

## 基于免疫记忆克隆的特征选择

朱虎明, 焦李成

(西安电子科技大学智能信息处理研究所, 710049, 西安)

**摘要:** 针对数据挖掘和模式识别等领域的高维数据降维问题, 提出了一种基于抗体克隆选择学说和免疫记忆理论的特征选择算法. 该算法利用抗体种群进行全局搜索, 通过设立记忆单元来保留历史最好个体, 并对其嵌入可控制搜索深度的局部搜索算子, 用以加快抗体亲和力成熟速度, 同时对抗体种群和记忆单元采用不同的亲和度函数以获得更好的搜索能力. 将该算法用于几个高维数据集进行特征子集选择, 然后进行最近邻分类并采用留一法验证, 结果表明, 与标准遗传算法相比, 新算法具有更低的复杂度和更好的搜索能力, 其鲁棒性也优于经典的串行浮点前向搜索算法.

**关键词:** 数据挖掘; 分类; 特征选择; 遗传算法

**中图分类号:** TP18 **文献标志码:** A **文章编号:** 0253-987X(2008)06-0679-04

## Feature Selection via Immune Memory Clone

ZHU Huming, JIAO Licheng

(Institute of Intelligent Information Processing, Xidian University, Xi'an 710071, China)

**Abstract:** Focusing on the problem of dimension reduction in data mining and pattern recognition, a novel algorithm for feature selection was proposed based on antibody clonal selection and immune memory principle (ICMFS). The antibody population is used for global search, and the memory unit, which only reserves the best individuals, with embedded local search operations is designed for fine-tune search. Different fitness functions for the antibody population and the memory unit are used to improve search performance. The fitness of an individual is determined by evaluating the nearest neighbor classifier with leave-one-out cross-validation. Experiment results on several standard high-dimension datasets show that the proposed algorithm outperforms a conventional genetic algorithm and the classical sequential floating forward search algorithm in terms of classification accuracy and robustness.

**Keywords:** data mining; classification; feature selection; genetic algorithm

高维数据在数据挖掘时容易引起维数灾难问题<sup>[1]</sup>, 因此, 如何对高维数据进行降维就成为一个关键的预处理步骤. 特征选择是常用的一种降维方法, 虽然很多和特征选择相关的问题被发现是 NP-hard 问题, 但通过采用启发式搜索算法, 还是可以在运算效率和特征子集质量间寻找到一个好的平衡点. 由于标准遗传算法在应用中容易过早收敛, 因此出现了新的智能计算方法如免疫克隆算法(ICMFS)、自适应遗传算法、量子遗传算法、协同进化等<sup>[2-3]</sup>, 但这

些算法或复杂度高或局部搜索能力有限且不容易控制. 近年来, 借鉴免疫学中抗体克隆选择学说的主要机理和相关免疫记忆理论的免疫记忆克隆算法取得了较好的结果<sup>[4]</sup>. 本文提出了一种基于免疫记忆克隆的封装式特征选择算法, 利用抗体种群进行全局搜索, 通过设立记忆单元来保留历史最好个体, 并对其嵌入可控制深度的局部搜索算子, 用以加快抗体亲和力成熟速度. 这种算法既可以降低复杂度, 又可以取得较好的特征子集.

# 1 基于免疫记忆克隆的特征选择算法设计

将免疫记忆克隆算法用于特征选择,关键是编码、亲和度函数和局部搜索算子的设计。

## 1.1 编码设计

对于特征选择问题使用二进制编码。1个个体表示1个特征子集,1个长度为 $n$ 的个体对应于1个 $n$ 维的二进制特征矢量。个体中的1表示对应特征项包含于所选特征子集中,0表示不包含。例如:1个抗体编码为 $A_k=\{1001100110\}$ ,就表示所选特征子集为 $X=\{1,4,5,8,9\}$ ,由 $n$ 个抗体组成抗体群 $A=\{A_1,A_2,\dots,A_n\}$ ,正整数 $n$ 称为抗体种群规模。在特征选择问题中,抗原就是所需要的特征子集,亲和度函数的值就是这个特征子集对应的分类精确率,显然其值越大,就表示所选择的特征子集越好。

