

# 分类规则挖掘的免疫算法

王自强, 冯博琴

(西安交通大学电子与信息工程学院, 710049, 西安)

**摘要:** 为了高效地从数据库中挖掘分类规则, 提出了一种基于免疫算法的分类算法. 该算法的核心思想为: 对规则的前件进行固定长度编码, 适应度函数的计算由分类规则的较小分类错误率、简洁性、一致性和训练实例的覆盖性构成, 通过把适应度最小的个体作为先验知识来修改个体的某些分量的方法进行疫苗接种, 并通过检测个体是否出现退化和模拟退火来实现免疫选择, 同时还采用了基于信息增益的规则剪枝策略. 在美国加州大学标准数据集中的 5 个数据集上将该算法与 RISE 和 OCEC 算法进行了实验比较, 结果表明该算法不仅具有更快的收敛速度, 而且获得了更高的预测准确率及更小的规则集.

**关键词:** 数据挖掘; 分类规则; 免疫算法; 信息增益

**中图分类号:** TP18 **文献标识码:** A **文章编号:** 0253-987X(2005)02-0111-04

## Mining of Classification Rule Based on Immune Algorithm

Wang Ziqiang, Feng Boqin

(School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China)

**Abstract:** To efficiently mine the classification rule from databases, a novel classification algorithm based on immune algorithm was proposed. The core of the immune classification algorithm is as follows. The rule antecedent is encoded as fixed-length chromosome; The fitness function is calculated according to minor misclassification ratio, simplicity and consistency of rules, and coverage ratio of training examples; A vaccination is accomplished by modifying genes on some bits in accordance with minimal fitness function which serves as prior knowledge; Immune selection is accomplished by testing whether a serious degeneration has happened in the evolutionary process and annealing selection. Meanwhile, a rule pruning procedure based on information gain was designed for improving the comprehensibility of classification rule mined. The algorithm has been compared with RISE and OCEC algorithms with five benchmark datasets from UCI data set repository. Experimental results show that the proposed algorithm not only has faster convergence speed, but also can achieve higher prediction accuracy with less number of rules.

**Keywords:** data mining; classification rule; immune algorithm; information gain

分类是数据挖掘中的一个重要研究课题, 其目标是根据数据已有的类别归纳出每一类的一般性描述. 最常用的分类算法是决策树方法, 该方法对于小的数据集是适合的, 当数据集非常大时, 由于计算量太大而无法应用. 相关的改进方法有统计和粗糙集方法、神经网络方法、贝叶斯方法等, 其不足之处是: 需要额外的专家知识才能有效地工作<sup>[1]</sup>, 尤其是当

分类规则中条件参数的不同组合太多时, 由于计算量太大而无法应用.

近年来, 有关学者提出了基于遗传算法(GA)<sup>[2]</sup>的分类方法<sup>[3-5]</sup>并取得了较好的分类效果, 但是在进化过程中可能产生退化现象, 将导致迭代次数过多以及预测准确率不高. 为了克服 GA 的上述不足, 本文提出了基于免疫算法(IA)<sup>[6]</sup>的分类规则挖掘算法.

# 1 基于 IA 的分类规则挖掘算法

## 1.1 个体的编码

本文采用的方法是,只对规则的前件(IF部分)进行固定长度编码,编码结构如图1所示.图中: $A_i$ 表示第*i*个属性; $V_{ij}$ 表示属性 $A_i$ 的第*j*个属性值; $O_i$ 是和属性 $A_i$ 相关的逻辑/关系运算符; $F_i$ 表示第*i*个属性条件是否出现在规则中的标识符,其取值是1或者0,当 $F_i=1$ 时表示第*i*个属性 $A_i$ 出现在规则的前件中, $F_i=0$ 表示 $A_i$ 没有出现在规则的前件中.

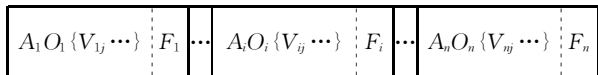


图1 个体的编码结构

## 1.2 适应度的计算

为了提高分类规则的挖掘效果,本文的适应度函数主要由算法的分类错误率、规则的简洁性、规则的一致性以及使发现的规则尽可能多地覆盖训练实例而构成.

(1)分类错误率.确保较小的分类错误率是衡量分类算法的一个重要标准,在此采用的分类错误率计算式为

$$E = \frac{N_{m-tr}}{N_{tr}} \quad (1)$$

式中: $N_{m-tr}$ 表示错误分类的实例数目; $N_{tr}$ 表示所有的训练实例数目.

