基于免疫编码的图像特征选择方法

曹 琼,郑 红,李行善
(北京航空航天大学自动化与电气工程学院,北京 100191)

摘 要: 针对目标与背景两类图像模式识别问题,在已有的特征选择方法基础上,提出了一种新颖的基于免疫分子编码机理的图像特征选择方法(Immune Antibody Construction Algorithm, IACA). 该方法借鉴生物免疫系统的抗体分子编码机理,在对样本进行参数估计情况下,提出熵度量单个特征对于目标和背景的识别敏感度;从集合的角度研究并且定义了特征之间的包含和互补关系;并且基于组成抗体分子氨基酸结合能量最小原则,提出了关于图像目标的免疫抗体构建规则;最终实现了寻找最优特征子集的算法 IACA,该特征子集的维数通过算法自动获得无需人为设定,选择结果为目标的"免疫抗体",能很好的从背景中识别目标. 利用归纳法证明了用 IACA 得到的特征子集的最优性. 与其他特征选择方法比较,测试结果显示该算法具有较低的计算复杂度和错误识别率,表明了该方法的优越性和先进性.

关键词: 图像目标识别; 特征选择; 生物免疫
中图分类号: TP751 **文献标识码**: A **文章编号**: 0372-2112 (2009) 03-0562-05

Image Feature Selection Method Based on Immune Encoding Mechanism

CAO Qiong, ZHENG Hong, LI Xing-shan
(School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China)

Abstract: Aiming at two-classes image pattern recognition problem of object and background, a novel image feature selection method, named immune antibody construction algorithm (IACA) is proposed, inspired by the biological immune antibody encoding principle. In the case of sample parameter estimation, IACA considers entropy to measure individual feature's sensitivity of object and background, and defines the inclusion and complementary formulas about multi-features in set theory perspective. Guided by the minimum energy principle, image immune antibody construction rules and corresponding algorithm are proposed to find an optimized feature subset as object immune antibody. Furthermore, the dimension of the subset can be automatically determined without prior setting. The induction proved the result was the optimal feature subset. Data testing result shows that IACA has a lower computational complexity and error recognition rate than other methods, which has verified the superiority and the advanced nature of the method.

Key words: image object recognition; feature selection; biological immune

1 引言

基于图像信息的目标识别方法是模式识别、计算机视觉、人工智能领域的重要研究方向之一. 图像目标的识别是对于目标和背景两类模式进行识别的过程,而图像特征选择是决定识别效果的关键,其主要目标是:在给定的众多图像特征中,选择其中的重要特征以减少特征数量,同时尽量保留分类信息. 因此如何选择具有充分识别信息的最小特征子集,对于图像目标识别显得尤为重要.

目前已有的特征选择方法可分为四大类:穷举法、启发式方法、随机方法和神经网络方法. 其中穷举法遍历特征空间中所有的特征子集寻找最优子集,其计算复杂度为 $O(2^N)$ (N 为特征空间维数),因计算量太大而不实用;启发式方法采用人工机器调度规则,递增产生特征子集,虽然计算简单快速,但是只能够获得近似最优解,常见的启发式方法有顺序向前(向后)选择、决策树法[1]、Relief 方法[2]等;随机方法通过机器随机产生子集,需要通过设置最大迭代次限制算法复杂度,以此决定能否得到最优解,常见的随机方法有 Tabu 搜索法[3]、遗传算法[4]、模拟退火算法等. 另外,神经网络[5]方法也被用来进行特征的选择和删减. 这些方法中除了穷举法,其他方法均以搜索结果的近似为代价换取算法的简化,因此,都不能保证获取结果的最优. 究其原因,主要

收稿日期:2008-04-01;修回日期:2008-11-25
基金项目:国家自然科学基金(No. 60543066)

是这些方法都拘泥于特征之间相互关系未知的情况下,以整体结果的全局误差为目标函数,进行寻优运算,因此,它们无法利用不同特征相互作用的关系.而这些相互关系则是决定最优识别集合的关键,所以,现有各类算法在可接受计算复杂度下选取最优特征集合的根本问题仍未能得到很好解决.

