基于流形距离的人工免疫无监督分类与识别算法

公茂果 [1] 焦李成 [1] 马文萍 [1] 张向荣 [1]

摘 要 将一种新的流形距离作为相似性度量测度, 提出了一种用于无监督分类与识别的人工免疫系统方法. 通过基于流形距离的相似性度量, 有效利用样本集固有的全局一致性信息, 充分挖掘无类属样本的空间分布信息, 对样本进行类别划分. 新方法将免疫响应过程建模为一个四元组 $AIR = (G, I, R, A)$, 其中 G 为引发免疫响应的外界刺激, 即抗原; I 为所有可能抗体的集合; R 为抗体间相互作用的规则集合; A 为支配抗体反应、指导抗体进化的动态算法. 针对无监督分类问题, 将抗体编码为代表各类别的典型样本序号的排列, 利用动态算法 A 搜索能代表各类别的典型样本的最佳组合. 将新方法与标准的 K- 均值算法、基于流形距离的进化聚类算法以及 Maulik 等人提出的基于遗传算法的聚类算法进行了性能比较. 对 6 个人工数据集及手写体数字识别问题的仿真实验结果显示, 新方法对样本空间分布复杂的无监督分类问题和实际的模式识别问题具有较高的准确率和较好的鲁棒性.

关键词 人工免疫系统, 流形, 无监督分类, 聚类, 模式识别
中图分类号 TP18

Unsupervised Classification and Recognition Using an Artificial Immune System Based on Manifold Distance

GONG Mao-Guo[1] JIAO Li-Cheng[1] MA Wen-Ping[1] ZHANG Xiang-Rong[1]

Abstract In this study, a novel artificial immune system algorithm for unsupervised classification and recognition is proposed by using a novel manifold distance based dissimilarity measure which can measure the geodesic distance along the manifold. The new method formulizes the immune response as a quaternion $AIR = (G, I, R, A)$, where G denotes exterior stimulus or antigen, I denotes the set of valid antibodies, R denotes the set of reaction rules describing the interactions between antibodies, and A denotes the dynamical algorithm describing how the reaction rules are applied to antibody population. In order to solve unsupervised classification problems, the new method encodes each antibody as a sequence of real integer numbers representing the cluster representatives, and searches the optimal cluster representatives from a combinatorial optimization viewpoint using the dynamical algorithm A. Experimental results on six artificial datasets with different manifold structures and the USPS handwritten digit datasets show that the novel algorithm has the ability to identify complex non-convex clusters, compared with the K-means algorithm, a genetic algorithm-based clustering proposed by Maulik, and an evolutionary clustering algorithm with the manifold distance.

Key words Artificial immune systems, manifold, unsupervised classification, clustering, pattern recognition

人工免疫系统 (Artificial immune systems, AIS) 是受免疫学启发, 模拟免疫学功能、原理和模型来解决复杂问题的自适应系统[1], 已经成为人工智能领域理论和应用研究的热点[2]. 近几年, 国内外很多学者相继提出了自己的人工免疫系统算法或模型[3−6]. 早在上世纪 80 年代中期, Farmer 等[7] 率先基于免疫网络学说给出了免疫系统的动态模型, 并探讨了免疫系统与其他人工智能方法的联系, 开始了人工免疫系统的研究. 但是, 此后的研究成果比较少见. 直到 1996 年 12 月, 在日本举行了基于免疫系统的国际专题讨论会, 首次提出了 "人工免疫系统" 的概念. 随后, 人工免疫系统的相关研究迅速展开, 有关论文和研究成果逐年增加. 1997 和 1998 年 IEEE Systems, Man, and Cybernetics 国际会议组织了相关专题讨论, 并成立了人工免疫系统及其应用分会. 随后, 一些人工智能领域著名的国际会议 (如 International Joint Conference on Artificial Intelligence (IJCAI), International Joint Conference on Neural Networks (IJCNN), IEEE Congress on Evolutionary Computation (CEC), Genetic and Evolutionary Computation Conference (GECCO) 等) 也相继开辟了人工免疫系统专题. 从 2002 年开始, 在英国、意大利、加拿大等地连

收稿日期 2007-07-03 收修改稿日期 2007-09-21
Received July 3, 2007; in revised form September 21, 2007
国家自然科学基金 (60703107), 国家高技术研究发展计划 (863 计划) (2006AA01Z107), 国家重点基础研究发展计划 (973 计划) (2006CB705700), 西安电子科技大学研究生创新基金 (创 05004) 资助
Supported by National Natural Science Foundation of China (60703107), National High Technology Research and Development Program of China (863 Program) (2006AA01Z107), National Basic Research Program of China (973 Program) (2006CB705700), and Graduate Innovation Fund of Xidian University (05004)
1. 西安电子科技大学智能信息处理研究所 西安 710071
1. Institute of Intelligent Information Processing, Xidian University, Xi′an 710071
DOI: 10.3724/SP.J.1004.2008.00367

续召开了六届人工免疫系统国际会议. 总之, 人工免疫系统结合了分类器、神经网络和机器推理等系统的一些优点, 具有解决复杂问题的潜力[2], 已经逐渐受到广大研究者的重视.

