Journal of Zhejiang University SCIENCE A
ISSN 1009-3095 (Print); ISSN 1862-1775 (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Immune algorithm for discretization of decision systems in rough set theory*

JIA Ping, DAI Jian-hua†, CHEN Wei-dong, PAN Yun-he, ZHU Miao-liang

(*Institute of Artificial Intelligence, Zhejiang University, Hangzhou 310027, China*)

†E-mail: jhdai@zju.edu.cn

Received July 8, 2005; revision accepted Nov. 20, 2005

Abstract: Rough set theory plays an important role in knowledge discovery, but cannot deal with continuous attributes, thus discretization is a problem which we cannot neglect. And discretization of decision systems in rough set theory has some particular characteristics. Consistency must be satisfied and cuts for discretization is expected to be as small as possible. Consistent and minimal discretization problem is NP-complete. In this paper, an immune algorithm for the problem is proposed. The correctness and effectiveness were shown in experiments. The discretization method presented in this paper can also be used as a data pre-treating step for other symbolic knowledge discovery or machine learning methods other than rough set theory.

Key words: Rough sets, Discretization, Immune algorithm, Decision system
doi:10.1631/jzus.2006.A0602 **Document code:** A **CLC number:** TP18

INTRODUCTION

As a new effective mathematical tool to deal with vagueness and uncertainty, the rough set theory was first proposed by Pawlak in 1982. Theoretical research on rough set theory is characterized by many achievements (Dai, 2004b; 2005; Dai *et al*., 2004). In recent years rough set theory has been successfully applied in many fields such as machine learning, pattern recognition, decision support and data mining.

But rough set theory cannot deal with continuous attributes although a very large proportion of real datasets include continuous variables. One solution to this problem is to partition numerical variables into a number of intervals and treat each interval as a category. This process of partitioning continuous variables into categories is usually termed discretization. Discretization in rough set theory has new content when considering indiscernibility relation. Generally speaking, we should seek possibly minimum number of discrete intervals, and at the same time should not weaken the indiscernibility ability. In other words, we want to get consistent and minimum discretization. Dai (2004a) presented a genetic algorithm for the problem.

Immune algorithm (IA) (Chun *et al*., 1997; 1998) is a new optimization algorithm imitating the immune system to solve many problems in the real world. In this paper, we propose an immune algorithm for discretization of decision systems in rough set theory.

DESCRIPTION OF DESCRETIZATION BASED ON ROUGH SET THEORY

Let $S=<U, A\cup d, V, f>$ be a decision system, while $U=\{x_1, x_2, \ldots, x_n\}$ is objects set, $A=\{a_1, a_2, \ldots, a_k\}$ is condition attributes set, d is decision attribute. For any $a\in A$, there is information mapping $U\rightarrow V_a$, where V_a is the value domain. We assume $V_a=[l_a, r_a]\subset\mathbb{R}$. We also assume that S is a consistent decision system in this paper.

* Project supported by the National Basic Research Program (973) of China (No. 2002CB312106), China Postdoctoral Science Foundation (No. 2004035715), the Science & Technology Program of Zhejiang Province (No. 2004C31098), and the Postdoctoral Foundation of Zhejiang Province (No. 2004-bsh-023), China

Definition 1 Any pair (a, c), where $a\in A$ and $c\in\mathbb{R}$, defines a partition of V_a into left-hand-side and right-hand-side interval. And the pair (a, c) is called a cut on V_a.

Let us fix an attribute $a\in A$. Any set of cuts

$$D_a=\{(a, c_1^a), (a, c_2^a), \ldots, (a, c_{k_a}^a)\}, \quad (1)$$

where $k_a\in\mathbb{N}$, and $l_a=c_0^a<c_1^a<c_2^a<\ldots< c_{k_a}^a < c_{k_a+1}^a=r_a$, defines a partition on V_a into sub-intervals, i.e.,

$$V_a=[c_0^a, c_1^a)\cup[c_1^a, c_2^a)\cup\ldots\cup[c_{k_a}^a, c_{k_a+1}^a]. \quad (2)$$

So, any set of cuts on condition attributes $D=\bigcup_{a\in A} D_a$ transforms the original decision system S into discrete decision system $S^D=<U, A\cup d, V^D, f^D>$, where $f^D(x_a)=i\Leftrightarrow f(x_a)\in[c_i^a, c_{i+1}^a)$, $x\in U$, $i\in\{0, 1, \ldots, k_a\}$. After discretization, the original decision system is replaced with the new one. And different sets of cuts will construct different new decision systems.

