第33卷 第23期 计 算 机 工 程 2007年12月
Vol.33 No.23 Computer Engineering December 2007
·软件技术与数据库· 文章编号：1000—3428(2007)23—0051—03 文献标识码：A 中图分类号：TP18

基于免疫原理的粗糙集属性约简

张 旭[1,2]，郭 晨[2]
(1. 大连交通大学机械工程学院，大连 116028；2. 大连海事大学自动化与电气工程学院，大连 116026)

摘 要：在基于粗糙集理论的知识发现中，属性约简是其中重要的研究内容之一，已经被证明是NP完全问题。基于生物免疫原理，提出了一种新型粗糙集属性约简算法。该算法由记忆细胞获取、克隆选择、超变异和群体更新4种算子构成。算法设计的重点在于将分类精度和约简中所含属性个数集成为一个统一的亲合度成熟目标，并通过抗体更新和抗体相似性抑制来维持群体的多样性，以获得多个符合分类质量要求的属性约简集。实验结果证明了该算法的有效性。

关键词：免疫原理；粗糙集；属性约简

Reduction of Rough Set Attribute Based on Immune Mechanism

ZHANG Xu[1,2], GUO Chen[2]
(1. School of Mechanical Engineering, Dalian Jiaotong University, Dalian 116028;
2. School of Automation and Electrical Engineering, Dalian Maritime University, Dalian 116026)

【Abstract】 Attribute reduction that has been proved to be a NP-hard problem is one of the important issues of the KDD based on the rough set theory. A novel attribute reduction algorithm of rough set based on the vertebrate immune mechanism is proposed. The main operators of the algorithm include memory cells producing, clone selection, hyper-mutation and population updating. The key of its design is to integrate discernible ability and the elements in the condition attribute set into one unified affinity maturation objection. The different attribute reduction sets that can maintain the ability of classification can be found through maintaining the diversity of antibody population with renewal of antibody and similar antibodies suppression. The experimental results show that it is effective.

【Key words】 immune mechanism; rough set; attribute reduction

1982年，波兰数学家Z. Pawlak提出了粗糙集(rough set)理论[1]，并于1991年形成理论体系。该理论近年来得到了迅速发展，目前已在机器学习、人工智能、模式识别、图像处理和传感数据分析等众多领域得到应用，为数据挖掘和知识发现领域提供了一种有效而新颖的理论。

属性约简是粗糙集理论中的一个核心部分[2]，其计算的复杂性随着数据表的增大呈指数增长，已经证明它是一个NP-hard问题[3]，现有的如基于信息熵、基于差别矩阵等算法虽然取得了相当的成效，但到目前为止，还没有一个公认的、高效的属性约简算法[4~5]。另外这些方法都没有考虑到数据领域知识的特殊性以及用户需求的灵活性。如在机器诊断的应用中，某些特征属性的获取困难或计算过于复杂，包含这些特征的属性约简集合不一定是最优选择。因此，多个属性约简集合的获取是非常必要的，再由领域专家根据实际情况来进行优化选择，可提高粗糙集理论的实际应用能力。

人工免疫系统是近年来的新兴智能系统，具有高效的分布式自学习性，既能实现抗体的快速优化，又能保持不同抗体和谐并存，其独特的工作机理对开发高效的多模态优化算法具有借鉴意义。因此，本文利用基于免疫原理的多模态优化策略来获得粗糙集的多属性约简集合，并给出了实际的应用效果。

1 粗糙集基本概念

定义1 四元组$S=(U, A, V, f)$是一个知识表达系统，其中，U表示对象的非空有限集合，称为论域；$A=C\cup D$，$C\cap D=\varnothing$，C称为条件属性集，D称为决策属性集；$V=\bigcup_{a\in A}V_a$是属性a的值域；f表示$U\times A\to V$是一个信息函数，它为每个对象的每个属性赋予一个信息值，即：$\forall a\in A$，$x\in U$，$f(x,a)\in V$。将具有条件属性和决策属性的知识表达系统称为决策表。

