A Compression Method of Decision Table Based on Matrix Computation

Laipeng Luo, Ergen Liu
（School of Basic Sciences, East China Jiaotong University, Nanchang,P.R.China）
E-mail:loulp@tom.com

Abstract

A new algorithm of attribute reduction based on boolean matrix computation is proposed in this paper. The method compresses the valid information stored in table into a binary tree, at the same time deleting the invalid information and sharing a branch about the same prefix information. Some relative concepts such as local core attributes, local attribute reduction and global core attributes, global attribute reduction are introduced. The conclusions that the global core set is the union of all local core sets and the global attribute reduction sets are the union of respective local attribute reduction sets are proved. The attribute reduction steps of the algorithm are presented. At last, The correctness and effectiveness of the new algorithm are also shown in experiment and in an example.

Key words: Rough Set; equivalence matrix; attribute reduction; information compression

1. Introduction

Rough set theory, introduced by Zdzislaw Pawlak in the early 1980s[1,2], is a new mathematical tool to deal with vagueness and uncertainty. This approach seems to be of fundamental importance to artificial intelligence and cognitive sciences, especially in the areas of machine learning, knowledge acquisition,deci-sion analysis, knowledge discovery from databases, expert systems, decision support systems,inductive reasoning, and pattern recognition[3,4,5].

Attribute reduction is one of the main applications of Rough set.The general problem of finding all reductions are NP-hard, thus it is important for attribute reduction of Rough set to design algothms with lower price and investigate new computation method.A matrix computation method of Rough set was proposed by the author in[6,7]for information system.A matrix was seen as an internal representation of equivalence relations.By defining the operation of the equivalence matrix ,matrices are applied to define dependencies between two subsets of attributes,significance of an attribute ect.The approach presents a series of algorithms and their time complexity of attribute reduction.Howevre,there are still several problems to be solvered for the method:(1)The number of objects has great influence on time complexity of these algorithms;(2)These algorithms need too many computations of matrices;(3)The method only discusses matrix computation for information system;（4）How to apply the method to variable precision Rough set model.

With the above-mentioned motivation, in another paper[8],aiming at the problem that how to apply the matrix computation to variable precision Rough set model,we have proposed a measure and computation approach based on matrix about studying Rough set theory.In this paper,we will provide a new insight into attributes reduction in decision table.A compression method of decision table based on matrix computation is proposed.We shall prove the feasibility of the compression method in theory and show its effectiveness in experiment and in an agriculture example.

2.Equivalence matrix of decision table

Definition2.1. Let U be a non-empty finite set of objects and R be equivalence relation on U,We denote the partition gived by the R as follows:U/R=$\{X_1, X_2,...X_n\}$

Definition2.2.Let R be equivalence relation on U,then R is expressed in terms of a binary n×n requivalence matrix $M_R=[m_{ij}]_{n\times n}$ where n=$|U|$, $m_{ij}=\begin{cases}1 & x_i R x_j \\ 0 & or\end{cases}$.

Definition2.3.Let $P=(p_{ij})_{n\times n}$,$Q=(q_{ij})_{n\times n}$ be two binary n×n matrices.The intersection P∩Q of matrices P and Q is defined as follows:$P\cap Q=[t_{ij}]_{n\times n}$,where $t_{ij}=\min\{p_{ij},q_{ij}\}$

Definition2.4. Let R={r_1,r_2,⋯,r_m},then $M_R = \prod_{i=1}^{m} M_{r_i}$

Definition2.5.LetC,D be equivalence relation and $M_C = (c_{ij})_{n\times n}$, $M_D = (d_{ij})_{n\times n}$be their matrices respectively.If for arbitrary positive integer i,j, $c_{ij} \leqslant d_{ij}$,then $M_C \le M_D$

Definition2.6.Let S=(U,R=C∪D,V,f) be a decision table,where C is condition attributes and D is decision attributes,Then S is consistent if and only if $M_C \le M_D$.

Definition2.7. Let S=(U,R=C∪D,V,f, $M_C \le M_D$) be a decision table,Given an attribute a∈C,then attribute a is nonsignificant in C if $M_{\{C-a\}} \le M_D$.

Definition2.8. Let S=(U,R=C∪D,V,f, $M_C \le M_D$) be a decision table.The set of all attributes $C'' \subseteq C$ which are significant in S is called the core set of C.

Definition2.9. Let S=(U,R=C∪D,V,f, $M_C \le M_D$) be a decision table.A subset T' of C is said to be a attribute reduction of C if and only if T' satisfies:(1) $M_{T'} \le M_D$;(2)if $T'' \subset T'$, then $: M_{T''} > M_D$.