## 1.2 亲和度函数设计

免疫记忆克隆算法中的抗体种群和记忆单元的目标不一样,因此它们的亲和度函数有一些差异,用来进行全局搜索的抗体种群亲和度定义为

$$p_i = A_{acc}(i) + \left( \frac{1}{n_c} \sum_{j=1}^{n_c} w_{ij} \right) \quad (1)$$

式中: $p_i$ 是第 $i$ 个个体的亲和度; $A_{acc}(i)$ 是用种群中的个体进行特征选择后按最近邻分类并采用留一法(LOO)验证的分类正确率; $n_c$ 是克隆规模; $w$ 是归一化的海明距离,即

$$w_{ij} = \frac{d_{ij} - d_{\min}}{d_{\max} - d_{\min}} \quad (2)$$

式中: $d_{ij}$ 表示个体 $i$ 和个体 $j$ 之间的海明距离; $d_{\max}$ 和 $d_{\min}$ 表示每一代中海明距离的最大值和最小值。

## 1.3 局部搜索函数设计

如果某个个体为 $p=1001100110$ ,它对应的特征子集为 $A=\{1,4,5,8,9\}$ ,又设 $B$ 为剩余的特征集合,则 $B=\{2,3,6,7,10\}$ 。特征子集 $A$ 上的评价函数 $J(A)$ 的值为LOO评价的最近邻分类器的分类精确率,那么可以对个体 $p$ 完成如下的局部搜索。在 $B$ 中寻找元素 $b=\arg \max_{x \in B} J(A \cup \{x\})$ ,将 $b$ 从集合 $B$ 迁移到集合 $A$ ,称这个过程为个体 $p$ 的1次前向搜索,在 $A$ 中寻找1个元素 $a=\arg \max_{x \in A} J(A - \{x\})$ ,将 $a$ 从集合 $A$ 迁移到集合 $B$ ,称这个过程为个体 $p$ 的1次后向搜索。算子 $S_{add}(i,p)$ 表示对个体 $p$ 做 $i$ 次前向搜索,算子 $S_{del}(i,p)$ 表示个体 $p$ 做 $i$ 次后向搜索。假设 $b=6$ ,那么执行

$S_{add}(1,p)$ 后,元素6从集合 $B$ 迁移到集合 $A$ ,集合 $A$ 变为 $A_1=\{1,4,5,6,8,9\}$ ,集合 $B$ 变为 $B_1=\{2,3,7,10\}$ 。假设 $a=4$ ,那么执行 $S_{del}(1,p)$ 后,元素4从集合 $A$ 迁移到集合 $B$ ,集合 $A$ 变为 $A_2=\{1,5,8,9\}$ ,集合 $B$ 变为 $B_2=\{2,3,4,6,7,10\}$ 。用不同的 $i$ 值来调节局部搜索的深度, $i$ 越大,局部搜索的深度就越大,消耗的时间也就越多。本文使用 $local(i,p)$ 表示局部搜索函数,其功能是对个体 $p$ 对应的特征子集进行深度为 $i$ 的前向和后向搜索,返回搜索到的最大值。所以有

$$local(i,p) = \{S_{add}(i,p); S_{del}(i,p)\}$$

# 2 基于免疫记忆克隆算法的特征选择框架

设计好编码、亲和度函数和局部搜索算子后,就可以利用下面的算法框架进行特征选择。

(1)初始化。设定算法终止条件,给定变异概率 $p_m$ 、克隆规模 $n_c$ 、抗体种群规模 $n$ 、记忆单元规模 $s$ 、抗体种群更新率 $T$ 、抗体编码长度 $l$ ,随机产生初始抗体群 $A(0)$ ,从 $A(0)$ 中选取 $s$ 个抗体构成记忆单元 $M$ ,进化代数 $g=0$ 。

