基于免疫原理和Boosting机制的模糊分类规则挖掘算法

张雷[1,2]，李人厚[1]
（1.西安交通大学系统工程研究所，710049，西安；2.河南科技大学电子与信息工程学院，471003，洛阳）

摘要：基于免疫原理和Boosting机制，提出了一种模糊分类规则挖掘算法.该算法主要借鉴于自然免疫系统中的克隆选择原理，通过抗体种群的演化来优化模糊规则.模糊规则库通过增量的方式产生，算法每次运行得到一个规则.Boosting机制用于调整训练数据的权值，使得新生成规则集中于当前未被覆盖或误分类的数据实例.仿真实验表明，所提算法可根据规则的分类精度来调整训练数据的权值，促进了模糊规则之间的协作关系，避免了规则之间相互冲突，提高了系统的分类精度.

关键词：免疫原理；模糊规则；分类规则；挖掘算法
中图分类号：TP18 **文献标识码**：A **文章编号**：0253-987X(2007)08-0927-04

Algorithm for Mining Fuzzy Classification Rules Based on Immune Principles and Boosting Mechanism

Zhang Lei[1,2], Li Renhou[1]
(1. Institute of System Engineering, Xi′an Jiaotong University, Xi′an 710049, China; 2. School of Electronics and Information Engineering, Henan University of Science & Technology, Luoyang 471003, China)

Abstract: An algorithm is proposed for mining fuzzy classification rules by using natural immune principles and boosting mechanism. The proposed algorithm is mainly inspired by the clonal selection principle of biological immune systems, and the population of antibodies is evolved to optimize fuzzy classification rules. The fuzzy classification rule base is generated in an incremental way, in which one rule is obtained in each run of the proposed algorithm. When one rule is generated, boosting mechanism is used to change the weights of the training instances so as to mine new rules that are focused on currently uncovered or misclassified instances. The proposed algorithm can promote the cooperation and avoid the conflict among fuzzy rules by using boosting mechanism. Compared to other relevant algorithms the proposed algorithm has better predictive accuracy.

Keywords: immune principle; fuzzy rule; classification rule; mining algorithm

模式分类是数据挖掘中最为常用的一个应用领域，分类模型可作为一种解释性工具，用于区分不同类别的数据，并能确定未知的数据类别.基于模糊规则的分类系统，其规则可以语言变量的形式来表示，这种系统具有较好的透明性和可理解性.

近来，许多研究逐渐关注从训练数据集中自动获取模糊规则[1-2]，如基于理论免疫学和免疫系统的人工免疫算法[3]，资源有限的人工免疫识别系统[4]，挖掘模糊分类规则的人工免疫算法[5]等.文献[5]实质上是采用了“分而治之”的策略，其规则之间会发生冲突，影响系统的分类精度.针对这一问题，本文提出了一种人工免疫算法，主要是借鉴了免疫系统中的克隆选择原理[6]，并利用Boosting机制[7]，使得新规则集中于当前未被覆盖或误分类的数据，以

收稿日期：2006-11-17. 作者简介：张雷(1974～)，男，博士生；李人厚(联系人)，男，教授，博士生导师.

促进模糊规则之间的协作关系，提高分类性能.

1 模糊分类规则和 Boosting 机制

本文所采用的模糊规则的类型为

$$R_i:\text{if } x_1 \text{ is } A_{i1} \text{ and } \cdots x_n \text{ is } A_{in} \text{ then } y=C_i \quad (1)$$

式中：A_{ij} 为规则 R_i 中对应第 j 个属性的模糊集合；$C_i\in\{C_1,\cdots,C_m\}$ 为 R_i 的类别. 式(1)的前件部分表示由数据属性构成的空间中的一个特定区域，而后件部分表示规则的类别.