聚类, 即无监督分类, 是一种重要的数据分析方法, 已经被广泛应用于计算机视觉、信息检索、数据挖掘和模式识别等领域. 在现有的聚类方法中, 基于目标函数的聚类算法由于把聚类问题归结为一个优化问题, 具有深厚的泛函基础, 是聚类算法研究的重要分支之一, K- 均值算法[8] 就是其中最典型的方法. 由于 K- 均值算法的聚类目标函数是高度非线性和多峰的函数, 因此, 标准的 K- 均值算法用梯度下降法优化目标函数时, 搜索方向总是沿着能量减小的方向, 使算法很容易陷入局部极值点, 只有当初始化较好时, 算法才能收敛到全局最优解. 作为一类有效的全局优化技术, 进化计算已经被很多学者用于聚类问题[9−12]. 在设计基于进化计算的聚类算法时, 最核心的两个问题就是进化个体的编码以及相似性度量. 针对聚类问题的个体编码方式有很多, 其中使用较多的是借用于 K- 均值算法的编码方式, 即每个个体只对 K 个聚类中心进行编码, 然后对数据样本按照其与聚类中心的相似性进行类别划分. 因此, 相似性度量对这类算法的性能有重要影响. 最简单的相似性度量应该是欧氏距离. 但是以欧氏距离作为相似性度量的进化聚类算法虽然在全局最优化性能上较传统的基于梯度下降的 K- 均值算法有较大提高, 但是同样存在一个重要的缺点, 即它们只对空间分布为球形或超球体的数据具有较好的性能, 而对空间分布复杂的流形结构的数据效果很差, 这是基于欧氏距离的相似性度量的缺陷导致的必然结果[13]. 因此, 为这类聚类算法设计一个更加合理的相似性度量是一项非常必要的工作. 为此, Su 和 Chou[13] 基于点对称概念提出了一种非公制测度 (Nonmetric measure) 作为相似性度量的准则. 它对具有明显的对称结构的复杂流形结构的数据具有很好的性能. 最近, Charalampidis[14] 针对含有方向信息的向量聚类问题, 设计了一种基于可变旋转向量的相似性度量准则, 并应用于 K- 均值算法.

针对聚类 (即无监督分类) 和无监督模式识别问题 (通称为无监督分类和识别问题), 本文通过引入一种基于流形距离的相似性度量和一种新的抗体编码方式, 提出了一种基于人工免疫响应模型的解决方案. 通过对 6 个具有不同流形结构的人工数据聚类问题的仿真实验, 考察了新算法与标准的 K- 均值算法[8]、基于流形距离的进化聚类算法[15] 以及 Maulik 等提出的基于遗传算法的聚类算法[10] 相比所具有的新特性. 最后, 本文将新方法应用于一个典型的模式识别问题, 即手写体数字识别问题中, 获得了较满意的结果.

1 流形距离及人工免疫响应模型的定义

1.1 流形距离

在现实世界的聚类问题中, 数据的分布往往具有不可预期的复杂结构, 导致了基于欧氏距离的相似性度量无法反映聚类的全局一致性 (即位于同一流形上的数据点具有较高的相似性[16]). 从图 1 所示的例子中可以形象地看出, 我们期望数据点 1 与 3 的相似性要比数据点 1 与 2 的相似性大, 这样才有可能将数据点 1 和 3 划分为同一类. 但是, 按照欧氏距离进行相似性度量时, 数据点 1 与 2 的欧氏距离要明显小于数据点 1 与 3 的欧氏距离, 从而导致了数据点 1 与 2 划分为同一类的概率要大于数据点 1 与 3 划分为同一类的概率. 也就是说, 用欧氏距离作为相似性度量时, 根本无法反映图 1 中所示数据的全局一致性. 因此, 对于现实世界中复杂的聚类问题, 简单地采用欧氏距离作为相似性度量会严重影响聚类算法的性能.