(2)规则的简洁性.根据数据挖掘中的“奥卡姆剃刀”原则,为了使挖掘的分类规则尽可能地简洁,可采用规则简洁性度量函数

$$f_1 = \frac{n_c}{n} \quad (2)$$

式中: $n_c$ 表示实际出现在规则中的条件数; $n$ 表示在编码中允许的最大条件数. $f_1$ 的值越小,说明规则的简洁性越好.

(3)规则的一致性.为了确保分类规则的一致性,减少其不确定性,可采用类似于信息论中的熵的计算方法来度量规则的一致性,因为信息熵的值越小,系统的不确定性就越小.

设 $p_c$ 表示单个个体中规则的后件(结论)与训练实例的目标值相一致的概率,由于一个个体是由数个规则构成的,于是规则的一致性度量函数

$$f_2 = \frac{-p \ln p_c - (1-p_c) \ln(1-p_c)}{N_{tr}} \quad (3)$$

$f_2$ 的值越小,分类规则的一致性就越好,即不确定

性程度越小.

(4)训练实例的覆盖性.为了使算法能够应用,必须使训练实例的覆盖率达到一定的程度,于是定义了衡量训练实例的覆盖性函数

$$f_3 = 1 - \frac{|\bigcup_i P_i^+| + |\bigcup_i N_i^-|}{N_{tr}} \quad (4)$$

式中: $P_i^+$ 表示实例的目标值是正的,并且至少被个体中的一条规则分类为正; $N_i^-$ 表示实例的目标值是负的,并且至少被个体中的一条规则分类为负.

综合以上因素,本文采用的适应度计算函数为

$$f = E + f_1 + f_2 + f_3 \quad (5)$$

## 1.3 免疫分类算法中的算子设计

(1)采用轮盘式选择算子<sup>[2]</sup>.

(2)在个体与个体之间的同类规则的同基因位进行交叉,即一条规则中的某个属性或者几个属性对应的边界值与另外一条规则的相应部位进行交叉,这种算子的交叉率为75%.

(3)随机地把个体中的一个属性值变换为属于同一属性域中的另一个属性值,这种变异算子的变异率为2%.

(4)免疫算子的实现由接种疫苗和免疫选择两部分构成.接种疫苗,即给定个体 $v$ 并对它进行接种,也就是按照一定的先验知识修改 $v$ 的某些分量(规则中前件的属性条件),以便能以较大的概率获得更好(适应度值变小)的适应度.可把上次(第*k*-1次)迭代中适应度最小的个体(最好的个体)作为当前(第*k*次)迭代中的先验知识.接种疫苗需满足的条件是,若 $v$ 已是最优个体,则 $v$ 以概率1转移到 $v$ .设有种群 $V=(v_1, v_2, \dots, v_{n_p})$ ,对 $V$ 进行接种疫苗操作就是在 $V$ 中按比例 $\alpha$ 随机抽取 $N_a = \alpha n_p$ 个个体接种.免疫选择分两步完成:一是免疫检测,即对接种了疫苗的个体进行检测,若适应度大于父代(适应度定义值越小越好),说明在交叉、变异过程中出现了严重的退化现象,这时该个体被父代中对应的个体取代;二是模拟退火选择,即在当前的*k*代群体 $V_k=(v_1, v_2, \dots, v_{n_p})$ 中以概率 $p(v_i) = \frac{e^{f(v_i)/T_k}}{\sum_{i=1}^{n_p} e^{f(v_i)/T_k}}$ 选择个体 $v_i$ 进入新的父代群体,得到新一代父本,其中 $f(v_i)$ 为个体 $v_i$ 的适应度, $\{T_k\}$ 是趋于0的温度控制序列, $T_k = \ln\left(\frac{T_0}{k} + 1\right)$ .

## 1.4 规则的剪枝

为了使生成的规则更加简洁、易于用户理解,可采取基于信息增益<sup>[7]</sup>的规则剪枝策略.使用这种策

略的原因在于,它仅需搜索规则中条件属性所在的部分空间,所需的额外信息较少,所以具有很快的计算速度,而且经信息增益剪枝后的规则,具有更加简洁的表现形式及更低的错误分类率.下面给出基于信息增益的剪枝算法的描述.

**算法 1** 设  $n$  表示实际出现在规则前件(条件)中的属性数目,数组 Info\_Gain[]用来存储属性条件的信息增益值,数组 Sort\_Cond[]用来存储信息增益值较小的条件标号,则剪枝算法步骤如下.

步骤 1: 令规则中的最小条件数 Min\_Cond=1.