(2)计算亲和度。

(3)算法终止条件判断。如果满足就结束算法,否则继续。

(4)进化代数 $g$ 加1,抗体种群和记忆单元进行下一代操作。

(5)抗体种群更新。将抗体种群中亲和度弱的 $T$ 个抗体用 $T$ 个亲和度强的记忆单元和抗体种群中的个体代替,产生新的抗体种群 $A'(g)$ 。

(6)克隆。对当前第 $g$ 代父本种群 $A'(g)$ 进行克隆操作,得到种群 $Y(g)$ 。

(7)变异。依据概率 $p_m$ 对克隆后的种群 $Y(g)$ 进行变异得到种群 $Z(g)$ 。

(8)克隆选择操作。从种群 $Z(g)$ 中以一定比例选择亲和度较大的个体得到下一代抗体种群 $A(g+1)$ 。

(9)对记忆单元 $M(g)$ 进行局部搜索。从克隆选择完成后的抗体种群 $A(g+1)$ 中选择一个亲和度最大的个体进行局部搜索。如果返回搜索到的最大值大于记忆种群中的最小亲和度,则替换最小亲和度的个体,对 $M(g)$ 局部搜索操作完成后得到下一代记忆单元 $M(g+1)$ 。

(10)返回步骤(3)。

### 3 仿真实验及结果分析

#### 3.1 数据集和实验参数

实验所用数据集来自机器学习数据集(UCI)<sup>[5]</sup>,特征值均为数值,特征维数为13~60,详细信息见表1.

表1 UCI数据集

| 数据集           | 样本数量 | 维数 | 类别数 |
|---------------|------|----|-----|
| Ionosphere    | 351  | 34 | 2   |
| Lungcancer    | 32   | 56 | 3   |
| Sonar         | 208  | 60 | 2   |
| Soybean-large | 683  | 35 | 19  |
| Wine          | 178  | 13 | 3   |
| Wpbc          | 194  | 33 | 3   |

对于有数据丢失的Lungcancer数据集,由于样本比较少,所以删掉了2个丢失数据值的特征,Wpbc数据集删掉了丢失数据的4个样本.Soybean-large训练集为307个样本,测试集为376个样本,除了Soybean-large数据集,其余数据集均采用LOO验证,LOO是 $k$ 倍交叉验证的一个特例,即LOO中的 $k$ 等于样本个数.

进化算法的参数选取有很大的技巧和经验,本文选用的参数为:抗体种群规模 $n=20$ ,克隆规模 $n_c=40$ ,记忆单元规模 $s=10$ ,变异概率 $p_m=0.05$ ,抗

体编码长度 $l$ 等于数据集维数,抗体种群更新率 $T=10$ .进化算法的结果取15次计算的平均值.算法仿真软件为Matlab 7.0.1.

#### 3.2 时间复杂度和分类正确率

用 $T(K)$ 表示评价特征维数为 $K$ 时所花费的时间代价,又把评价一个特征花费的时间称为原子时间,记为 $t$ ,那么有 $T(K) \approx Kt$ .如一个个体为 $p=1001100110$ ,那么数据集维数 $D_p$ 等于抗体编码长度 $l$ ,即 $D_p=10$ ,特征子集的维数 $k_p$ 等于个体 $p$ 中1的数目,即 $K_p=5$ ,评价这个特征子集所花费的时间就为 $K_p t$ .对于局部搜索算子 $S_{add}(1, p)$ , $T=(K_p+1)(D_p-K_p)$ ;对于局部搜索算子 $S_{del}(1, A)$ , $T=(K_p-1)K_p$ .最后将每代中的时间求和就是总的时间复杂度,实际得到的结果见表2.

由于顺序浮动前向搜索算法(SFFS)<sup>[6]</sup>是经典的搜索能力很强的特征选择算法,所以在比较性能时也加入了这个算法.表3总结了5种算法在6个UCI数据集上的分类精确率.

图1给出了ICMFS1算法在Sonar数据集上的某一次运行结果,表明算法运行代数与LOO评价的特征子集的关系.记忆单元中的新最大值表示经过局部搜索后新发现的最好特征子集.