生物免疫系统可以通过学习,对外来"非我"抗原产生相应抗体,抵御病毒侵扰.它极强的免疫识别功能主要原因在于所有生物免疫抗原均由 20 多种氨基酸通过不同的排列编码构成,生物抗原特异性由氨基酸编码形式体现.因此,氨基酸的种类、性质、排序及化学结合方式构成了一系列生物免疫识别的基础.对于外来"非我"抗原,免疫系统可由其特异性学习编码结构,产生抗体中和"非我"抗原的抗体,并形成对于同类"非我"抗原的长期免疫.问题的实质是生物抗体编码的基础是对有限氨基酸性质及其编码的整体充分了解,使得生物抗原描述具有完全并且统一的表述基础——氨基酸组合排列方式[6].

借鉴生物免疫编码的这种特性,本文提出了一种图像特征的"抗体编码"选择方法 IACA(Immune Antibody Construction Algorithm),将特征相结合的过程类比组成抗体分子的"氨基酸"结合的过程;定义了一系列概念、表达和定理,并且设计相应算法来获得最优特征子集;从理论上证明了结果的正确性.最后设计实验,与启发式方法中的顺序法,随机方法中的模拟退火法以及神经网络方法作比较,从识别率和时间效率上进一步说明该方法的优点和先进性.

2　IACA 方法

2.1　若干概念的定义

特征选择　从 N 个特征的集合中选出尽可能小的 M 个特征组成子集,满足 $M \leq N$,所选特征子集应具有不显著减低分类精度、不影响类分布、稳定和适应性强的特点[7~9].

图像模式类　用(w_1, w_2)表示目标和背景两类模式,其中,w_1 为目标类,w_2 为背景类.

图像特征完全集合　从不同的物理和数学角度统计得到的所有能对图像目标和背景进行区分的特征组成的集合,记为:

$$F = \{ f_1, f_2, \cdots, f_D \} \tag{1}$$

其中:f_i 为第 i 个特征且 $i = 1, 2, \cdots, D$, D 为特征个数.

图像样本集合　将每幅图像样本按照一定尺度分割,得到总的训练样本个数为 N,其中目标样本数 N_1,背景样本数 N_2,对这两类样本分别进行编号存储,得到目标尺度样本集合:

$$O = \{ o_1, o_2, \cdots, o_{N_1} \} \in w_1 \tag{2}$$

其中 o_i 为目标尺度块, $i = 1, 2, \cdots, N_1$.

背景尺度样本集合:

$$B = \{ b_1, b_2, \cdots, b_{N_2} \} \in w_2 \tag{3}$$

其中 b_i 为目标尺度块, $i = 1, 2, \cdots, N_2$.

类敏感度　每个特征所具有的识别目标和背景类别的能力,可用信息量表示.

类可分性　特征或特征集合区分图像目标和背景的能力.

2.2　特征分类性能的集合表达

特征 f_i 对样本数据进行分类,得到 4 类子集:

·正确归入到目标类的目标样本子集 O_{ri}

·错误归入到背景类的目标样本子集 O_{ei}

·正确归入到背景类的背景样本子集 B_{ri}

·错误归入到目标类的背景样本子集 B_{ei}

下标中 r 指正确样本, e 指错误样本.设 t 为分类阈值,对于目标类则有:

$$O_{ri} = \{ o_{ki} \mid f(o_{ki}) \leq t, k = 1, 2, \cdots, N_{ri}^1 \} \tag{4}$$

$$O_{ei} = \{ o_{ki} \mid f(o_{ki}) > t, k = 1, 2, \cdots, N_{ei}^1 \} \tag{5}$$

且满足: $N_{ri}^1 + N_{ei}^1 = N_1$, $O_{ri} \cup O_{ei} = O$, $O_{ri} \cap O_{ei} = \varnothing$;

