarrow E2_i=\sum_{j=1}^{t}E1_{ij}\Rightarrow E3_i=\begin{cases}0,\ E2_i=1\\ E2_i,\ E2_i>1\end{cases}.$$

To furthest divide the conditional attribute set in the discourse domain- that is the the system reduction set obtained when $D=\varnothing$ establishes- is actually to figure out the refined core attributes (or selected attribute subsets) as many as possible in the attribute set to gain the result of minimum reduction attribute number and maximum division[8]. If make the divided block as basic information granule, we can establish the heuristic rule that adopts the maximum difference degree as the attribute significance compared to the core attribute (or the current selected attribute subset R). The value of the maximum difference degree depends on whether the product division of the divided attribute that is to be selected and the one that has be selected can generate maximum increase form R set rank (greedy algorithm).

Definition 4. The maximum difference degree attribute significance of attribute c about R is showed as follows:

$$sig\ (c)=\sum_{i=1}^{r}E3_i \quad (2)$$

Or, based on ***E*** matrix and embedding the function relation definition, the maximum difference degree attribute significance of attribute c about R is showed as follows:

$$sig\ (c)=\sum_{i=1}^{r}g(\sum_{j=1}^{t}f(E_{ij})) \quad (3)$$

$$\text{Where } f(E_{ij})=\begin{cases}0, & E_{ij}=\varnothing\\ 1, & E_{ij}\neq\varnothing\end{cases},\quad g(*)=\begin{cases}0, & \sum f(E_{ij})=1\\ \sum f(E_{ij}), & else\end{cases}.$$

From the definition 4 we can see that: the larger division difference between attribute c and set R, the greater the significance is, and it should be taken into first consideration.

Proposed Attribute Reduction Algorithm

Algorithm concept: firstly, set $R=\emptyset$ and figure out the difference attribute significance of each attribute in C-R. and then, selecting the attributes according to the difference degree and reduce them in a heuristic way. After R set select a a attribute, its indiscernibility relation will change because the division determined by R set in the discourse domain can not refine the division block with merely one objective; thus reserving it from the discourse domain, which greatly reduces the inference on sequenced attribute evaluation and optimizes the heuristic reduction algorithm. The algorithm pseudo code is showed as follows:

Input: information system $S=<U,C,V,f>$, $C=\{c_1, c_2,\ldots, c_m\}$.

Output: attribute reduction set red(C) .

Step 1. R=Ø.

Step 2. Compute the ind(C) of S.

Step 3. Figure out the core attribute of S, R=core attribute set.

Step 4. $C'=C-R$, IF $C'=\emptyset$ THEN red(C)=R ,output red(C), the algorithm stop. ELSE go to Step5 .

Step 5. IF $R=\emptyset$ THEN $R=c_r(c_r=\{c_i|\text{rank}(c_i)=\max(\text{rank}(c_i)),c_i\in C'\})$.

Step 6. Compute ind(R),$U = U - x(x\in U, [x]_R=\{x\})$.

Step 7. Compute and simplify $\boldsymbol{E}$ matrix on U, compute sig(c_i), $c_i\in C'$ according to the formula (3).

Step 8. $c_{max}=\{c_i \mid \max \text{sig}(c_i), c_i\in C'\}$, $R=R\cup c_{max}$.

Step 9. IF ind(R)=ind(C) THEN red(C)=R, output red(C) , the algorithm stop. ELSE go to Step10.

Step 10. $C'= C'-c_{max}$, go to Step6.

Compared with decision table attribute reduction, the no decision one makes each objective as a category, thus requiring much more attributes to be refined.

The time complexity of the algorithm includes such two parts as: one is to compute the system product division; the other is to define the attribute significance. Both of parts require to compute the intersection of the equivalence relations with the time complexity of $O(|U|^2)$; thus, the worst time complexity is $O(|C||U|^2)$ when solving each conditional attribute, which can, to some extent, resolve the attribute reduction problem of no decision attributes information systems and extend application of Rough Set Theory.

Algorithm Example

Reducing the attributes without decision attribute CTR dataset showed in table 1 according to the algorithm in this paper.

Table 1. CTR Dataset without Decision Attributes

U	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	U	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9
1	0	1	1	1	1	1	1	1	0	11	1	0	0	1	1	0	0	0	0
2	0	1	0	1	1	1	1	0	0	12	0	0	0	0	1	0	0	0	0
3	0	0	1	1	1	1	1	0	1	13	1	0	1	1	0	1	1	0	0
4	0	1	0	1	1	0	0	0	0	14	1	0	0	0	0	0	2	0	0
5	0	1	0	0	1	0	0	1	2	15	0	0	1	1	1	0	1	0	0
6	0	1	0	1	1	0	1	0	2	16	0	1	0	1	1	0	1	1	0
7	1	0	0	0	0	1	2	0	1	17	0	0	0	1	1	0	1	1	0
8	0	0	0	0	0	1	2	0	0	18	1	0	0	1	0	1	0	0	0
9	0	0	0	0	0	1	0	1	0	19	0	0	0	1	0	1	0	0	0
10	1	0	0	1	0	1	2	0	1										

Step 1-Step 5. Core(C)=$\{c_1,c_2,c_4\}$, R=Core(C).