每个属性子集$P\subseteq A$决定了一个二元不可区分关系$IND(P)$：

$$IND(P)=\{(x,y)\in U\times U\,|\,\forall a\in A, f(x,a)=f(y,a)\} \quad (1)$$

关系$IND(P)$ $(P\subseteq A)$构成了U的一个划分，用$IND(P)$表示，简记为U/P。$U/IND(P)$中的任何元素$[x]_P$称为等价类。$U/IND(P)$表示知识表达系统$S=(U,A,V,f)$中与P相关的知识，是最小不可识别对象集，也体现该知识表达系统的最高辨别能力。

定义2 设$S=(U,A,V,f)$是一个知识表达系统，对于每个子集$X\subseteq U$和一个等价关系$R\subseteq A$，称子集$\underline{R}X=\{x\in U\,|\,[x]_R\subseteq A\}$为$X$的$R$下近似集。

X的R下近似集也称作X的R正域，记作$POS_R(X)$，即$POS_R(X)=\underline{R}X$。$\underline{R}X$或$POS_R(X)$是由那些根据知识R判定肯定属于X的U中元素组成的集合。

基金项目：国家自然科学基金资助项目(60474014)；教育部高等学校博士学科点专项基金资助项目(20040151007)

作者简介：张 旭(1972－)，男，博士研究生，主研方向：免疫优化计算，数据挖掘，复杂机械的智能故障诊断；郭 晨，教授、博士生导师

收稿日期：2006-12-07 **E-mail：**fspuzx@newmail.dlmu.edu.cn

定义 3 设 $S=(U,A,V,f)$ 是一个知识表达系统，$P,Q\in A$，则称子集 $POS_R(Q)=\bigcup_{X\in U/Q}\underline{P}X$ 为 Q 的 P 正域。Q 的 P 正域是 U 中所有根据分类 U/P 的信息可以准确地划分到关系 Q 的等价类中去的对象集合。

定义 4 设 $S=(U,A,V,f)$ 是一个知识表达系统，$P,Q\in A$，$R\in P$，如果 $POS_P(X)=POS_{(P-\{R\})}(X)$，则称 R 为 P 中 Q 不必要的；否则 R 为 P 中 Q 必要的。

不必要属性在知识表达系统中是多余的，若将它从系统中去掉，不会改变系统分类能力。

定义 5 设 $S=(U,A,V,f)$ 是一个知识表达系统，$P,Q\in A$，如果每个 $R\in P$ 在 P 中都是 Q 必要的，则称 P 为 Q 独立的；否则，称 P 为 Q 依赖的。

对于相依赖的属性来说，其中包含有多余的属性，可以对其约简。

定义 6 设 $S=(U,A,V,f)$ 是一个知识表达系统，$P,Q\in A$，P 中所有 Q 必要的原始属性构成的集合称为 P 的 Q 核，简称相对核，记为 $core_Q(P)$。

定义 7 设 $S=(U,A,V,f)$ 是一个知识表达系统，$P,Q\in A$，$K\subseteq P$，如果：(1) $core_K(Q)=core_P(Q)$；(2) K 是 Q 独立的，则称 K 是 P 的一个 Q 约简。P 的 Q 约简也简称为相对约简；显然，相对约简不是唯一的。可以证明相对核是所有相对约简的交集。

定义 8 设 $S=(U,A,V,f)$ 是一个知识表达系统，$A=C\cup D$，$C\cap D=\varnothing$，C 为条件属性集，D 为决策属性集；若 $U/C=\{X_1,X_2,\cdots,X_n\}$，$U/D=\{Y_1,Y_2,\cdots,Y_m\}$，则决策属性 D 关于条件属性 C (或称 C 对 D)的依赖度定义如下：

$$k=\frac{1}{|U|}\sum_{i=1}^{m}|\underline{R}Y_i|=\frac{1}{|U|}\sum_{i=1}^{m}|POS_C(Y_i)|,\quad Y_i\in U/D \tag{2}$$

其中，$|\bullet|$ 表示集合包含的元素个数。k 表明知识 C 对整体决策 U/D 的依赖程度，通常 $0\le k\le 1$。若 $k=1$，则称知识 D 完全依赖于知识 C，或者说对于全域的所有元素都能用 C 来分类于 U/D 的概念中；若 $0<k<1$，则称知识 D 部分依赖于知识 C，或者说在已知条件 C 下，只能将 U 上那些属于正区域的个体分类于 U/D 的概念中；若 $k=0$，则称知识 D 全不依赖于知识 C，或者说全域中的所有元素都不能利用条件 C 分类于 U/D 的概念中。