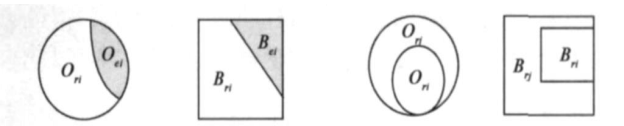
同样地,对于背景类则有:

$$B_{ri} = \{ b_{ki} \mid f(b_{ki}) \geq t, k = 1, 2, \cdots, N_{ri}^2 \} \tag{6}$$

$$B_{ri} = \{ b_{ki} \mid f(b_{ki}) < t, k = 1, 2, \cdots, N_{ei}^2 \} \tag{7}$$

且满足: $N_{ri}^2 + N_{ei}^2 = N_2$, $B_{ri} \cup B_{ei} = B$, $B_{ri} \cap B_{ei} = \varnothing$.

上述各式用集合表示如图 1 所示:



(a) 对目标的划分　(b) 对背景的划分　(a) $O_{ri} \subseteq O_{rj}$　(b) $B_{ri} \subseteq B_{rj}$

图1　特征f_i对样本集合的划分　　图2　特征f_j包含f_i

图 1 中,空白部分为正确分类样本,阴影部分为错误分类的样本.特征 f_i 和 f_j 之间是基于识别的集合关系描述,它们的物理性能、几何性质可以完全没有关系,但是,从将目标与背景分离的角度分析,它们之间的关系可以用识别结果的集合关系描述如下.

包含:对于特征 f_i 能够完全正确区分的样本,如果特征 f_j 也能够正确区分,即同时满足 $O_{ri} \subseteq O_{rj}$ 和 $B_{ri} \subseteq B_{rj}$,则称 f_i 包含于 f_j(或者 f_j 包含 f_i),如图 2 所示.

互补性:特征 f_i 不能正确区分的样本中,特征 f_j 能区分的样本集合称为特征 f_j 与特征 f_i 的互补集合,记为 C_{ij},有:

$$C_{ij} = (O_{ci} \cap O_{rj}) \cup (B_{ei} \cap B_{rj}) \cup (O_{ej} \cap O_{ri}) \cup (B_{ej} \cap B_{ri}) \tag{8}$$

集合 C_{ij} 越大,说明 f_j 与 f_i 的互补性越大.

已知单个特征对样本集合的划分,通过集合运算,可得到特征集合 $I=\{f_1,f_2,\cdots,f_{k-1}\}$ 对目标和背景样本集合的分类:

目标错误归类样本集合:$O_{eI}=\bigcap_{i=1}^{k-1}O_{ei}$ (9)

目标正确归类样本集合:$O_{rI}=O-O_{eI}$ (10)

背景错误归类样本集合:$B_{eI}=\bigcap_{i=1}^{k-1}B_{ei}$ (11)

背景正确归类样本集合:$B_{rI}=B-B_{eI}$ (12)

两两特征之间的互补性定义扩展到特征 f_k 和 I 时,它们的互补集合定义为:

$$C_{Ik}=(O_{eI}\cap O_{rk})\cup(B_{eI}\cap B_{rk})\cup(O_{ek}\cap O_{rI})\cup(B_{ek}\cap B_{rI}) \tag{13}$$

2.3 特征集合运算和性质

数据预处理:对所有图像样本,计算 F 中特征 f_i 的值,得到 D 个 N 维的特征值向量:

$$(x_{1i},x_{2i},\cdots,x_{Ni})=(f_i(o_1),\cdots,f_i(o_{N_1}),f_i(b_1),\cdots,f_i(b_{N_2})) \tag{14}$$

其中:x_{ji} 为第 j 个样本关于特征 f_i 的值,$i=1,2,\cdots,D$,$j=1,2,\cdots,N$. 分别对它们进行归一化处理[10],使每个特征向量均具有零均值和单位方差. 在有限样本情况下,计算各个向量中目标样本和背景样本特征,并统计其频数分布直方图,然后曲线拟合特征 f_i 关于模式 w_1 的类条件概率密度函数 $p(x_i|w_1)$ 曲线和模式 w_2 的类条件概率密度函数 $p(x_i|w_2)$ 曲线. 得到这两条曲线的交点为 t.