假定训练数据集为 $\{(x^1,C_1),\cdots,(x^N,C_N)\}$，它包含了 N 个数据. 其中，$x^k=\{x_1^k,\cdots,x_n^k\}$ 表示第 k 个数据，包含了 n 个属性，而 $C_k\in\{C_1,\cdots,C_m\}$ 表示该数据所对应的类别. 数据 x^k 对规则 R_i 的激励强度可定义为

$$\mu_{R_i}(x^k)=\min_{j=1,\cdots,n}\mu_{A_{ij}}(x_j^k) \quad (2)$$

用它可描述数据与规则之间的匹配程度. 模糊推理机制根据模糊规则库可以确定未知数据的类别，例如单一优胜规则或投票方案. 对于投票方案，数据 x^k 可由

$$C_{\max}=\arg\max_{j=1,\cdots,m}\Big(\sum_{R_i|C_i=C_j}\mu_{R_i}(x^k)\Big) \quad (3)$$

来确定类别，即对所有类别为 $C_i=C_j(j=1,\cdots,m)$ 的规则激励强度进行求和，而数据 x^k 则被分类到最大值对应的类别 $C_{\max}$.

Boosting 机制将多个不同的、分类精度不高的分类器集成为一个总的分类器，再由投票方案结合各分类器的结果进行分类，总的分类器可以得到比各子分类器更优的分类精度. 本文算法中的每个模糊规则可视为一个子分类器，而模糊规则库就相当于总的分类器.

2 基于免疫原理和 Boosting 机制的模糊分类规则挖掘算法

本文提出了一种人工免疫算法，用于挖掘模糊分类规则，它主要借鉴了自然免疫系统中的克隆选择原理. 克隆选择属于自然选择的一种类型，即抗原具有选择抗体的功能，只有那些能够有效识别抗原的抗体可进行克隆增殖和超变异.

2.1 适应度函数

算法中每个抗体表示一个模糊规则，而每个抗原表示一个训练实例. 抗原 g^A 与抗体 b_i^A 之间的亲和度定义为

$$A_f(b_i^A,g^A)=\mu_{R_i}(x^k)=\min_{j=1,\cdots,n}\mu_{A_{ij}}(x_j^k) \quad (4)$$

式中：R_i 为 b_i^A 对应的模糊规则. 式(4)表示训练实例与模糊规则之间的匹配程度.

抗体的适应度函数可反映模糊规则分类性能的优劣，定义为

$$\text{fit}(b_i^A)=\prod_{i=1}^{3}f_i \quad (5)$$

它包含了 3 个部分. 第 1 部分用于评价规则 R_i 对训练数据集中同类数据的覆盖能力，定义为

$$f_1=\frac{\sum_{k|c^k=c_i}W^k\mu_{R_i}(x^k)}{\sum_{k|c^k=c_i}W^k} \quad (6)$$

式中：W^k 表示数据 x^k 的权值. 第 2 部分用于度量模糊规则的一致性，定义为

$$f_2=\begin{cases}0, & \sum_{k|c_k=c_i}\mu_{R_i}(x^k)<\sum_{k|c_k\neq c_i}\mu_{R_i}(x^k)\\ \dfrac{\sum_{k|c_k=c_i}\mu_{R_i}(x^k)-\sum_{k|c_k\neq c_i}\mu_{R_i}(x^k)}{\sum_{k|c_k=c_i}\mu_{R_i}(x^k)}, & \text{其他}\end{cases} \quad (7)$$

它要求一个规则尽可能多地覆盖本类数据，同时尽可能少地覆盖其他类数据，且与数据的权值无关. 第 3 部分用于评价规则的分类精度，定义为

$$f_3=\frac{T_P+T_N}{T_P+T_N+F_P+F_N} \quad (8)$$

式中：T_P 表示被规则覆盖并且与其类别相同的实例数目；F_P 为被规则覆盖但与其类别不同的实例数目；T_N 为未被规则覆盖也与其类别不同的实例数目；F_N 为未被规则覆盖但与其类别相同的实例数目. 它们分别定义为

$$T_P=\sum_{k|c^k=c_i}^{N}W^k\mu_{R_i}(x^k) \quad (9)$$

$$F_P=\sum_{k|c^k\neq c_i}^{N}W^k\mu_{R_i}(x^k) \quad (10)$$

$$T_N=\sum_{k|c^k\neq c_i,\mu_{R_i}(x^k)=0}^{N}W^k \quad (11)$$

$$F_N=\sum_{k|c^k=c_i,\mu_{R_i}(x^k)=0}^{N}W^k \quad (12)$$

式(9)～式(12)中：N 为训练数据的数目.