#### 3.3 实验结果分析和讨论

从表2和表3可以看出,本文提出的免疫记忆

表2 进化算法在UCI数据集上的时间复杂度及运行代数

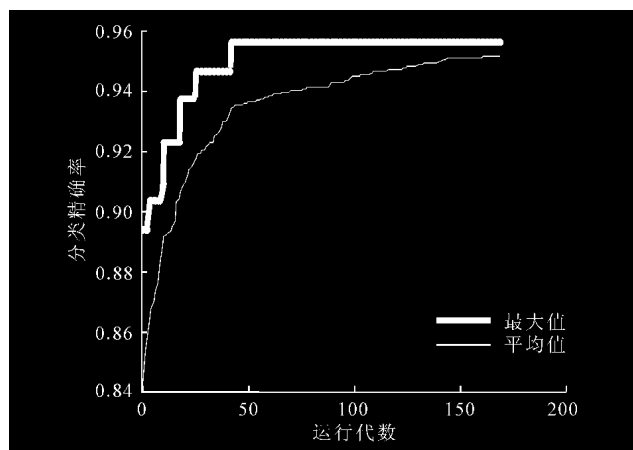
| 数据集           | 算法复杂度   |         |         | 算法运行代数 |        |        |
|---------------|---------|---------|---------|--------|--------|--------|
|               | SGA     | ICMFS1  | ICMFS2  | SGA    | ICMFS1 | ICMFS2 |
| Ionosphere    | 163 600 | 124 402 | 132 950 | 680    | 130    | 90     |
| Lungcancer    | 651 121 | 603 096 | 632 151 | 1 350  | 250    | 170    |
| Sonar         | 699 540 | 519 453 | 468 957 | 1 200  | 170    | 100    |
| Soybean-large | 332 650 | 195 648 | 211 734 | 700    | 110    | 80     |
| Wine          | 12 379  | 7 360   | 4 711   | 100    | 20     | 10     |
| Wpbc          | 209 948 | 122 751 | 122 860 | 600    | 130    | 90     |

注:SGA表示标准遗传算法<sup>[6]</sup>;ICMFS1、ICMFS2表示本文ICMFS算法,其局部搜索算子深度分别是1和2.

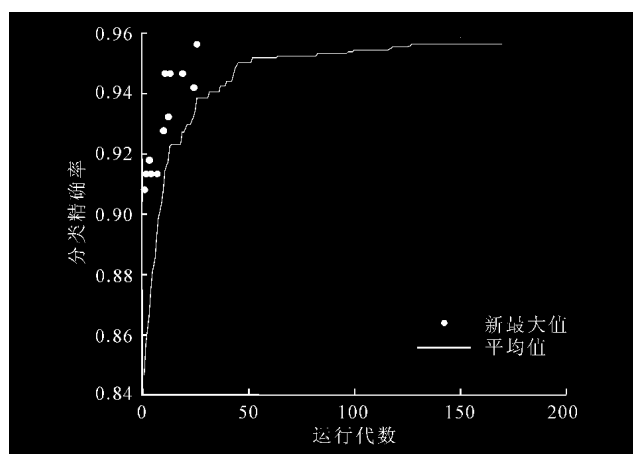
表3 不同算法在UCI数据集上的分类精确率

| 数据集           | 分类精确率/% |       |              |              |        |
|---------------|---------|-------|--------------|--------------|--------|
|               | LOO     | SFFS  | SGA          | ICMFS1       | ICMFS2 |
| Ionosphere    | 86.61   | 94.87 | 94.81        | 94.97        | 95.05  |
| Lungcancer    | 43.75   | 93.75 | <b>95.00</b> | <b>95.00</b> | 94.12  |
| Sonar         | 82.69   | 93.75 | 95.05        | <b>95.60</b> | 95.22  |
| Soybean-large | 90.69   | 94.95 | 95.14        | <b>95.48</b> | 95.35  |
| Wine          | 76.97   | 95.51 | 95.51        | 95.51        | 95.51  |
| Wpbc          | 69.07   | 75.77 | 75.97        | <b>76.53</b> | 76.49  |

注:LOO所在列的数据结果是没有特征选择的分类精确率;黑体数据表示最好的结果.