类敏感度计算:用香农熵计算特征对目标和背景两类模式的敏感性,记 f_i 对目标的敏感度为 H_{i1},对背景的敏感度为 H_{i2}.$|H_{i1}|$ 或 $|H_{i2}|$ 值越大,说明该特征对于该类更敏感. 计算公式为:

$$H_{ij}=-\int p(w_j|x_i)\log_2 p(w_j|x_i)\,\mathrm{d}x,\ j=1,2 \tag{15}$$

其中:$p(w_j|x_i)=\dfrac{p(x_i|x_j)\,p(w_j)}{\sum_{j=1}^{2}p(x_i|w_j)\,p(w_j)}$,$p(x_i|x_j)=\dfrac{k_x}{N_j}$,且 $p(w_j)=\dfrac{N_j}{N}$.

特征的类可分性计算:定义度量特征 f_j 与特征 f_i 的类可分性公式如下:

$$J_{ij}=-\log_2 e_{ij} \tag{16}$$

其中 e_{ij} 为分类误差,当 $i=j$ 时,

$$e_i=\frac{|O_{ei}\cup B_{ei}|}{|O\cup B|} \tag{17}$$

(集合的个数称为集合的基数,用$|\ |$表示). 当 $i\neq j$ 时,

$$e_{ij}=\frac{|(O_{ei}\cup B_{ei})\cup(O_{ej}\cup B_{ej})-C_{ij}|}{|O\cup B|}=\frac{|(O_{ei}\cap O_{ej})\cup(B_{ei}\cap B_{ej})|}{|O\cup B|} \tag{18}$$

特别的,当特征子集 $I=\{f_1,f_2,\cdots,f_{k-1}\}$ 与特征 f_k 结合时,公式(16)仍然适用,将 e_{ij} 调整为联合分类误差:

$$e_{(I)(k)}=\frac{|(\bigcap_{i=1}^{k}O_{ei})\cup(\bigcap_{i=1}^{k}B_{ei})|}{|O\cup B|} \tag{19}$$

以上从集合论的角度对单个特征及多特征的分类能力和相关互补性能进行了定义,从识别角度解释了特征的分类能力以及多个特征之间的关系,这也是下述 IACA 方法的数学基础.

2.4 IACA 算法实现

比照生物免疫抗体编码原理,IACA 将图像目标比作生物免疫系统的"抗原",把图像的特征比作免疫抗体基元,而图像特征选择的过程则看成是得到相应"免疫抗体"的过程. 它是逐次递增的方法,下次特征选择在很大程度上依赖于上一次的结果,它们是一组序列相关的动作,每次都考虑所选特征对已入选特征之间的互补性,只有在能达到一定互补效果时,才增加新的特征,然后用特征的类别可分离性函数 J_n 度量新的特征子集也即抗体"氨基酸"分子结合的亲和度,并且,用类敏感度作为"氨基酸"分子的"构象"选择下一步与之结合的特征,使抗体基元组合能进一步"中和"外来"抗原",即新的特征子集能够更好的识别图像目标. 算法流程图如图 3.