2.2 基于训练数据集的模糊规则挖掘

基于免疫原理和 Boosting 机制的模糊分类规则挖掘算法是针对每个类别分别进行的，每次运行生成一个规则，然后调整训练数据的权值，并再次进行新规则挖掘，直到未被覆盖的数据数目小于设定

的阈值. 该算法的步骤如图 1 所示.

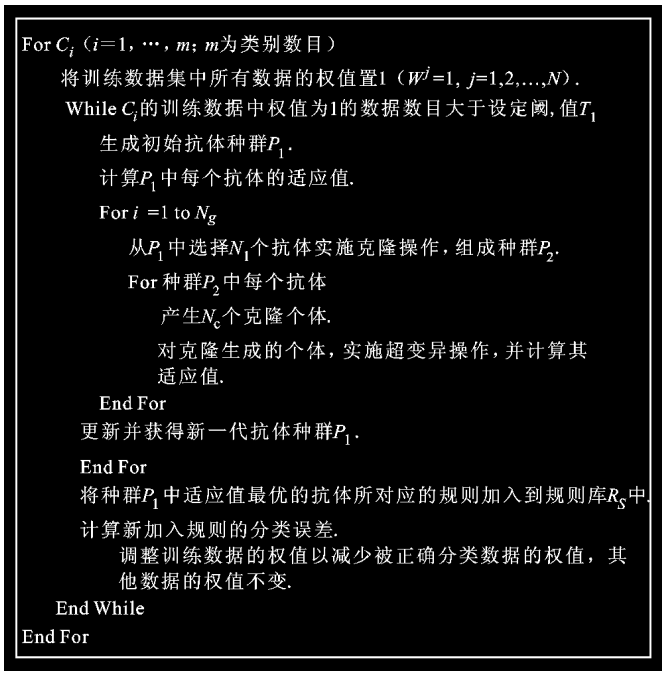


图 1　基于免疫原理和 Boosting 机制的模糊分类规则挖掘算法

模糊规则集合 R_S 初始时为空集，算法每次迭代结束后，将其中适应值最优的规则加入到 R_S 中，同时得到该规则的分类误差. 算法分别针对不同类别的规则进行优化搜索操作，每次对一个新类别规则进行挖掘时，都需将所有训练数据的权值置 1.

每个抗体对应于式(1)的模糊规则，其编码串包含 $n+1$ 个基因，即包含规则的 n 个前件条件和规则的 1 个类别 C. 每个前件条件 i 的基因包括 2 个部分，第 1 部分是其对应隶属度函数的参数，第 2 部分为一个布尔标志位 B_i，表示规则中是否包含前件条件 i，$B_i=1$ 则表示包含了该条件. 所有隶属度函数均为三角形函数，它们可用 3 个参数表示.

为了提高搜索效率，本文利用训练数据作为原型来获得初始种群中抗体的隶属度函数参数. 对于每个抗体，首先从指定类别的训练数据中随机选择 2 个数据 x^1 和 x^2，它们被选择的概率与它们的权值成正比，然后按照

$$a_j=\min\{x_j^1,x_j^2\};\quad c_j=\max\{x_j^1,x_j^2\};\quad b_j=0.5(x_j^1+x_j^2) \tag{13}$$

确定抗体第 j 个前件的隶属度函数 A_{ij} 的参数.

克隆操作从当前种群中选择 T 个适应值最高的抗体，同时要求它们互不相同，以此作为克隆的个体. 这种策略能够保证抗体之间的多样性，避免陷入局部最优解. 对每个克隆的抗体，其产生克隆的数目 N_c 与该抗体的适应值成正比，即

$$N_c=\max\{N_{\min},\mathrm{round}(N_{\max}\times \mathrm{fit}(i)/f_{\max})\} \tag{14}$$

式中：$N_{\min}$ 和 $N_{\max}$ 分别表示抗体克隆的最小和最大数目；$\mathrm{fit}(i)$ 表示抗体 i 的适应度函数值；$f_{\max}$ 表示当前种群中抗体的最大适应度函数值.

对每个新生成的克隆个体实施超变异操作，抗体的适应值越高，其变异率越低，计算式为

$$r(i)=(1-\mathrm{fit}(i)/f_{\max})(r_{\max}-r_{\min})+r_{\min} \tag{15}$$

式中：$r_{\max}$ 和 $r_{\min}$ 是分别设定的最大和最小变异率.