It is obvious that discretization process is associated with loss of information. Usually, the task of discretization is to determine a minimal set of cuts from a given decision system and keeping the discernibility between objects. And the rationality of the selected cuts can be evaluated by the following criteria (Nguyen, 1997; 1998):

(a) Consistency. For any objects u, $v\in U$, satisfying: if (u, v are discerned by A) then (u, v are discerned by D).

(b) Minimum. There is no $D'\subset D$, satisfying the consistency.

(c) Optimality. For any D' satisfying consistency, it follows that $card(D)\leq card(D')$, then D is optimal cut.

Theorem 1 Optimal Discretization Problem is NP-complete (Nguyen, 1997).

IMMUNE ALGORTIHM FOR DISCRETIZAION OF DECISION SYSTEMS IN ROUGH SET THEORY

Immune algorithm (IA) is a kind of effective searching and optimizing technique and has been applied to various fields. The immune system is a basic defense system against bacteria, viruses and other disease-causing organisms and has dramatic and complex mechanisms that recombine genes to cope with the invading antigens, producing antibodies against antigens. By using this mechanism, IA performs well as an optimization algorithm. We design an immune algorithm for optimal discretization problem mentioned above.

Frame of algorithm:
determine candidate cuts
construct the new decision table $S^*=<U^*, R^*, V^*, f^*>$ from the original decision table $S=<U, A\cup d, V, f>$
initialize the antigen and antibodies $pop(t)$;
t=1;
while (not terminate) do
{
  work out the affinities between antigen and antibodies in $pop(t)$;
  work out the affinities between antibodies and the density values in $pop(t)$;
  change memory cells $mem(t)$;
  crossover $pop(t)$;
  mutate $pop(t)$;
  $t=t$+1;
}

The antigen and antibody of any immune system correspond to objective function and optimal solution of immune algorithm respectively. Memory cells $mem(t)$ constitute a part of the population $pop(t)$.

Determination of candidate cuts

Let $S=<U, A\cup d, V, f>$ be a decision system. An arbitrary condition attribute $a\in A$, defines a sequence $v_1^a < v_2^a < \ldots < v_{n_a}^a$, where $\{v_1^a, v_2^a, \ldots, v_{n_a}^a\} = \{a(x): x\in U\}$, then the set of all possible cuts on a is defined by:

$$C_a=\left\{\left(a, \frac{v_1^a+v_2^a}{2}\right), \left(a, \frac{v_2^a+v_3^a}{2}\right), \ldots, \left(\frac{v_{n_a-1}^a+v_{n_a}^a}{2}\right)\right\}. \quad (3)$$

The set of possible cuts on all attributes is denoted by:

$$C_A=\bigcup_{a\in A} C_a. \quad (4)$$

Definition 2 Let us assume that the objects in U are sorted in ascending order over the values of a, (a, c) is a cut of a, (a, c) is called a bound cut if and only if the following are satisfied:

(1) $\exists x_i, x_j \in U$, $d(x_i) \neq d(x_j)$, satisfying $a(x_i) < c < a(x_j)$;

(2) There is no $x \in U$ between x_i and x_j, satisfying $a(x_i) < a(x) < a(x_j)$.

Theorem 2 Let S be a decision system, assuming that the set of bound cuts of cuts C_A defined by Eq.(2) is BC_A, BC_A can discern all the objects from different decision classes.

Based on Theorem 2, we get the candidate cuts BC_A constituting the set of all the bound cuts in C_A. The Theorem 2 details can be found in (Dai and Li, 2002).

Construct the new decision table S^* from S

Let $S=<U, A\cup d, V, f>$ be a decision system, and construct the new decision system $S^*=<U^*, BC_A, V^*, f^*>$ by the following:

$U^*=\{(x_i, x_j)\in U\times U | d(x_i)\neq d(x_j)\}$

$BC_A=\{P_r^a | a\in A\}$, P_r^a is the rth bound cut (a, c_r^a) on attribute a

For any P_r^a, if $c_r^a \in \{\min[a(x_i), a(x_j)], \max[a(x_i), a(x_j)]\}$, then $f^*(P_r^a, (x_i, x_j))=1$,

else $f^*[P_r^a, (x_i, x_j)]=0$

To describe the procedure above clearly, we discuss an example decision table (Table 1) in which the condition attribute set $C=\{a_1, a_2\}$ and the decision attribute is d. It is easy to find that the two condition attributes must be discretized.