2 基于免疫原理的粗糙集属性约简

2.1 免疫原理

免疫系统[6]是一个分布式、自组织和具有动态平衡能力的自适应复杂系统。其主要运行机制有：(1)免疫响应。随着对高亲和度抗体(Ab)的免疫正选择和抗原(Ag)驱动的免疫细胞超变异的循环过程，抗体对抗原的亲和度不断增升直至达到亲和度成熟。(2)免疫记忆。免疫系统通过学习抗原产生优秀抗体，并对有优良特性的抗体给予奖励(Baldwin效应)，利用克隆选择机制产生记忆细胞。(3)免疫调节。免疫系统内部各免疫细胞之间形成一个相互作用的动态平衡网络。当有外界抗原入侵时，通过免疫调节，达到新的免疫平衡；在无抗原入侵时，抗体间的相互促进与抑制作用可维持适当数量的必要抗体，以维持免疫平衡。抗体一旦产生，将分组进入淋巴系统实施进化，直至达到亲合度成熟，即抗体与抗原达到最佳匹配。从上述工作机理看，免疫优化实质上是一种并行分布的局部优化方式。免疫系统利用这种方式最终达到各种优化模式共生的目的。

2.2 编码方式

求最小属性约简就是在 N 个条件属性中寻找最小的属性子集。编码方案采用长度为 N 的二进制串表示每个抗体，每一个位对应一个条件属性，1 表示所选子集中含有对应属性，0 则表示不含对应属性。例如：假设有 10 个条件属性 $\{a_1,a_2,\cdots,a_{10}\}$，有一个可能的约简为 $\{a_2,a_3,a_5,a_8\}$，则它应表示为 $v=0110100100$。

2.3 亲合度函数的确定

由属性约简的定义可知，抗体的亲合度主要取决于两个方面：所含条件属性的个数 n 和决策属性对其依赖度 k。对某一属性子集而言，其包含属性个数越少，决策属性对其依赖度越大，则越有可能成为最小约简。因此，求解最小约简实际上是一个多目标优化问题，而多目标优化大多都具有多个 Pareto 最优解。构造亲合度函数如下：

$$f(R)=\alpha\square k+(1-\alpha)\frac{card(C)-card(R)}{card(C)} \tag{3}$$

其中，k 为决策属性对该个体的支持度；$card(C)$ 为条件属性的总个数；$card(R)$ 为个体中包含的条件属性个数；α 为调整参数，用以调整分类质量以及个体中属性数对亲合度的贡献程度。

求解知识表达系统中最小相对约简，实际就是要在保持整体条件属性支持度不变的情况下寻找所含条件属性最少的约简，而笔者构造的亲合度函数恰好从这两方面满足了问题的求解要求。这样，就将多目标优化问题转化为单目标优化问题。

2.4 变异算子

变异算子采用单点变异，同时按一定的启发信息进行变异。即当个体的依赖度 $k=1$ 时，说明该个体是一个可行解，但不一定是最优解，此时应尽量减少个体中的属性个数。所以只进行从“1”到“0”的变异；当个体的依赖度 $k<1$ 时，说明该个体还不是一个可行解，则随机地选择参与变异的个体的位，对该位取反。

2.5 算法描述

Step1 随机产生初始抗体集合，抗体个数为 p；设定终止条件，这里为达到最大进化代数 S 或记忆细胞集中亲合度最高个体连续 10 代未发生变化。

Step2 计算初始抗体集中每个抗体的亲合度。

Step3 从中选取亲和度高且编码各异的抗体，克隆复制到记忆细胞库，记忆细胞集规模为 m。

Step4 对每个记忆细胞进行克隆操作，产生临时抗体集 C；再对临时抗体集 C 进行变异操作，产生新抗体集 D。

Step5 计算新抗体集 D 中每个抗体的亲合度。

Step6 对新抗体集 D 联同原记忆细胞集进行抗体抑制，清除相似抗体，保留 m 个亲合度高的抗体形成新的记忆细胞集。

Step7 判断程序终止条件，满足则输出记忆细胞，其中依赖度 $k=1$ 的记忆细胞即为优化的最简属性约简集合，算法结束；否则补充一定数量的随机抗体进入记忆细胞集，转至 Step 4。