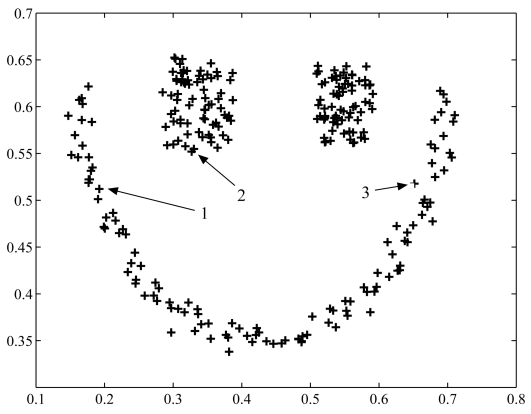


图 1 欧氏距离无法反映样本的全局一致性

Fig. 1 The Euclidian distance metric can not reflect the global consistency

基于以上考虑, 本文尝试设计一种能反映聚类全局一致性的相似性度量, 期望新的相似性度量能够打破在欧氏空间 “两点之间直线最短” 的定理, 使得两点间直接相连的路径长度不一定最短, 也就是说新的相似性度量并不一定满足欧氏距离下的三角不等式定理. 为了达到这一目的, 本文首先定义一个流形上的线段长度.

定义 1. 流形上的线段长度.

空间任意两点 x_i 与 x_j 之间流形上的线段长度 $L(x_i, x_j)$ 按下式计算

$$L(x_i, x_j) = \rho^{dist(x_i, x_j)} - 1 \tag{1}$$

其中, $dist(x_i, x_j)$ 为 x_i 与 x_j 之间的欧氏距离, $\rho < 1$ 为伸缩因子.

显然, 这样定义的线段长度可以满足上面的性质, 从而可以用来描述聚类的全局一致性. 根据流形上的线段长度, 我们可以进一步定义一个新的距离测度, 称为流形距离. 将数据点看作是一个加权无向图 $G = (V, E)$ 的顶点 V, 边集合 $E = \{W_{ij}\}$ 表示的是在每一对数据点间定义的流形上的线段长度, 则流形距离测度可定义如下:

定义 2. 流形距离测度.

将数据点看作是图 $G = (V, E)$ 的顶点, 令 $p \in V^l$ 表示图上一个长度为 $l = |p| - 1$ 的连接点 p_1 与 $p_{|p|}$ 的路径, 其中边 (p_k, p_{k+1}), $1 \le k < |p|$. 令 P_{ij} 表示连接数据点 x_i 与 x_j 的所有路径的集合, 则 x_i 与 x_j 之间的流形距离按下式计算

$$D(x_i, x_j) = \min_{p \in P_{i,j}} \sum_{k=1}^{|p|-1} L\left(p_k, p_{k+1}\right) \tag{2}$$

其中, $L(a, b)$ 表示两点间流形上的线段长度.

显然, 新的距离测度满足测度的四个条件, 即:

1) 对称性: $D(x_i, x_j) = D(x_j, x_i)$;

2) 非负性: $D(x_i, x_j) \ge 0$;

3) 三角不等式: 对于任意的 x_i, x_j, x_k, $D(x_i, x_j) \le D(x_i, x_k) + D(x_k, x_j)$;

4) 自反性: $D(x_i, x_j) = 0$, 当且仅当 $x_i = x_j$.

流形距离测度可以度量沿着流形上的最短路径, 这使得位于同一流形上的两点可以用许多较短的边相连接, 而位于不同流形上的两点要用较长的边相连接, 从而实现了放大位于不同流形上的数据点间的距离, 而缩短位于同一流形上的数据点间的距离的目的.

1.2　人工免疫响应模型

生物免疫系统的免疫响应过程可以简化为如下过程[6, 17−18]: 由抗原引发, 在免疫系统的控制下多种免疫细胞 (包括抗体和 T 细胞) 经过一系列反应, 逐步亲和度成熟, 产生相应的免疫效应, 清除抗原. 因此, 人工免疫响应可以被描述为一个四元组 $AIR = (G, I, R, A)$, 其中 G 为引发免疫响应的外界刺激, 即抗原; I 为所有可能抗体 (本文将抗体和 T 细胞不作区分, 统一抽象为抗体) 的集合; R 为抗体间相互作用的规则集合; A 为支配抗体反应、指导抗体进化的算法.

1) 抗原 G

在免疫学中, 抗原是一类能够诱导机体免疫响应并能与相应抗体或 T 细胞受体发生特异反应的物质. 在人工免疫响应四元组模型中, 抗原一般指问题及其约束. 与免疫学中抗原的作用类似, 它是诱导人工免疫响应的始动因子.

2) 抗体空间 I

集合 $I = \{\boldsymbol{b}_1, \boldsymbol{b}_2, \cdots, \boldsymbol{b}_n\}$ 被称作抗体空间, 是针对抗原 G 所有可能出现的抗体的集合, 其中 n 可以为无穷大的整数. 针对不同的抗原 G, 抗体 b 的表现形式不同, 例如可以是二进制码串、实数序列、抽象的符号序列、特征序列等. 抗体是人工免疫响应的基础, 其表现形式对抗体间相互作用的规则集合 R 的设计起着决定作用.