(a) 抗体种群运行结果



(b) 记忆单元运行结果

图1 ICMFS1算法在Sonar数据集上的某一次运行结果

克隆算法由于嵌入了局部搜索算子,提高了搜索效率,所以用比遗传算法更少的评价次数,取得了更好的特征子集,采用特征选择算法后,分类的精确率都有了很大的提高,基于进化的算法都取得了与SFFS相当或者更好的结果.还可以看出:在数据集维数最小的Wine数据集上,几个特征选择算法都找到了同样分类精确率的特征子集;在大一些的Ionosphere数据集上,遗传算法和SFFS算法性能接近,ICMFS1和ICMFS2好于遗传算法和SFFS算法;在维数最大的Sonar数据集上,遗传算法好于经典的SFFS,ICMFS1和ICMFS2又好于遗传算法,表明ICMFS在高维数据集上具有很好的搜索能力.从图1可以清楚地发现,比较好的解是首先在记忆单元中发现的,这表明局部算子具有较好的搜索能力.

为了比较算法的鲁棒性,本文采用文献[7]使用过的准则:一种算法在某个数据集上的相对性能用 $b_m$ 表示, $p_m$ 表示这种算法在某个数据集上的平均

性能,即 $b_m = \frac{p_m}{\max_k p_k}$ ,把一个算法在所有数据集上的 $b_m$ 相加求和,如果和的值大,就表明这种算法的鲁棒性强.图2显示了前面5种算法在UCI数据集上的鲁棒性.

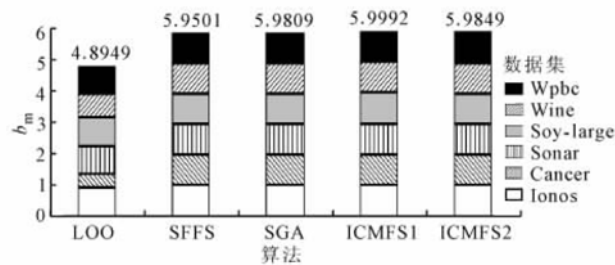


图2 算法鲁棒性比较

在图2的柱形图最上面的数值表示 $b_m$ 之和,下面标注的是算法,可以清楚看出本文提出的免疫记忆克隆算法ICMFS1和ICMFS2的鲁棒性好于SFFS和SGA.

## 4 结束语

本文提出了一个基于免疫记忆克隆的封装式特征选择算法,利用抗体种群进行全局搜索,记忆单元嵌入了可控制搜索深度的局部搜索算子,用来加速收敛速度,在UCI数据集上的仿真结果验证了本文算法的鲁棒性.当然,算法还有许多问题,如免疫记忆克隆算法的参数选择、局部搜索的深度和方向等有待继续研究.

## 参考文献:

- [1] RAYMER M L, PUNCH W F, GOODMAN E D, et al. Dimensionality reduction using genetic algorithms [J]. IEEE Trans Evolutionary Computation, 2000, 4 (2): 164-171.
- [2] 张向荣,焦李成. 基于免疫克隆选择算法的特征选择 [J]. 复旦学报(自然科学版), 2004, 43(5): 926-929. ZHANG Xiangrong, JIAO Licheng. Feature selection based on immune clonal selection algorithm [J]. Journal of Fudan University (Natural Science), 2004, 43 (5): 926-929.
- [3] DASH M, LIU H. Feature selection for classification. [J]. Intelligent Data Analysis, 1997, 1(3): 131-156.
- [4] 杜海峰,公茂果,焦李成,等. 用于高维函数优化的免疫记忆克隆规划算法 [J]. 自然科学进展, 2004, 14(8): 925-933. DU Haifeng, GONG Maoguo, JIAO Licheng, et al.

- Immune memory clonal programming algorithm for high-dimensional function numerical optimization [J]. *Progress in Natural Science*, 2004, 14(8): 925-933.
- [5] BLAKE C L, KEOGH E, MERZ C J. UCI repository of machine learning databases [M]. Irvine, CA, USA: University of California, 1998.
- [6] OH I S, LEE J S, MOON B R. Hybrid genetic algorithms for feature selection [J]. *IEEE Trans on Pattern Analysis and Machine Intelligence*, 2004, 26(11): 1424-1437.
- [7] VLACHOS M, DOMENICONI C, GUNOPULOS D, et al. Non-linear dimensionality reduction techniques for classification and visualization [C]//*Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, USA: ACM Press, 2002: 645-651.

(编辑 刘杨)