3 实例分析

3.1 实例 1

文献[7]给出了某电厂汽轮发电机组汽轮机轴系某测点下由振动状态构成的知识表达系统，它包括低频故障(油膜涡动)、高频故障(不平衡、不对中和动静碰摩)和正常状态等 5 种状态。每组样本利用频谱特征提取了 8 个条件属性和 1 个决策属性 D，其中条件属性 $\{x1,x2,\cdots,x8\}$ 分别表示测点振动

数据在频谱区间(0～014、014～016、016～110、1、2、3、4倍频和大于4倍频)中的最大幅值。采用了模糊C均值聚类方法进行离散化处理，将每个属性离散成高、中、低3种状态(3、2、1)。

如表1所示，对决策表采用本文属性约简算法进行约简，取最大进化代数为300，记忆细胞数为20，调整参数 α 设为0.9，与个体中条件属性的个数相比给与分类质量以更大的关注度。

表1　汽轮机轴系振动离散化决策表

N	$x1$	$x2$	$x3$	$x4$	$x5$	$x6$	$x7$	$x8$	D
1	1	1	1	3	1	2	2	2	1
2	2	1	2	2	2	1	2	2	1
3	1	1	1	2	1	1	1	1	1
4	1	1	1	2	1	1	1	1	1
5	1	1	1	3	3	3	2	3	1
6	1	1	1	3	3	3	2	3	1
7	1	1	1	3	3	3	2	3	1
8	1	1	1	3	1	2	1	1	1
9	1	1	1	3	2	1	1	2	1
10	1	1	1	2	2	1	1	2	1
11	1	1	2	2	1	1	1	1	1
12	1	1	1	2	3	1	1	2	2
13	1	1	1	1	2	1	1	1	2
14	1	1	1	2	3	1	1	2	2
15	1	1	1	2	2	1	1	2	2
16	1	1	1	1	2	1	1	2	2
17	1	1	1	1	2	1	1	2	2
18	1	1	1	1	2	1	1	2	2
19	1	1	1	1	2	1	1	2	2
20	1	1	1	1	2	1	1	2	2
21	1	1	1	2	3	1	1	2	2
22	1	2	1	1	1	1	1	1	3
23	1	2	1	1	1	1	1	1	3
24	1	3	1	1	1	1	1	2	3
25	1	2	1	1	1	1	1	1	3
26	1	2	2	1	1	1	1	1	3
27	1	2	1	1	1	1	1	1	3
28	1	3	2	1	1	1	1	2	3
29	1	3	1	1	1	1	1	1	3
30	2	3	2	1	1	1	2	1	3
31	2	2	2	1	1	1	1	1	3
32	2	1	3	1	1	2	3	3	4
33	3	2	3	2	3	3	2	3	4
34	2	2	3	2	2	3	3	3	4
35	3	2	3	2	3	3	3	3	4
36	2	2	3	2	3	2	2	2	4
37	1	1	1	1	1	1	1	1	5
38	1	1	1	1	1	1	1	1	5
39	1	1	1	1	1	1	1	1	5
40	1	1	1	1	1	1	1	1	5

计算结果获得了3个最小的属性约简集合。其中，除文献[7]中获得的 $\{x1,x2,x4,x5\}$ 属性约简集合外，还有2个属性约简集合 $\{x2,x3,x4,x5\}$ 和 $\{x2,x4,x5,x7\}$，这3个集合中都包含了 $\{x2,x4,x5\}$ 特征，说明它们是反映表1中5种状态的关键特征，而单一的约简集合则是无法反映的。考虑到属性 $x7$ 的表征为4倍频幅值，其影响力远不如其他倍频分量。因此，$\{x2,x4,x5,x7\}$ 属性集合可以不予考虑。而属性 $x1$ 和属性 $x3$ 分别代表了2个分倍频信息，它们对现有表1知识表达系统中的分类作用相同，所以最优的属性约简集合应为 $\{x1,x2,x4,x5\}$ 和 $\{x2,x3,x4,x5\}$。