Table 1 A decision system

U	a_1	a_2	d
x_1	0.8	2	1
x_2	1.0	0.5	0
x_3	1.3	3	0
x_4	1.5	1	1
x_5	1.4	2	0
x_6	1.6	3	1
x_7	1.3	1	1

Based on the discussion above, we can get the possible cuts and the candidate cuts as follows:

$C_A=\{(a_1, 0.9), (a_1, 1.15), (a_1, 1.35), (a_1, 1.45), (a_1, 1.55), (a_2, 0.75), (a_2, 1.5), (a_2, 2.5)\}$,

$BC_A=\{(a_1, 0.9), (a_1, 1.15), (a_1, 1.35), (a_1, 1.45), (a_2, 0.75), (a_2, 1.5), (a_2, 2.5)\}$.

By using the candidate cuts BC_A, we get the new decision table (see Table 2) from the original decision system (Table 1).

Table 2 The new decision table for Table 1

U^*	$P_1^{a_1}$	$P_2^{a_1}$	$P_3^{a_1}$	$P_4^{a_1}$	$P_1^{a_2}$	$P_2^{a_2}$	$P_3^{a_2}$
(x_1, x_2)	1	0	0	0	1	1	0
(x_1, x_3)	1	1	0	0	0	0	1
(x_1, x_5)	1	1	1	0	0	0	0
(x_2, x_4)	0	1	1	1	1	0	0
(x_2, x_6)	0	1	1	1	1	1	1
(x_2, x_7)	0	1	0	0	1	0	0
(x_3, x_4)	0	0	1	1	0	1	1
(x_3, x_6)	0	0	1	1	0	0	0
(x_3, x_7)	0	0	0	0	0	1	1
(x_4, x_5)	0	0	0	1	0	1	0
(x_5, x_6)	0	0	0	1	0	0	1
(x_5, x_7)	0	0	1	0	0	1	0

Representation of antibodies

To calculate the minimal cuts is to find the minimal subset maintaining the discretization ability of the whole N candidate cuts. It is easy to represent an antibody as a binary string of length N, where N is the number of candidate cuts, i.e. $N=card(BC_A)$. 1 means that the corresponding attribute is present, and 0 means it is not.

Taking Table 2 as an example, we have 7 candidate cuts $\{P_1^{a_1}, P_2^{a_1}, P_3^{a_1}, P_4^{a_1}, P_1^{a_2}, P_2^{a_2}, P_3^{a_2}\}$ i.e. $\{(a_1, 0.9), (a_2, 1.15), (a_1, 1.35), (a_1, 1.45), (a_2, 0.75), (a_2, 1.5), (a_2, 2.5)\}$. Antibody 0111010 represents the determined cuts set as follows:

$\boldsymbol{D}=\{(a_1, 1.15), (a_1, 1.35), (a_1, 1.45), (a_2, 1.5)\}$

Computing of affinities and densities

Let $Ab_k=<Ab_{k1}, Ab_{k2}, \ldots, Ab_{kN}>$ be an antibody, where $Ab_{kl}=0$ or 1, $l=1,2,\ldots,N$. Then the affinity between Ab_i and Ab_j is defined as:

$$Aff(Ab_i, Ab_j) = 1 - \left(\sum_{n=1}^{N} | Ab_{in} - Ab_{jn} | \right) / N. \qquad (5)$$

According to the definition of optimal discretization, we know that the fitness function depends on the number of cuts (which we wish to keep as low as

possible) and the consistency (which we wish to keep as high as possible). So, we can define the affinity between antigen Ag and an antibody Ab_k as:

$$Aff_{Ag}(Ab_k)=\alpha\cdot\left(N-\sum_{n=1}^{N}Ab_{kn}\right)/N+\beta\cdot\frac{D_{Ab_k}}{card(U^*)},\quad(6)$$

where Ab_{kn} is the nth bit of antibody Ab_k, D_{Ab_k} is the objects pairs set discerned by antibody Ab_k, $card(U^*)$ is the number of the objects pairs in the new decision system S^*, α and β are weight factors. Based on the definition of the affinity between antigen and antibodies, we can define the density $Den(Ab_k)$ of an antibody Ab_k as

$$Den(Ab_k)=\frac{card(\{Ab_l\mid Aff(Ab_k,Ab_l)\geq\mu\})}{\#antibodies},\quad(7)$$

i.e., the proportion of the antibodies Ab_l satisfying $Aff(Ab_k,Ab_l)\geq\mu$, l=1,2,…,$\#antibodies$, $0.8\leq\mu\leq1$, to the whole antibody population.

Crossover and mutation

We use classical, one-point crossover process that affects antibody selected to reproduce with probability of P_c.

We use strategy self-adaptive mutation rate in mutation process. By this strategy, we use a higher probability of mutating from "1" to "0" than that in the opposite direction, since our goal is minimal discretization.