步骤 2: 用 for  $i=1$  to  $n$  do 来计算每个条件的信息增益值 Info\_Gain[ $i$ ].

步骤 3: 对 Info\_Gain[ $i$ ]中的值按增序排序.

步骤 4: 用 for  $i=1$  to  $n$  do 把第  $i$  个信息增益值最小的条件号 id 赋给 Sort\_Cond[ $i$ ].

步骤 5:  $t\_id=1$ ,用于记录当前的条件标号.

步骤 6: 计算规则中标识符为 1 的条件个数,并把该值赋给 Num\_Cond.

步骤 7: 运行条件 while (Num\_Cond > Min\_Cond) and ( $t\_id < n$ ), 则:

(1) 在 0 和 1 之间随机产生一个随机数 Rand\_Num;

(2) 如果 Rand\_Num < Info\_Gain [Sort\_Cond[ $t\_id$ ]], 那么条件标号为 Info\_Gain [Sort\_Cond[ $t\_id$ ]]的属性条件留在规则中,否则条件标号为 Info\_Gain [Sort\_Cond[ $t\_id$ ]]的属性条件从规则中去掉.

步骤 8: while 循环后结束.

### 1.5 分类规则挖掘的免疫算法描述

在综合以上关键技术分析的基础上,下面给出本文分类规则挖掘的免疫算法描述.

**算法 2** 分类规则挖掘的免疫算法步骤如下.

步骤 1: 按照 1.1 节的编码方案对个体进行编码,并随机产生初始化父代种群  $A_1$ .

步骤 2: 按照式(5)计算个体的适应度.

步骤 3: 判断是否满足停机条件(停机条件可设定为适应度函数所能达到的阈值或迭代次数),若条件满足,则转向步骤 8,否则继续.

步骤 4: 对当前的  $k$  代种群  $A_k$  进行交叉操作,得到种群  $B_k$ .

步骤 5: 对  $B_k$  进行变异操作,得到种群  $C_k$ .

步骤 6: 对  $C_k$  进行疫苗接种,得到种群  $D_k$ .

步骤 7: 对  $D_k$  进行免疫选择,得到新一代父本  $A_{k+1}$ ,转至步骤 2.

步骤 8: 按照算法 1 对生成的规则进行剪枝,并输出剪枝后的分类规则,结束.

## 2 实验结果

为了验证本文提出的免疫分类算法的性能,采用了美国加州大学机器学习数据集中的 5 个数据集<sup>[8]</sup>,即 Iris、Dermatology、Pima、Breast 和 Hepatitis 作为测试数据.另外,由于本文算法挖掘的是分类规则,因而首先利用 C4.5-Disc 方法<sup>[7]</sup>对连续属性离散化,预测准确率度量采用 10 次交叉验证法<sup>[9]</sup>.

将本文方法与 RISE<sup>[10]</sup>、OCEC<sup>[5]</sup> 2 种方法进行了比较,其中 RISE 是一种高效的非遗传方法,OCEC 是一种新的遗传分类方法.实验中,RISE、OCEC 2 种分类方法的参数设置可参见文献<sup>[5, 10]</sup>,本文方法的参数设置为:交叉率  $p_c=0.75$ ;变异率  $p_m=0.02$ ;接种疫苗中的比例选择  $\alpha=1.5$ ;编码长度  $n$  为数据集中的最多属性条件数;停机条件为迭代 600 次.

表 1 列出了 3 种方法预测准确率的平均值与标准方差,表 2 列出了各方法产生规则的平均个数,表 3 给出了收敛于最优解的运行时间比较.从表 1 的实验数据可以看出,本文方法在所有测试数据集上的预测准确率都大于其他 2 种方法,而标准偏差也都小于另 2 种方法,并且标准偏差值也较小.这说明,本文方法的稳定性较好,对同一数据集的不同训练数据,预测准确率不会产生较大的波动.从表 2 的实验结果可以看出,本文方法发现的规则数目在所有的测试数据集上都小于另 2 种方法,因而用我们提出的方法发现的规则列表更加简洁,从而更便于用户的理解.从表 3 的运行时间看,本文提出的免疫分类算法的运行时间不仅小于 OCEC 所用的时间,而且也小于 RISE 所用的时间,这说明本文算法能有效地克服遗传算法进化过程中的退化现象,从而使得发现最优解的时间大大减少,所以本文算法具有更快的收敛速度.