3.2　实例2

文献[8]给出了某柴油机燃油喷射系统的柱塞在正常和磨损状态下经过多次采样并经AR时序建模而得到的压力波形特征参数。试验是在柴油机喷油泵试验台上进行的。试验时，柴油机为空载，凸轮轴转速为500 r/min，压力波形由串接在喷油泵端的压阻式压力传感器测取，经CS2092DH数据采集器进行AD转换，用Matlab语言编程计算得到AR参数。

表2所示为柴油机燃油喷射系统的柱塞在不同状况下的9个经Kohonon网络进行连续属性的离散化后的标准样本决策表，其中 $\{a1,a2,\cdots,a10,nj\}$ 为条件属性，d 为决策属性(取柱塞正常为1，磨损为2)。

表2　柱塞故障离散化决策表

N	$a1$	$a2$	$a3$	$a4$	$a5$	$a6$	$a7$	$a8$	$a9$	$a10$	nj	d
1	1	0	0	0	1	0	0	1	1	1	1	1
2	0	1	1	0	0	0	1	1	1	0	0	1
3	0	1	0	0	0	0	1	0	0	1	1	2
4	1	0	0	1	1	1	0	0	1	1	0	2
5	0	1	1	1	0	1	1	0	0	1	1	2
6	0	1	1	0	0	1	0	1	0	1	1	2
7	1	0	1	1	0	0	1	0	1	0	0	1
8	1	0	1	1	0	0	1	0	1	0	1	1
9	1	0	1	1	1	1	1	1	1	0	0	2

应用本文的属性约简方法获得的最小属性集合有 $\{a6,a9\}$，$\{a3,a5,a6\}$，$\{a3,a6,a8\}$，$\{a4,a6,a8\}$，$\{a1,a7,a8\}$，$\{a4,a5,a9\}$，$\{a2,a6,a10\}$，$\{a6,a8,a10\}$，$\{a5,a9,a10\}$ 共9种。相关文献中只计算出了 $\{a3,a4,a5,a6\}$，$\{a6,a7,a8,a9\}$，$\{a4,a5,a6,a7,a8\}$ 共3种组合，远少于本文的结果，且这3个约简都不是最小约简。

4　结束语

属性约简是粗糙集理论中的一个核心部分，目前还没有高效的最佳属性约简算法。对一个知识表达系统而言，不同的约简可以从不同的角度对数据进行浓缩和简化，从而为用户提供更多的信息。因而求解出多组不同的最小约简是有必要的。借鉴生物免疫原理，本文提出了一种新型粗糙集属性约简的新方法。并通过抗体更新和亲和力抑制以维持群体的多样性，从而求出决策表中存在的多种不同知识约简，为特征提取、决策支持和数据挖掘等提供了更多的信息，有着非常广泛的应用前景。实验表明，该算法是正确、有效的。

参考文献

1 Pawlak Z. Rough Set——Theoretical Aspects of Reasoning About Data[M]. London: Kluwer Academic Publishers, 1991: 1-110.

2 Pawlak Z, Slowinski R. Rough Set Approach to Multiattribute Decision Analysis[J]. European Journal of Operational Research, 1994, 72(33): 443-459.

3 Wong S K M, Ziarko W. On Optional Decision Rules in Decision Tables[J]. Bulletin of Polish Academy of Science, 1985, 33(11/12): 693-696.

4 苗夺谦，王　珏. 粗糙集理论中概念与运算的信息表示[J]. 软件学报, 1999, 10(2): 113-116.

5 王　珏. Rough Sets约简与数据浓缩[J]. 高技术通讯, 1997, 7(11): 40-45.

6 Lydyard P M, Whelan A, Fanger M W. Instant Notes in Immunology [M]. Beijing: Science Press, 2001: 1-40.

7 于达仁，胡清华，鲍　文. 融合粗糙集和模糊聚类的连续数据知识发现[J]. 中国电机工程学报, 2004, 24(6): 205-210.

8 曹龙汉，曹长修. 基于粗糙集理论的柴油机神经网络故障诊断研究[J]. 内燃机学报, 2002, 20(4): 357-361.