抗体种群的更新操作是，从新生成的抗体中选择 (N_p-N_1) 个适应值最高的抗体，并与种群 P_2 中的抗体共同组成下一代抗体种群 P_1，而种群 P_2 中的抗体相当于精英个体.

当抗体种群的演化完成以后，从中选择适应度函数值最优的抗体，并将其对应的规则 R_i 加入到模糊规则库中，同时得到该规则的分类误差

$$E(R_i)=\frac{\sum\limits_{k|c_k\neq c_i}W^k\mu_{R_i}(x^k)}{\sum\limits_{k}W^k\mu_{R_i}(x^k)} \tag{16}$$

由于模糊规则 R_i 只能对那些被覆盖的实例进行分类，因而定义中考虑了实例与规则之间的匹配程度 $\mu_{R_i}(x^k)$ 以及实例的权值 W^k.

调整训练数据的权值，即对每个训练数据的权值基于规则的分类误差 $E(R_i)$ 重新进行调整，调整式为[7]

$$W^k(t+1)=\begin{cases}W^k(t) & C_i\neq C_k\\ W^k(t)\beta^k & C_i=C_k\end{cases} \tag{17}$$

式中：β^k 因子定义为

$$\beta^k=\left(\frac{E(R_i)}{1-E(R_i)}\right)^{\mu_{R_i}(x^k)} \tag{18}$$

由于减少了被规则 R_i 正确分类数据的权值，而那些未能正确分类或者未被规则覆盖的数据的权值仍保持不变，因而 Boosting 机制相当于增加了这些数据的相对权值，使得新规则的挖掘更为有效.

2.3　测试数据集的分类

针对训练数据集，用本文算法所获得的模糊规则集合对测试数据集进行分类，可评价算法的分类性能. 对于每个测试数据 x^k，由下式确定其类别，即

$$C_{\max}=\arg\max_{j=1,\cdots,m}\Big(\sum_{R_i|C_i=C_j}(\ln(\beta_i))\mu_{R_i}(x^k)\Big) \tag{19}$$

式中：β_i 定义为

$$\beta_i=\frac{E(R_i)}{1-E(R_i)} \tag{20}$$

3 仿真实验

采用了UCI机器学习数据库[8]中的4个数据集进行实验，它们分别是Wisconsin Breast Cancer、Glass、Iris和Wine，其中：Wisconsin Breast Cancer包含9个属性，699个实例，共分为2类；Glass包含9个属性，214个实例，分为6类；Iris包含4个属性，150个实例，分为3类；Wine包含13个属性，178个实例，分为3类.

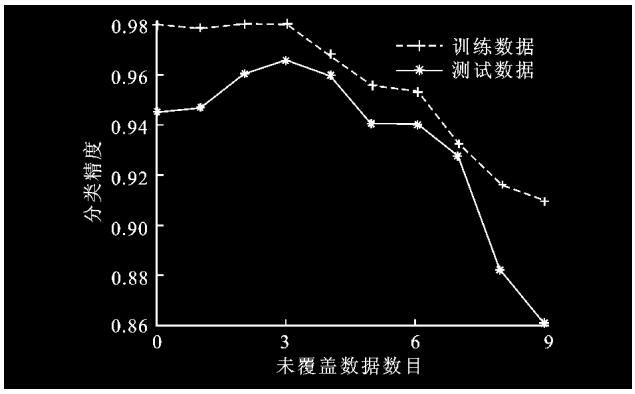
本文算法采用10阶交叉验证实验来评价算法的分类精度，即将数据集分成10个部分，每次取出其中的一个部分作为测试数据集，其余的作为训练数据集来优化模糊规则. 上述过程运行10次，取它们的平均值来评价算法的分类精度，同时与文献[5]中的算法进行分类性能比较.

表1列出了2种算法的平均分类精度以及标准差. 可以看出，对于4个数据集，本文算法在分类精度上均优于文献[5]算法，主要原因是：Boosting机制能根据规则的分类精度来调整训练数据的权值，而无需删除正确分类的数据，这样能够促进模糊规则之间的协作关系，避免文献[5]中的规则之间相互冲突的问题，从而提高了系统的分类精度.