在进行特征选择前,首先进行样本数据预处理,得到特征集合 F 中每个特征对 O 和 B 的集合划分,按照公式(16)和(17)计算类可分性 J_i,$i=1,2,\cdots,D$. 设其中最大的值为 J_1,则选取它对应的特征 f_1 为第一个免疫抗体基元,它的分子亲和度为 $J(1)=J_1$. 设第 $j(j\geq 1)$ 次特征选择结束后得到特征子集 $I_j=\{f_1,f_2,\cdots,f_j\}$,剩下的可选的特征组成的集合为定义为

$$A_j=F-I_j \tag{20}$$

设它的维数 D_j,并且定义互补性下限计算公式为:

$$\delta_j=\frac{N}{D_j\times 2^D} \tag{21}$$

计算 I_j 对样本集合的划分 $O=O_{rI}\cup O_{eI}$ 和 $B=B_{rI}\cup B_{eI}$. 当 $j=1$ 时,就是初始选择第一个特征 f_1 的情况,有 $O_1=D-1$,并且剩下的可选特征集合为 $A_1=F-f_1$.

3 IACA 最优性证明

IACA 算法结果最优性定理 对维数为 D 的完全特征集合 $F=\{f_1,f_2,\cdots,f_D\}$,利用目标和背景样本,进行 IACA 特征选择,得到的特征子集 $I_d=\{f_1,f_2,\cdots,f_j\}$ 为最优特征子集,即使得 IACA 定义的可分性判据 J 最

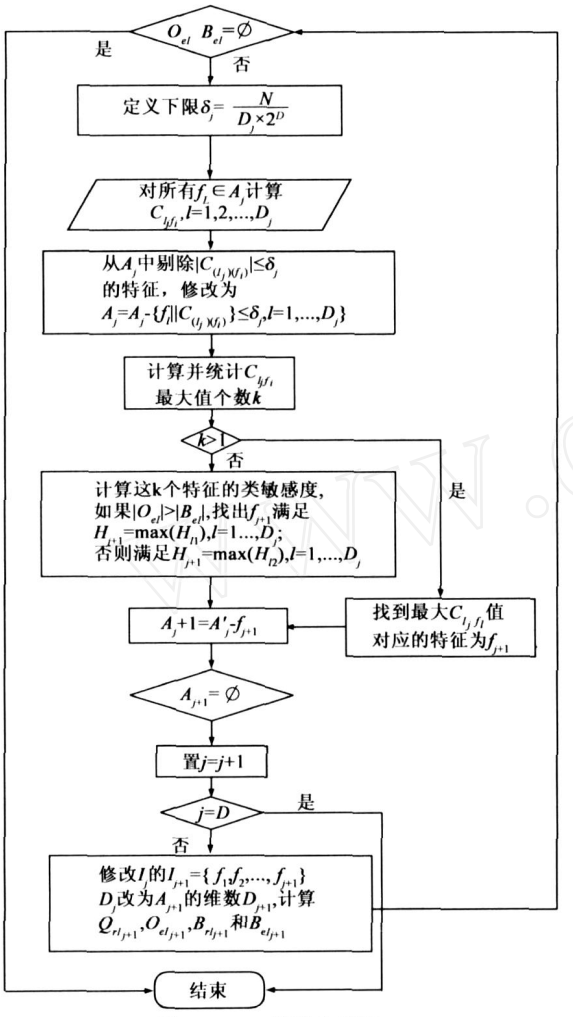


图3　IACA算法流程图

大.

证明:(1) 当满足算法结束条件 $O_{eI}\cap B_{eI}\Phi$ 时,由定义知,此时 $J(d)=\infty$,显然 I_d 为最优特征子集.

(2) 当满足条件 $A_{j+1}=\emptyset$ 时,算法结束,用归纳法证明特征子集的最优性.