表 1 3 种分类方法的预测准确率比较

数据集	预测准确率/%		
	RISE	OCEC	本文方法
Iris	92.67±0.06	98.00±0.02	99.01±0.01
Dermatology	91.47±0.45	93.42±0.12	95.47±0.03
Pima	65.63±0.30	75.00±0.04	77.26±0.02
Breast	91.85±0.07	95.42±0.02	96.38±0.01
Hepatitis	91.24±0.73	95.53±0.02	96.86±0.02

表2 3种分类方法挖掘的规则数目比较

数据集	挖掘率/%		
	RISE	OCEC	本文方法
Iris	11.90	4.60	3.20
Dermatology	9.59	6.85	5.66
Pima	22.90	17.70	9.89
Breast	32.80	15.50	11.26
Hepatitis	5.76	4.66	3.85

表3 3种分类方法的运行时间比较

数据集	运行时间/s		
	RISE	OCEC	本文方法
Iris	6.8	5.9	5.2
Dermatology	14.9	17.4	13.5
Pima	11.7	13.8	10.3
Breast	15.2	16.8	14.6
Hepatitis	23.7	26.5	21.4

### 3 结束语

为了较好地解决遗传分类算法在进化中产生的退化现象,进而导致迭代次数过多以及预测准确率不高的问题,提出了基于免疫算法的分类规则挖掘算法,并通过实验说明了该算法的可行性和高效性。

#### 参考文献:

[1] Han J W, Kamber M. Data mining: concepts and techniques [M]. San Mateo, USA: Morgan Kaufmann Publishers, 2000. 185-211.

- [2] 陈国良,王煦法,庄镇泉,等. 遗传算法及其应用 [M]. 北京:人民邮电出版社, 2001. 164-192.
- [3] Yang L Y, Widyantoro D H, Ioerger T, et al. An entropy-based adaptive genetic algorithm for learning classification rules [A]. The 2001 Congress on Evolutionary Computation, Seoul, South Korea, 2001.
- [4] Carvalho D R, Freitas A A. A hybrid decision tree/genetic algorithm for coping with the problem of small disjuncts in data mining [A]. The Genetic and Evolutionary Computation Conference, Las Vegas, USA, 2000.
- [5] 刘静,钟伟才,刘芳,等. 组织协同进化分类方法 [J]. 计算机学报, 2003, 26(4): 446-453.
- [6] Jiao L C, Wang L. A novel genetic algorithm based on immunity [J]. IEEE Transactions on Systems, Man & Cybernetics, 2000, 30(5): 552-561.
- [7] Kohavi R, Sahami M. Error-based and entropy-based discretization of continuous features [A]. Second International Conference on Knowledge Discovery and Data Mining, Menlo Park, USA, 1996.
- [8] Hettich S, Bay S D. The UCI KDD archive [EB/OL]. <http://kdd.ics.uci.edu>, 2000-04-26.
- [9] Weiss S M, Kulkowski C A. Computer systems that learn [M]. San Mateo, USA: Morgan Kaufmann Publishers, 1991. 75-93.
- [10] Domingos P. Unifying instance-based and rule-based induction [J]. Machine Learning, 1996, 24(2): 141-168.

(编辑 苗凌)

## 本刊再次荣获“百种中国杰出学术期刊”称号

2004年12月7日,科技部中国科技信息研究所在北京国际会议中心召开新闻发布会,公布了“中国科技论文统计结果”和“中国科技期刊引证报告”。据悉,《西安交通大学学报》(自然科学版)再次荣获“百种中国杰出学术期刊”称号。

“百种中国杰出学术期刊”的评定主要是以评价期刊学术影响力的指标——总引文频次、影响因子、他引总引比等为依据,结合专家意见评选出来的。这次《西安交通大学学报》(自然科学版)的学术影响指标又有了较大提高,总被引频次为638次(2003年公布的数字为521次),影响因子为0.335(2003年公布的数字为0.287),他引总引比为0.92(2003年公布的数字为0.91)。全校发表国内论文3446篇,《西安交通大学学报》(自然科学版)占9.5%;全校国内论文总被引频次4101次,《西安交通大学学报》(自然科学版)占15.57%。

在新的成绩面前,《西安交通大学学报》(自然科学版)的编辑们并没有停步,他们在坚持以往“出精品、创名牌,办一流学报;争时效、促交流,举科技人才”办刊目标的同时,今年年末在总结办刊经验的基础上,努力将学报工作融入学校的科研工作之中,为学校创办世界知名高水平大学作出更大的贡献。

(期刊中心供稿)