3.Theory analysis of matrix computation

Put $[R]_{ij}$ express the element in row i and in column j of M_R,where R={r_1,r_2,⋯,r_m}.Obviously, $[R]_{ij}$ has the following properties:

(1) $[R]_{ij} = [r_1]_{ij} \wedge [r_2]_{ij} \wedge \cdots \wedge [r_m]_{ij}$(i=1,2⋯n,j=1,2⋯n);(2)if $[R]_{ij}$ =1,then for arbitrary $r_k \in R$, $[r_k]_{ij}$ =1;(3) if $[R]_{ij}$ =0,then there at least exist an attribute $r_k \in R$,such that $[r_k]_{ij}$ =0.Referring to properties,we immediately derive the following facts:

Theorem 3.1 Let S=(U,R=C∪D,V,f, $M_C \le M_D$) be a decision table. A subset T' of C is said to be a attribute reduction if and only if for any $[D]_{ij}$=0 in M_D, $[T']_{ij}$ satisfies:(1) $[T']_{ij}$=0;(2)There no exist $T'' \subseteq T'$, such that $[T'']_{ij}$=0.

Theorem 3.2 Let S=(U,R=C∪D,V,f, $M_C \le M_D$) be a decision table where C={c_1,c_2,⋯,c_m}.c∈C is core attribute if and only if there at least exist positive integer i,j,(i=1,2⋯n,j=1,2…n) such that $[D]_{ij}$=0, $[c]_{ij}$=0 , but for any b∈C-c, $[b]_{ij}$=1.

Proof.Let c∈C be core attribute.If there exist some attribute b∈C(c≠b),such that [b]ij=0,for any$[D]_{ij}$=0 ,$[C]_{ij}$=min {$[c_1]_{ij}$, $[c_2]_{ij}$,…$[c_m]_{ij}$}=0,then after deleting attribute b in C,we have $M_{\{C-c\}} \le M_D$.Thus there exist attribute set $C' \subseteq$ {C-c},such that C' is attribute reduction of S which contradict that c is core attribute in S.

Conversely,if there exist positive integer i,j,(i=1,2…n,j=1,2…n), such that $[D]_{ij}$=0 , $[c]_{ij}$=0 , and $[b]_{ij}$=1 for every b∈C-c,then after deleting attribute b in C,we have $[C]_{ij}$=0≠$[C-c]_{ij}$=1.Thus c is core attribute in S by theorem 3.1

Definition3.1 Let S=(U,R=C∪D,V,f, $M_C \le M_D$) be a decision table.If there exist positive integer i,j,such that attribute c satisfies $[c]_{ij}$=0, $[C-c]_{ij}$=1 when $[D]_{ij}$=0 ,$[C]_{ij}$=0,then attribute c is called local core attribute of decision table S.

Theorem 3.3 If c_1,c_2⋯,c_k be all local core attribute of S=(U,R=C∪D,V,f, $M_C \le M_D$),then core attribute set C' of decision table S is $\bigcup_{i=1}^{k} c_i$ .That is, $C' = \bigcup_{i=1}^{k} c_i$.

Definition3.2 Core attribute set C' of decision table S is called global core attributes

Definition3.3 Let S=(U,R=C∪D,V,f, $M_C \le M_D$) be a decision table and C' be core attribute set.If $a_k \in C\text{-}C'$ satisfies that there exist positive integer i,j,such that $[C']_{ij}$=1, $[C]_{ij}$=0 and $[a_k]_{ij}$=0,then attribute set $C' \cup \{a_k\}$ is called a local attribute reduction of decision table S.

Obviously, local attribute reduction derived by $[C]_{ij}$=0 can has not only one. All local attribute reduction derived by $[C]_{ij}$=0 is called a local attribute reduction set.

Definition3.4 Attribute reduction set $T \subset C$ of decision table is called global attribute reduction.

Theorem 3.4 Let $B_1,B_2\cdots,B_k$ be all local attribute reduction sets of decision table S=(U,R=C∪D,V,f,$M_C \le M_D$),where $B_i=\{A_{i1},A_{i2},\mathbf{L},A_{ik_i}\}(1\le i\le k)$ and $A_{ij}(1\le j\le i_k)$ is a local reduction.If T_t is attribute reduction,then T_t satisfies:(1)If for arbitrary positive integer i,j, $B_i\cap B_j=\phi$,then $T_t=\bigcup_{i=1}^{k}A_{ip},1\le p\le i_k$;(2)If there exist positive integer i,j, $B_i\cap B_j$ ϕ ,then $T_t=\bigcup A_{ip},1\le p\le i_k$ where for any $A_{im},A_{jn}\in T_t$, A_{im},A_{jn} must satisfy $A_{im}\notin B_j$ or $A_{jn}\notin B_i$.