当 $j=1$ 时,对于第一个抗体基元 f_{k1} 的选择,由算法过程知,它具有最大的类别可分离性,所以它是从原始特征集合 F 中挑选一个特征时的最优结果;

当 $1<j<d$ 时,假设此时特征子集为 I_j 是最优的,则满足:从 F 中任意挑选 j 个特征组成的子集 I_j',均有 $J(I_j')<J(I_j)$. 选择第 $j+1$ 个特征时,对 $\forall f_i\in A_j'$,它和 I_j 构成的新特征子集 $I_{j+1}'=\{I_j,f_i\}$,由算法有:

$$J(I_{j+1}')=-\log_2 e_{(I_j)(f_i)} \tag{22}$$

又因为:

$$e_{(I_j)(f_i)}=e_{I_j}-|C_{(I_j)(f_i)}| \tag{23}$$

则:

$$\Delta J_i=J(I_{j+1}')-J(I_j)=\log_2\frac{e_{I_j}}{e_{(I_j)(f_i)}}=\log_2\frac{e_{I_j}}{e_{I_j}-|C_{(I_j)(f_i)}|} \tag{24}$$

所以 ΔJ 是关于互补集合大小的单调递增函数. 算法选择的特征 f_{j+1} 满足:

$$|C_{(I_j)(f_k)}|=\max(|C_{(I_j)(f_i)}|),\ i=1,2,\cdots,D_j \tag{25}$$

故 ΔJ_{j+1} 最大,即 f_{j+1} 使得 $J(I_{j+1})$ 最大,所以 I_{j+1} 是从 F 中选择 $j+1$ 个特征时的最优特征子集.

4　IACA 测试及与其它方法比较

4.1　测试数据

为了验证正文算法的性能,我们利用车载图像数据,对于图 3 中的图像,识别座位上有人与否.

(a)空座位　　(b)有人座位

图4　样本图像

对采集到的图像中的座位识别"有人"和"无人"两种模式[11],图 4 中条形区域(1,2,5),提取它们关于灰度和空间的纹理以及形状轮廓等共 16 个特征并进行编号,从①到⑯,如表 1.

表 1　所使用特征及其编号

编号	①-④	⑤-⑧	⑨-⑫
特征名称	灰度共矩阵的对比度、熵、相关性以及角二阶矩.	两水平方向链码的长度及角度.	两垂直方向链码的长度及角度.
编号	⑬-⑭	⑮	⑯
名称	熵/相关性	分形	二阶矩

4.2　测试结果

采用 200 幅"有人"模式图像样本,200 幅"无人"模式图像样本,我们分别用顺序法(前向),模拟退火法,BP 神经网络法和 IACA 来进行特征选择. 用顺序法,模拟退火法,神经网络法等方法进行特征选择之前需要人为设定待选择的特征子集的维数 d,计算过程中,我们设定 $d=6$. 而 IACA 不需要事先设定维数. 结果如下所示:

表 2　不同特征选择算法得到的特征子集结果

特征选择方法名称	是否给定子集维数	选择结果(特征编号)
顺序法(前向)	给定 $d=6$	①④⑦⑫⑯
模拟退火	给定 $d=6$	①④⑦⑨⑩
BP 神经网络	给定 $d=6$	③④⑤⑥⑦⑨
IACA	无需给定	③④⑮

4.3　性能比较

我们对这些方法结果进行性能测试,分别用表 2 中的特征子集来对不同于训练样本图像的另外 400 幅

车载图像进行分类,得到如下检测结果:

表 3 检测结果

	虚警率(%)	误警率(%)	识别率(%)	计算复杂度
顺序法	10	14.4	75.6	$O(D\log_2 D)$
模拟退火	6.6	8.3	85.1	$O(d^2t(D))$
神经网络	5.0	4.2	90.8	$O(D^2)$
IACA	0	1.1	98.9	$O(D\log_2 D)$

注:$t(D)$ 为 D 的多项式.

由于采集的图像背景复杂,光线变化剧烈,用常规方法选择得到的特征难以对目标进行精确的描述,从实验可以看出,IACA 法的计算复杂度与顺序法相当,小于其他两种方法,但是其特征数少于后者;同时 IACA 用最少的特征子集得到了最佳的分类效果,识别率最高.