以字符串 $\boldsymbol{b} = b_1 b_2 \cdots b_l$ 为例, 依生物学术语, 抗体 $\boldsymbol{b}$ 中, b_i 被称为等位基因, 其可能取值与编码方式有关. 例如, 抗体结构为 8 位二进制数, 则位串 "0-1-1-1-0-1-0-0" 即代表一个抗体. 抗体群 $B = \{\boldsymbol{b}_1, \boldsymbol{b}_2, \cdots, \boldsymbol{b}_m\}$ 为抗体 $\boldsymbol{b}$ 的 m 元组, 是抗体空间 I 的一个子集, 正整数 m 称为抗体群规模.

3) 规则集合 R

抗体间相互作用的规则集合 $R = \{r_1, r_2, r_3, \cdots, r_l\}$ 描述了抗体空间 I 中抗体之间可能存在的作用形式. 一个规则 $r_i \in R$ 可以是从生物免疫系统中抗原与抗体间、抗体与抗体间的相互作用中启发得到. 对抗体群 $B = \{\boldsymbol{b}_1, \boldsymbol{b}_2, \cdots, \boldsymbol{b}_n\}$, 一个规则 $r_i \in R$ 可以简略地表示为

$$\boldsymbol{b}_1 + \boldsymbol{b}_2 + \cdots + \boldsymbol{b}_n \xrightarrow{r_i} \boldsymbol{b}_1^{'} + \boldsymbol{b}_2^{'} + \cdots + \boldsymbol{b}_m^{'} \tag{3}$$

其中 n, m 为正整数, m 的大小由规则 r_i 决定, 这里符号 "+" 并不是传统意义上的操作算子, 只是为了对式 (3) 中各抗体之间进行分隔 (除特殊说明外, 下文中 "+" 均为此义).

式 (3) 表示 n 个抗体经过规则 r_i 的作用, 演变成 m 个抗体. 为了细致地模拟生物免疫响应过程, 应该尽量详尽地设计足够多的规则, 从而实现对免疫响应过程的模拟.

4) 动态算法 A

动态支配算法 A 是模拟免疫系统中抗体进化过程以及支配人工免疫响应过程中抗体相互作用的算法, 包括规则集合 R 作用在抗体空间 I 中某一抗体群 B 上的具体方式、抗体 — 抗原亲和度的计算法则以及人工免疫响应终止条件的判断等.

2　用于无监督分类与识别的人工免疫响应算法设计

将人工免疫响应模型成功应用于无监督分类与识别问题, 需要解决抗体表示、抗体亲和度定义、抗体操作规则集合 R 的设计及动态算法 A 的设计等关键技术.

2.1　抗体表示

我们从组合优化的角度来考虑聚类问题, 将指

定类别数 K 的聚类问题建模为一个从数据集中选择 K 个典型样本来代表 K 个类别的优化问题, 然后按照无类属样本与 K 个典型样本的相似性, 对数据集进行类别划分. 因此, 每个抗体代表一种典型样本序号的组合, 显然, 对于一个 K 类的聚类问题, 一个抗体的长度为 K, 第一个基因位为代表第一个类别的样本序号, 第二个基因位为代表第二个类别的样本序号, 依此类推. 为了更加具体地说明这种新的抗体表示方式, 考虑以下简单的例子:

假设数据集大小为 100, 类别数目为 5, 则个体 $(6, 19, 38, 64, 91)$ 表示第 6、第 19、第 38、第 64、第 91 样本分别代表第 1 至第 5 类. 需要注意的是, 为了减少搜索空间, 我们将个体中每个基因位按照从小到大的顺序排列, 也就是说, 个体 $(6, 19, 38, 64, 91)$ 与个体 $(6, 19, 64, 38, 91)$ 将被视为一个个体.

显然, 这种编码方式没有涉及数据的维数, 因此, 搜索空间的大小与数据维数无关. 在 Maulik 等人提出的遗传聚类算法[10] 中, 将 K 个聚类中心编码为个体, 这样对于 m 维的数据聚类问题, 其编码长度为 $m \times K$, 且该编码方式决定了该算法为一个连续空间的优化问题. 而我们提出的编码方式, 编码长度为 K, 与 m 无关, 且为离散空间的优化问题, 降低了搜索空间的大小.

2.2 抗体亲和度定义

抗体亲和度反映抗体与抗原之间的结合力. 在人工免疫系统中, 一般指候选解对问题的适应性度量. 针对聚类问题, 抗体的亲和度值即抗体对应的类别划分的目标函数值. 首先, 根据抗体表示的各类别的典型样本, 以流形距离作为相似性度量, 将所有无类属的样本数据划分到不同的类别中. 将点 x_i, $i = 1, 2, \cdots, n$ 划分到类 $C_j, j \in \{1, 2, \cdots, K\}$, 遵循下列原则:

$$j = \arg \min_{j=1,2,\cdots,K} (D(x_i, u_j)) \tag{4}$$

其中, u_j 为代表第 j 类的典型样本.