Proof (1)Put $T_t=\bigcup_{i=1}^{k}A_{ip}$ and let C' be core attribute set.There at least exist an attribute $b\in T_t$ for any $[D]_{ij}=0,[C]_{ij}=0$,such that $[b]_{ij}=0$.That is $[T_t]_{ij}=0$. On the other hand,for arbitrary $B\subset T_t$, Suppose $B\cup c_k=T_t$ and $C'\cup c_k\in B_m$。 By definition3.2 there exist positive integer p,q,such that $[C']_{pq}=1,[c_k]_{pq}=0$. While for arbitrary i,j,if $B_i\cap B_j=\phi$， then $[B]_{pq}=[T_t-c_k]\ne 0$ By theorem 3.1,attribute set B is not attribute reduction of system S.

(2)If there exist positive integer i,j,$B_i\cap B_j\ne\phi$，then $T_t=\bigcup_{i=1}^{k}A_{ip},1\le p\le i_k$ is not the optimal attribute reduction of system S.It is fact that if $B_i\cap B_j=A'$ and $A'\ne A_{jm}\in B_j$,then $A'\cup A_{jm}\subseteq T_t'$ is not the optimal attribute reduction. Hence ,by above (1), if for any $A_{im}\in B_i,A_{jn}\in B_j$ and $A_{im},A_{jn}\in T_t$, A_{im},A_{jn} satify $A_{im}\notin B_j$ or $A_{jn}\notin B_i$ then $T_t=\bigcup A_{ip},1\le p\le i_k$ is the optimal attribute reduction.

4.Compression of decision table

Because equivalence matrices of attribute set are symmetric,in practical application ,we pay attentin to the upper-triangle above the diagonal or the lower- triangle below the diagonal.Let C={c_1, c_2,⋯, c_r} be condition attribute set and D be decision attribute set. For arbitrary positive integer i,j,we obtain: $[C]_{ij}=[c_1]_{ij}\wedge[c_2]_{ij}\wedge\cdots\wedge[c_r]_{ij}$.By theorem 3.1,in practical application,we only devote our attention to $[D]_{ij}=0$ and $[C]_{ij}=0$ in the upper-triangle above the diagonal or the lower- triangle below the diagonal. In this paper,we compress information of $[D]_{ij}=0$ and $[C]_{ij}=0$ in the equivalence matrix into a binary tree where $[c_1]_{ij},[c_2]_{ij},\cdots\wedge[c_r]_{ij}$ are orderly arranged and nodes of the tree.

Please refer to below for details.First,create the root of the tree,labeled with “null”.Scan the elements of M_D a time and find all positive integer i,j which satisfy $[D]_{ij}=0$.The corresponding elements of each equivalence relation matrix of condition attributes orderly sorted lead to the construction of the first branch of the tree with r nodes where $[c_1]_{ij}$ is linked as a child of the root, $[c_2]_{ij}$ is linked to $[c_1]_{ij}$.The rest may be deduced by analogy.

The second $[c_1]_{mn},[c_2]_{mn},\cdots\wedge[c_r]_{mn}$would result in a branch where $[c_1]_{mn}$ is linked as a child of the root, $[c_2]_{mn}$ is linked to $[c_1]_{mn}$.the rest may be deduced by analogy.Howeve,this branch would share an existing path with other branchs if along the root node ,some branch has the common prefix.For example,if $[c_1]_{ij}=[c_1]_{mn},[c_2]_{ij}=[c_2]_{mn},[c_3]_{ij}\ne[c_3]_{mn}$,then the first two nodes of the branch which contains $[c_1]_{ij}$, $[c_2]_{ij}\wedge\cdots\wedge[c_r]_{ij}$ is the same as the branch which contains$[c_1]_{mn}$, $[c_2]_{mn}\wedge\cdots\wedge[c_r]_{mn}$.The rest branches may be constructed by analogy.

By theorem 3.2 ,definition 3.2,all local core attributes and all local attribute reducion are derived from these branches.By theorem 3.3,3.4, we get global core attributes and global attribute reduction.

5.Description of attribute reduction algorithm

Let S=(U,R=C∪D,V,f,$M_C \le M_D$) be a decision table.

Step1:Compute the equivalence matrices of decision attribute set and each of condition attributes ,and arrage the equivalence matrices of each of condition attributes in ordor;

Step2:According to $[D]_{ij}=0$,compress $[c_1]_{ij},[c_2]_{ij},\cdots\wedge[c_r]_{ij}$ into a binary tree where $[c_k]_{ij}$,(k=1,2,⋯r)is node of the

tree.