5 结论

对于图像两类模式的特征选择问题,本文首先基于集合理论讨论了样本空间特征可分性问题,通过类比生物免疫系统抗体分子编码原理给出了关于单个特征类别的集合可分性能定义及其判别条件,然后结合分子抗体和抗原结合过程,探讨了多特征之间的结合亲和度问题.以特征子集的有效率和特征之间补充性为切入点,提出了一种新的特征选择算法 IACA.

同其他的特征选择方法相比,IACA 方法有如下优点:目前的方法都是人为确定要选择的特征的个数,IACA 可以"主动"地寻找最优特征子集的大小 M,增加了合理性;目前方法中使用的可分性判据没有考虑新加入(或剔除)的特征的影响,每次都需要从头进行"黑盒"计算,IACA 定义了一种与选择序列相关的可分性判据,不仅考虑特征与特征之间的关系,还考虑特征与特征子集之间的相关性和补充性,这对于简化计算量,消除特征的冗余和不相关有更好的效果.

本文不仅从理论上证明了 IACA 算法的特征子集选择结果具有最优性,而且针对车载图像乘客识别的特征选择及识别问题,利用 IACA 算法与顺序法、模拟退火法、神经网络法分别进行特征选择及识别性能比较实验,验证了 IACA 算法在识别正确率和计算复杂度方面的有效性.

参考文献:

[1] Cardie C. Using decision trees to improve case-based learning [A]. Proc of 10th In'l Conf on Machine Learning [C]. Amherst, Massachusetts, USA: Morgan Kaufmann Publishers Inc, 1993. 1. 25 - 32.

[2] Kononenko I. Estimating attributes: analysis and extension of relief [A]. Proc of European Conf on Machine Learning [C]. Catania, Italy: Springer-Verlag GmbH & Company KG, 1994. 1. 171 - 182.

[3] Glover F, Hanafi S. Tabu search and finite convergence [J]. Discrete Applied Mathematics, 2002, 119(3): 3 - 36.

[4] Chakraborty B. Genetic algorithm with fuzzy fitness function for feature selection [A]. Proceedings of the 2002 IEEE International Symposium on Industrial Electronics [C]. L'Aquila, Italy: Institute of Electrical and Electronics Engineers, 2002. 1. 315 - 319.

[5] Wang W J, Jones P, Partridge D. A comparative study of feature-salience ranking techniques [J]. Neural Computation, 2001, 13(2): 1603 - 1623.

[6] Jin Bo-quan. Cellular and Molecular Immunology [M]. Beijing: Science Press, 2001. 96 - 98.

[7] Batiti R. Using mutual information for selecting features in supervised neural network learning [J]. IEEE Transactions on Neural Networks, 1994, 5(8): 537 - 550.

[8] 边肇祺,张学工. 模式识别(第 2 版) [M]. 北京:清华大学出版社, 2000: 215 - 260.

[9] Brunzell H, Erikcson J. Feature reduction for classification of multidimensional data [J]. Pattern Recognition, 2000, 33(4): 1741 - 1748.

[10] Sergios T, Konstantinos K. Pattern Recognition. Second Edition [M]. USA: Elsevier Science, 2003. 164 - 166.

[11] LI Hang-xi, ZHENG Hong, WANG Yang, A novel clustering-based algorithm for curve detection and its application to passenger recognition [A]. IEEE Conference on Industrial Electronics and Application [C]. Harbin, China: IEEE Press, 2007. 2758 - 2763.

作者简介:

曹 琼 女,1982 年 10 月出生于安徽省宿松县.博士.主要研究方向:图像处理、模式识别、人工智能等.
Email: caoqiong.gl@gmail.com

郑 红 女,1961 年 11 月出生于河北省.教授,博士生导师.主要研究方向:模式识别及智能检测. Email: julyanna@vip.sina.com.cn

李行善 男,教授,博士生导师.主要研究方向:检测技术与自动化装置,电力电子,微机应用.