类别划分完成之后, 则抗体的亲和度值按式 (4) 计算得到

$$Aff(b) = \frac{1}{1 + \sum\limits_{C_k \in C} \sum\limits_{i \in C_k} D(i, \mu_k)} \tag{5}$$

其中, C 为抗体 $\boldsymbol{b}$ 对应的类别划分, μ_k 为代表类别 C_k 的典型样本, $D(i, \mu_k)$ 为类别 C_k 中的第 i 个样本与 μ_k 之间的流形距离.

2.3 规则集合 $\boldsymbol{R}$ 的设计

受生物免疫响应过程的启发, 本文的规则集 R 主要包含克隆增殖操作 r_1、基因变异操作 r_2 和克隆选择操作 r_3.

2.3.1 克隆增殖操作 $\boldsymbol{r}_1$

在免疫学中, 克隆增殖是指通过无性繁殖 (如细胞丝分裂) 可连续传代并形成群体. 在人工免疫响应模型中, 对抗体种群 $B = \{\boldsymbol{b}_1, \boldsymbol{b}_2, \cdots, \boldsymbol{b}_n\}$ 的克隆增殖操作 r_1 定义为

$$\begin{aligned} &\boldsymbol{b}_1 + \boldsymbol{b}_2 + \cdots + \boldsymbol{b}_n \xrightarrow{r_1} \{\boldsymbol{b}_1^1 + \boldsymbol{b}_1^2 + \cdots + \boldsymbol{b}_1^{q_1}\} + \\ &\{\boldsymbol{b}_2^1 + \boldsymbol{b}_2^2 + \cdots + \boldsymbol{b}_2^{q_2}\} + \cdots + \{\boldsymbol{b}_n^1 + \boldsymbol{b}_n^2 + \cdots + \boldsymbol{b}_n^{q_n}\} \end{aligned} \tag{6}$$

其中, $\boldsymbol{b}_i^j = \boldsymbol{b}_i$, $i = 1, 2, \cdots, n$, $j = 1, 2, \cdots, q_i$, $q_i \in [1, n_c]$ 为一自适应参数, 也可以设定为一常数, n_c 为设定的克隆比例上限, $q_i = 1$ 表示对抗体没有进行克隆增殖操作. 可见, 上述克隆增殖过程与免疫学中的克隆增殖类似, 是简单的无性繁殖过程. 同一个抗体 $\boldsymbol{b}_i$ 经过克隆增殖后形成的子群体 $B_i(\boldsymbol{b}_i^1, \boldsymbol{b}_i^2, \cdots, \boldsymbol{b}_i^{q_i})$ 中的所有抗体与抗体 $\boldsymbol{b}_i$ 具有完全相同的属性.

2.3.2 基因变异操作 $\boldsymbol{r}_2$

基因变异操作 r_2 是对免疫系统学习识别外部模式、抗体基因变异和编辑过程的模拟. 对抗体种群 $B = \{\boldsymbol{b}_1, \boldsymbol{b}_2, \cdots, \boldsymbol{b}_n\}$ 的基因变异操作 r_2 定义为

$$\boldsymbol{b}_1 + \boldsymbol{b}_2 + \cdots + \boldsymbol{b}_n \xrightarrow{r_2} \boldsymbol{b}_1^{'} + \boldsymbol{b}_2^{'} + \cdots + \boldsymbol{b}_n^{'} \tag{7}$$

基因变异操作的基本内容是对抗体的某些基因位置上的基因值作变动. 本文中, 我们用以下简单的例子来说明抗体基因变异的具体操作.

设数据集大小为 100, 类别数目为 5, 抗体为 $(6, 19, 38, 64, 91)$, 随机产生一个 $[0, 1)$ 的随机数, 如果该随机数小于指定的变异概率, 且第二个基因位被选择为变异位, 则等概率将该个体变异为 $(6, 19 + \lfloor (100 - 19) * random + 1 \rfloor, 38, 64, 91)$ 或 $(6, 19 - \lfloor (19 - 1) * random + 1 \rfloor, 38, 64, 91)$, 其中 $random$ 表示 $[0, 1)$ 内均匀分布的随机数, $\lfloor \cdot \rfloor$ 表示向下取整. 如果变异过程中在同一抗体中产生了多个基因位数值相同, 则数值相同的基因位用变异前抗体中的相应基因位上的值代替, 保证变异后一个抗体中各个基因位上的数值不同, 且按照由小到大的顺序排列.