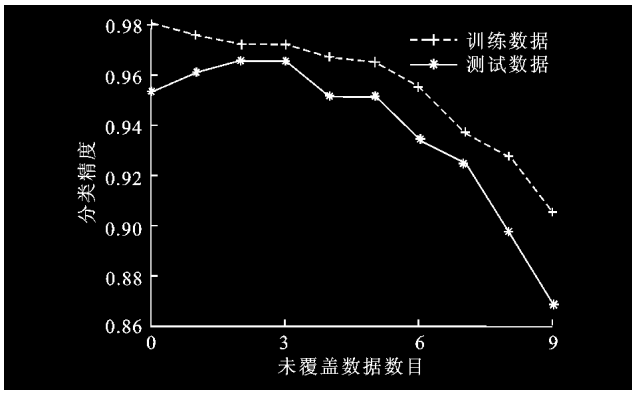
表1 算法分类精度的比较

数据集	分类精度/%	
	本文算法	文献[5]算法
Wisconsin Breast Cancer	95.6±0.8	92.8±1.5
Glass	72.2±1.9	45.7±4.3
Iris	96.1±1.1	95.1±1.7
Wine	95.8±1.6	86.2±2.2

在算法中，每类未被覆盖的数据数目阈值 T_1 是一个关键参数，针对该参数对算法性能的影响进行了实验. 对训练和测试数据的分类精度的影响如图2所示，其中图2a和图2b是分别针对Wine和Wisconsin Breast Cancer的仿真结果. 从中可以看出：对于训练数据，减少 T_1 能够提高分类精度；对于测试数据，在一定范围内减少 T_1 也能够提高分类精度，但当该参数小于一定数值时，分类精度反而下降. 这是由于 T_1 过小，对训练数据的学习出现了过拟合现象，这会降低所得模糊规则集合的泛化能力. 所以，应用时该参数不应取得太小，否则既增加了计算量，也会降低分类性能.



(a)Wine数据集



(b)Wisconsin Breast Cancer数据集

图2 分类精度与每类未被覆盖数据数目的关系曲线

4 结束语

本文提出了一种基于免疫原理和Boosting机制的模糊分类规则挖掘算法. 该算法利用克隆选择原理从训练数据集中确定模糊规则库，包括每个规则所对应的隶属度函数参数，同时算法使用了Boosting机制，通过减少那些被正确分类的数据的权值，使得新生成的规则集中于当前未被覆盖或误分类的数据，可以更好地实现模糊规则之间的协作，提高分类精度. 针对典型数据集与相关算法进行了性能比较，实验结果表明本文算法是有效的.

参考文献：

[1] Chang Xiaoguang, Lilly J H. Evolutionary design of a fuzzy classifier from data[J]. IEEE Transactions on Systems, Man, and Cybernetics, 2004, 34(4): 1894-1906.

[2] Roubos H, Setnes M. Compact and transparent fuzzy models and classifiers through iterative complexity reduction[J]. IEEE Transaction on Fuzzy Systems, 2001, 9(4): 516-524.

[3] Castro L N, Timmis J. Artificial immune systems: a

new computation intelligence approach [M]. Berlin: Springer-Verlag, 2002.

[4] Watkins A B, Boggess L C. A resource limited artificial immune classifier [C]// Proceedings of Congress on Evolutionary Computation. Piscataway, USA: IEEE, 926-931.

[5] Alves R T, Degado M R, Lopes H S, et al. An artificial immune system for fuzzy-rule induction in data mining[M]//Lecture Notes in Computer Science: Parallel Problem Solving from Nature. Berlin: Springer-Verlag, 2004: 1011-1020.

[6] Castro L N, von Zuben F J. Learning and optimization using the clonal selection principle [J]. IEEE Transaction on Evolutionary Computation, 2002, 6 (3): 239-251.

[7] Freund Y, Schapire R E. Experiments with a new boosting algorithm [C]// Proceedings of the 13th International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers, 1996: 148-156.

[8] Blake C, Merz C. UCI repository of machine learning databases (1998) [DB/OL]. [2006-01-12]. http://www.ics.uci.edu/mlearn/MLRepository.html.

（编辑　苗凌）