Step 6. Calculating U/ind(R),$U=U-x$ ($[x]_R=\{x\}$).

Step 7. Calculating and simplifying $\boldsymbol{E}$ matrix on U, according to the formula (3), we calculate sig(c_i), i=3,5,6,7,8,9 with the following result: c_3 c_5 c_6 c_7 c_8 c_9

5 6 1 2 4 3

Step 8-Step 10. $R=R\cup c_6$, but R does not meet the requirement of reduction set, go to Step 6

Step 6. Calculating U/ind(R), $U=U$-x ($[x]_R=\{x\}$).

Step 7. According to the formula (3), we calculate sig(c_i), i=3,5,7,8,9 with the following result: c_3 c_5 c_7 c_8 c_9

2 5 1 4 3

Step 8-Step 10. $R=R\cup c_7$, R still does not meet the requirement of reduction set, go to Step 6.

Step 6. Calculating U/ind(R),U=U – x ($[x]_R=\{x\}$).

Step 7. According to the formula (3), we calculate sig(c_i), i=3,5,8,9, with the following result:

c_3 c_5 c_8 c_9

1 3 2 3

Step 8. $c_{max}=c_3$, $R=R\cup c_{max}$.

Step 9. ind(R)=ind(C), red(C)=R=$\{c_1,c_2,c_3,c_4,c_6,c_7\}$ output red(C), STOP.

In the last round calculation of the attribute significance, the attribute c_3 and c_8 share the same significance and both meet the requirement of reduction; thus generating two reduction result while the other reduction set is red(C)= $\{c_1,c_2,c_3,c_4,c_6,c_7\}$.

Conclusion

Based on the Rough Set theory, the author studied attribute reduction for information system without decision attributes from the angle of discourse domain division and got such result as: the information system without decision attributes can obtain the maximum division of the discourse domain under the attribute feature description. Consequently, we can select the minimum attribute subset in the information system without decision attributes according to difference degree between the selected attribute and the one to be selected in such way as: the higher the difference degree, the greater significance the attribute possesses, thus should be taken into first consideration. Meanwhile, in the dynamic attribute selection process, we can quickly get the reduction set after reducing the influence of minimum division in discourse domain to attribute significance. Based on the above conclusion, this paper presented a heuristic attribute reduction algorithm. To a certain extent, the algorithm can resolve the attribute reduction problem for information system without decision attributes and extend application of Rough Set Theory.

Acknowledgment

This work is supported by the Foundation and Frontier Technologies Research Plan Projects of Henan Province of China (No. 102300410266), the Foundation and Frontier Technologies Research Plan Projects of Henan Province of China (No.122300410287) and a grant from the Ph.D.Research Funded Projects of Zhengzhou University of Light Industry (No. 2010BSJJ038). In addition, this work also received guidance from Huang De-Shuang who is a distinguished professor in Henan Province.

References

[1] Pawlak Z: Rough Set. International Journal of Computer and Information Sciences 11(15), 341-356(1982).

[2] Pawlak Z: Rough Set.Communications of ACM 38(11), 89 -95(1995).

[3] Pawlak Z: Vagueness and uncertainty-a rough set perspective. Computational Intelligence 11(2),227-232(1995).

[4] Liu Qing-zhen, Cai Jin-ding, and Wang Shao-fang: Fault diagnosis of power electronic circuits based on rough set-neural network System. Electric Power Automation Equipment 24(4),45-48(2004)

[5] Ding Hao, Ding Shi-Fei, and Hu Li-Hua: Research progress on Based on rough sets the attribute reduction. Computing and Engineering Science 32(6),92-94(2010)

[6] Zhang Xue-Pan, Gao She-Sheng, and Hu Pan: A new attribute reduction algorithm [J]. Computer Simulation 26(10), 195-197(2009).

[7] Zhang Wen-Dong, Li Ming-Zhuang, and Shi Xiao-Yan: Attribute reduction algorithm based on Rough set theory. Computer Engineering and Design 29(22), 5795-5797(2008).

[8] NI Zi-wei, CAI Jing-qiu. Discrete Mathematics. China Beijing: Science Press, 2002.10.

Published by Atlantis Press, Paris, France.