2.3.3 克隆选择操作 $\boldsymbol{r}_3$

克隆选择操作 r_3 是从抗体各自克隆增殖后的子代中选择优秀的个体, 从而形成新的种群, 是一个无性选择过程. 对抗体群 $B = \{\boldsymbol{b}_1^1, \boldsymbol{b}_1^2, \cdots, \boldsymbol{b}_1^{q_1}, \boldsymbol{b}_2^1, \boldsymbol{b}_2^2, \cdots, \boldsymbol{b}_2^{q_2}, \cdots, \boldsymbol{b}_n^1, \boldsymbol{b}_n^2, \cdots, \boldsymbol{b}_n^{q_n}\}$, 克隆选择操作 r_3 定义

如下

$$\{\boldsymbol{b}_1^1+\boldsymbol{b}_1^2+\cdots+\boldsymbol{b}_1^{q_1}\}+\{\boldsymbol{b}_2^1+\boldsymbol{b}_2^2+\cdots+\boldsymbol{b}_2^{q_2}\}+\cdots+\{\boldsymbol{b}_n^1+\boldsymbol{b}_n^2+\cdots+\boldsymbol{b}_n^{q_n}\}\xrightarrow{r_3}\boldsymbol{b}_1^{'}+\boldsymbol{b}_2^{'}+\cdots+\boldsymbol{b}_n^{'} \quad (8)$$

可见, 克隆选择操作是克隆增殖操作的逆操作. 同一个抗体 $\boldsymbol{b}_i$ 经过克隆增殖后形成的子群体在经过基因变异操作的编辑后, 通过克隆选择操作实现局部的亲和度升高. 具体地, $\forall i=1,2,\cdots,n,\exists j\in\{1,2,\cdots,q_i\}$, 若抗体 $\boldsymbol{b}_i^j$ 为子群体 $B_i(\boldsymbol{b}_i^1,\boldsymbol{b}_i^2,\cdots,\boldsymbol{b}_i^{q_i})$ 中亲和度最高的抗体, 则在子群体 $B_i(\boldsymbol{b}_i^1,\boldsymbol{b}_i^2,\cdots,\boldsymbol{b}_i^{q_i})$ 中 $\boldsymbol{b}_i^j$ 的选择压力最大, 即 $\boldsymbol{b}_i^{'}=\boldsymbol{b}_i^j$ 的概率最大, 本文设置 $p(\boldsymbol{b}_i^{'}=\boldsymbol{b}_i^j)=1$.

与生物进化系统有性别之分不同, 免疫系统是一个无性系统, 通过上述描述可以看出, 规则集 $R=\{r_1, r_2, r_3\}$ 中的三个规则均为无性操作过程, 可以理解为人工免疫响应对免疫系统这一特点的简单刻画.

2.4　动态算法 $\boldsymbol{A}$ 的设计

在人工免疫响应模型中, 将第 1.1 节描述的流形距离作为类别划分的相似性度量, 采用第 2.1 节设计的抗体编码方式, 支配规则集 $R=\{r_1,r_2,r_3\}$ 运作的动态算法的流程图如图 2 所示.

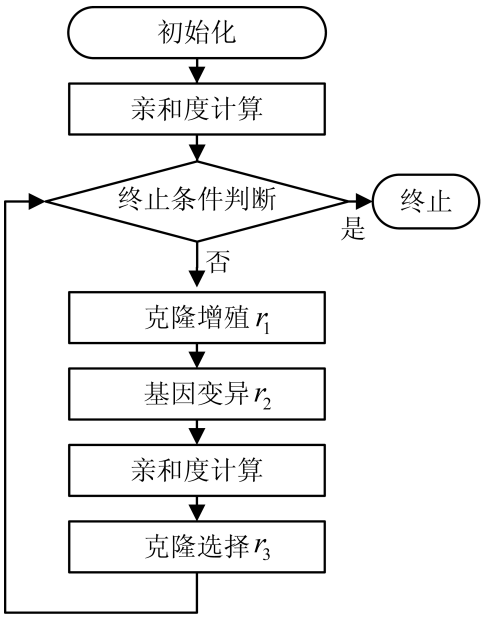


图 2　人工免疫响应无监督分类与识别算法流程图

Fig. 2　The flowchart of artificial immune unsupervised classification and recognition algorithm

初始化主要是指设置算法的初始参数, 包括抗体群规模、克隆比例、基因变异概率; 随机产生初始抗体群; 以及设定亲和度成熟条件. 亲和度计算及三个主要的操作 r_1、r_2、r_3 按照第 2.2 节和第 2.3 节的描述进行.

在文献 [17] 中, 我们分析了用于线性系统逼近的人工免疫响应算法是以概率 1 收敛到最优解集. 在本文中, 由克隆选择操作 r_3 的性质可知, 克隆选择后的最优个体至少比克隆增殖操作之前的最优个体好, 因此, 按照文献 [17] 中的算法收敛性分析方法, 同理可以证明人工免疫响应无监督分类与识别算法以概率 1 收敛到最优解集.