Let c_i be an antibody. We define $ones(c)$ is the number of "1" in c_i. And we define the mutation rate of an antibody c_j as:

$$mp_j=ones(c_j)/\sum_{i=1}^{popsize}ones(c_i),\quad(8)$$

where mp_j (j=1, 2, …, $popsize$) is the mutation probability.

EXPERIMENTAL STUDY

Nguyen (1997) proposed the named discretization approach based on rough set methods and Boolean reasoning. The main idea is to find possibly minimum number of discrete intervals, and at the same time not weaken the indiscernibility ability. In this section, we will make comparative experiment between our algorithm and Nguyen (1997)'s method.

From Table 1's decision system, we get the following result cuts by Nguyen's method:

$$D=\{(a_1, 1.15), (a_1, 1.35), (a_1, 1.45), (a_2, 1.5)\}.$$

The resulting 4 cuts and the discretized result are shown in Table 3.

Table 3 The discretized result of Table 1 by Nguyen's method

U	a_1	a_2	d
x_1	0	1	1
x_2	0	0	0
x_3	1	1	0
x_4	3	0	1
x_5	2	1	0
x_6	3	1	1
x_7	1	0	1

We also discretized Table 1's decision system by our method. The number candidate cuts determined the string length as 7. The population size, the size of memory cells, the parameter μ and the crossing-over rate P_c were 20, 5, 0.9 and 0.7 respectively. The weight factors were set as 1 and 2.5 respectively. The algorithm stops when there was no variation of the average fitness in certain number of generations. In the experiment, the antibodies all tended to be identical (0101010) and the corresponding cuts set was:

$$D=\{(a_1, 1.15), (a_1, 1.45), (a_2, 1.5)\}.$$

As there were only 3 cuts, the result is supposed to be the minimal one for this example, with the discretized result being as Table 4.

Table 4 The discretized result of Table 1 by our method

U	a_1	a_2	d
x_1	0	1	1
x_2	0	0	0
x_3, x_5	1	1	0
x_4	2	0	1
x_6	2	1	1
x_7	1	0	1

From the comparison between Table 3 and Table 4 we know that our method can get smaller discretized cuts and that the discretized result is smaller too.

DISCUSSION

Discretization based on rough set theory has new content when considering indiscernibility relation. Generally speaking, we should seek possibly minimum number of discrete intervals, and at the same time not weaken the indiscernibility ability. An immune algorithm for consistent and minimal discretization of decision system is proposed. The effectiveness is shown in the experiments.

References

Chun, K.S., Jung, H.K., Yahn, S.Y., 1997. Shape optimization of electromagnetic devices using immune algorithm. *IEEE Tran. Magnetics*, **33**(2):1876-1879. [doi:10.1109/20.582650]

Chun, J.S., Jung, H.K., Yahn, S.Y., 1998. A study on comparison of optimizing performances between immune algorithm and other heuristic algorithms. *IEEE Tran. Magnetics*, **34**(5):2972-2975. [doi:10.1109/20.717694]

Dai, J.H., 2004a. A Genetic Algorithm for Discretization of Decision Systems. Proceedings of the 3rd International Conference on Machine Learning and Cybernetics, IEEE Press, New Jersey, p.1319-1323.

Dai, J.H., 2004b. Structure of Rough Approximations Based on Molecular Lattices. Proceedings of the 4th International Conference on Rough Sets and Current Trends in Computing (RSCTC2004), LNAI 3066. Uppsala, Sweden, p.69-77.

Dai, J.H., 2005. Logic for Rough Sets with Rough Double Stone Algebraic Semantics. Proceedings of the Tenth International Conference on Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing (RSFDGrC 2005), LNAI 3641, Regina, Canada, p.141-148.

Dai, J.H., Li, Y.X., 2002. Study on Discretization Based on Rough Set Theory. Proceedings of the First International Conference on Machine Learning and Cybernetics, IEEE Press, New Jersey, p.1371-1373.

Dai, J.H., Chen, W.D., Pan, Y.H., 2004. A minimal axiom group of rough set based on quasi-ordering. *Journal of Zhejiang University SCIENCE*, **5**(7):810-815. [doi:10.1631/jzus.2004.0810]

Nguyen, S.H., 1997. Discretization of Real Value Attributes: Boolean Reasoning Approach. Ph.D Thesis, Warsaw University, Poland.

Nguyen, S.H., 1998. Discretization Problems for Rough Set Methods. *In*: Polkowski, L., Skowron, A.(Eds.), Proceedings of the First International Conference on Rough Sets and Current Trend in Computing (RSCTC'98), Lecture Notes on Artificial Intelligence, Springer-Verlag, Berlin, **1424**:545-552.