Step3:Scan the tree and find the only zero value node in every branch.We get local core of every branch. Core set of system S is union of all local core.

Step4: Prune the branch that includes local core and at the same time, retain shareable prefix part.

Step5: Travel every branch of binary tree pruned and find local attribute reduction sets of all branchs

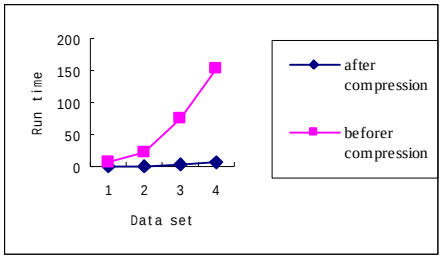
Step6: Compute attribute reduction of system S according to $\{T_t | T_t = \bigcup A_{ip}, 1 \le p \le i_k\}$

6.Algorithm analysis

The time complexity of algorithm in [7] for finding all core attributes is $o(|C|^2|U|^2)$, and for finding all attribute reduction is $o(2^{|C|}|C||U|^2)$. The time complexity of algorithm in this paper for finding all core attributes is at most $o(2^{|C|+1})$, and for finding all attribute reduction is at most $o(2\times 2^{|C|+1})$. In general, $|U| >> |C|$, hence,the method presented in this paper has an advantage over the method in [7].

Next,to compare the two methods,We made the relevant experiments on monks datasets in UCI database.The datasets have one decision attribute ,six conditon attributes and 423 records.We did four experiments with the first 100,150,300,423 recordes of standard datasets monks datas. The experiment environment is Petium4 2.1GMHZ,RAM512M, windows XP .The results are as follows:

Table 1 Run time of two algorithms



The chart shows that the method presented in this paper is efficient and scalable for finding core attributes set and attribute reduction sets ,and is faser than the method in [7].

7.An Application Example in Agriculture

The following is knowledge representation system of cotton diseases.The condition attributes are a-“diseased spot color”, b-“disease site”,c-“disease shape”,d-“feature” and the decision attribute is e-“the type of disease”.{1,2,3} represent the different value of each attribute.

Table 2 Knowledge representation system

U	a	b	c	d	e
1	3	2	3	2	2
2	3	2	3	2	2
3	1	1	1	1	2
4	2	2	3	2	3
5	2	2	3	2	3
6	2	1	3	3	1
7	3	1	2	2	1
8	3	2	3	2	2
9	1	1	1	1	2

According to the method proposed in this paper,we can obtain that the core attribute is {a} and the attribute reduction set are {a,b} and {a,c,d}.The conclusion is the same as the other method.That is to say that diseased spot is chief factor to judge the type of disease and diseased spot color, disease site or diseased spot color, disease shape,

feature may judge exactly the type of disease.

8.Conclusion

In this paper,We further discuss the approach of matrix computation about Rough set and applied it to dicision table.In theory,we have proved the relation between the elements of equivalence matrice and core attributes,attribute reduction .At the same time,we suggest an attribute reduction algorithm based on a storage structure of binary tree which can compress the invalid and the same prefix information.The algorithm designed is lower price.we also find that by changing the order of condition attributes sorted,the algorithm is more efficient.

References

[1] Z.Pawlak.Rough Sets[J].International Jounal of Information and Computer Science,l982,11(5),pp341-356.

[2]Z.Pawlak.Rough Sets:Theoretical Aspects of Reasonimg about data,Kluwer Academic Publishers,Dordrecht,Netherlands,1991

[3]Wang G Y, Hu F,Huang H, Wu Y. A granular computing model based on tolerance relation.The Journal of China Universities of Post s and Telecommunications,2005,12(3):86-90.

[4]S.Greco,A.G.Wojna,R.Slowinski.Fuzzy rough sets and multiple-premise gradualdecision rules,International Journal of Approximate Reasoning 41(2)(2006)179-211.

[5]Lin T Y,Yin P.Heuristically fast finding of the shortest reducts//Proceeding of the Rough Sets and Current Trends in Computing(RSCT2004).Uppsala,Sweden,2004:465-470

[6]J.W.Guan,D.A Bell. Matrix computational method for information systems[J].Artificial Intelligence,105(1998)77-103.

[7] J.W.Guan,D. A Bell, Z.Guan.Matrix computation for information systems[J].Information Sciences,131(2001)129-156.

[8]Laipeng LUO,Ergen LIU,Yi Ceng.Matrix approach to the study of rough set theory.Systems Engineering and Electronics,31 (4)2009,pp859-862(in chinese).