3　实验分析

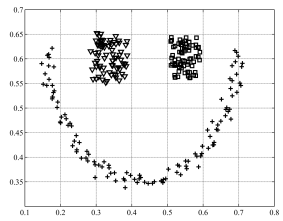
3.1　人工数据集聚类实验

为了能够直观地考察人工免疫响应无监督分类与识别算法 (简称 AIR) 的性能, 我们首先将新算法应用于 6 个人工数据集的聚类问题. 这 6 个数据集分别命名为 Line-blobs, Size5, Spiral, Square4, Sticks 和 Three-circles. 如图 3 (见下页) 所示, 它们具有不同的流形结构, 能够用来考察算法对不同结构数据的聚类性能. 本文将新算法与原始的 K- 均值算法[8](KM), 基于流形距离的进化聚类算法[15](DSEC), 以及 Maulik 等人提出的遗传聚类算法[10](GAC) 进行性能比较. 其中, AIR 的参数设置如下: 亲和度成熟条件为迭代次数 100, 抗体种群规模为 10, 克隆比例为 5, 基因变异概率为 1; DSEC 和 GAC 的参数设置如下: 算法中止条件为迭代次数 100, 种群规模为 50, 交叉概率为 0.8, 变异概率为 0.1; DSKM 与 KM 的最大迭代次数设置为 500, 停止阈值设置为 10^{-10}. 基于上述 6 个人工数据集的参数敏感性实验表明, 收缩因子 ρ 在 $(1,\mathrm{e}^{18}]$ 范围内取值时, 对 AIR 性能的影响很不明显, 在以下实验中令 $\rho=\mathrm{e}^2$.

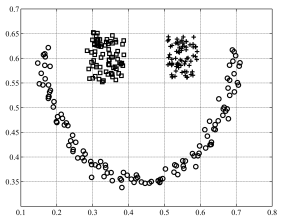
对于算法的聚类性能, 我们采用指标 Adjusted Rand Index[12] 来衡量, 它将类别划分看作是样本之间的一种关系, 每一对样本要么被划分在同一类, 要么在不同类, 通过统计正确决策对数来评价聚类算法的性能. 对于一个有 n 个样本的数据集, Adjusted Rand Index 可以按照以下公式计算

$$R(U,V)=\frac{\sum_{lk}\binom{n_{lk}}{2}-[\sum_{l}\binom{n_{l\cdot}}{2}\cdot\sum_{k}\binom{n_{\cdot k}}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_{l}\binom{n_{l\cdot}}{2}+\sum_{k}\binom{n_{\cdot k}}{2}]-[\sum_{l}\binom{n_{l\cdot}}{2}\cdot\sum_{k}\binom{n_{\cdot k}}{2}]/\binom{n}{2}} \quad (9)$$

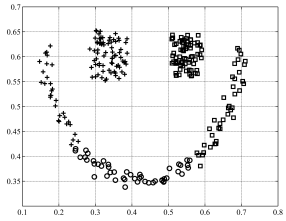
其中, n_{lk} 表示被划分到类属 l 和类属 k 的样本的个数. $R(U,V)\in[0,1]$, 其数值越大, 说明聚类划分的正确率越高.



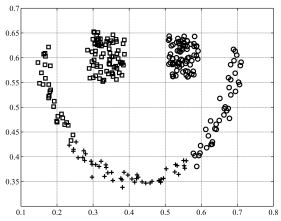
(a) AIR 的聚类结果 (a) Clustering results of AIR



(b) DSEC 的聚类结果 (b) Clustering results of DSEC



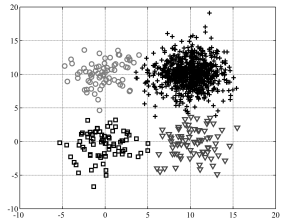
(c) GAC 的聚类结果 (c) Clustering results of GAC



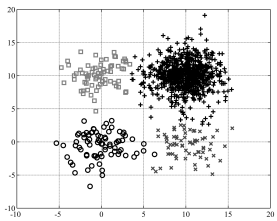
(d) KM 的聚类结果 (d) Clustering results of KM

(I) 四种算法对数据集 Line-blobs 的典型聚类结果

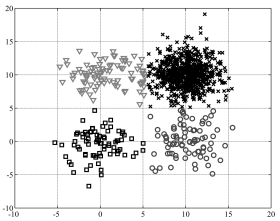
(I) The typical implementation results on Line-blobs obtained from the four algorithms



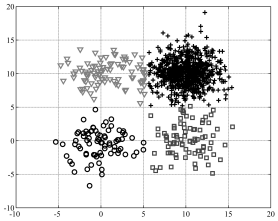
(e) AIR 的聚类结果 (e) Clustering results of AIR



(f) DSEC 的聚类结果 (f) Clustering results of DSEC



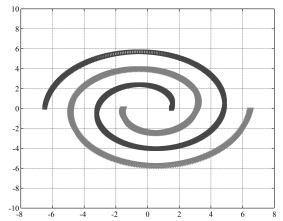
(g) GAC 的聚类结果 (g) Clustering results of GAC



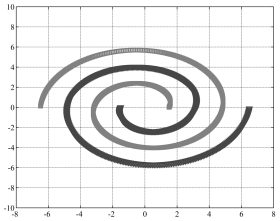
(h) KM 的聚类结果 (h) Clustering results of KM

(II) 四种算法对数据集 Size5 的典型聚类结果

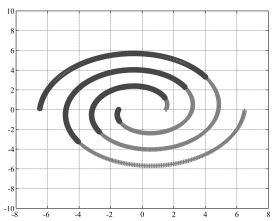
(II) The typical implementation results on Size5 obtained from the four algorithms



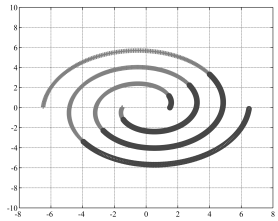
(i) AIR 的聚类结果 (i) Clustering results of AIR



(j) DSEC 的聚类结果 (j) Clustering results of DSEC



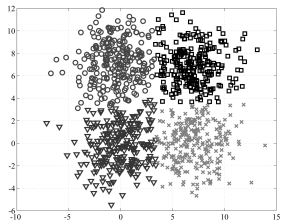
(k) GAC 的聚类结果 (k) Clustering results of GAC



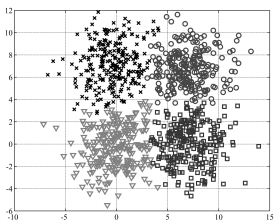
(l) KM 的聚类结果 (l) Clustering results of KM

(III) 四种算法对数据集 Spiral 的典型聚类结果

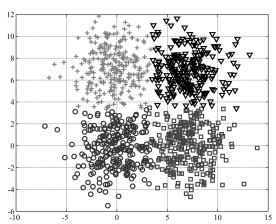
(III) The typical implementation results on Spiral obtained from the four algorithms



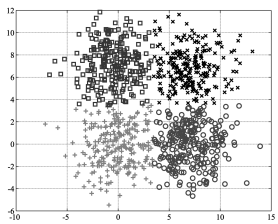
(m) AIR 的聚类结果 (m) Clustering results of AIR



(n) DSEC 的聚类结果 (n) Clustering results of DSEC



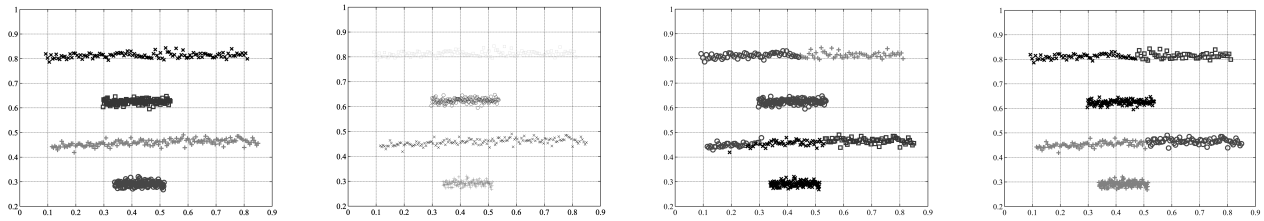
(o) GAC 的聚类结果 (o) Clustering results of GAC



(p) KM 的聚类结果 (p) Clustering results of KM

(IV) 四种算法对数据集 Square4 的典型聚类结果

(IV) The typical implementation results on Square4 obtained from the four algorithms

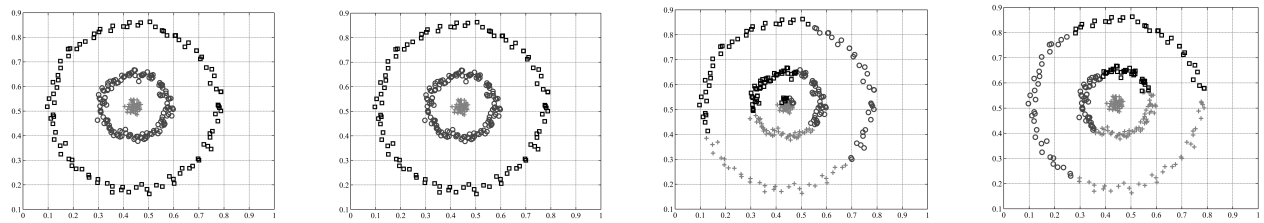


(q) AIR 的聚类结果　(r) DSEC 的聚类结果　(s) GAC 的聚类结果　(t) KM 的聚类结果

(q) Clustering results of AIR　(r) Clustering results of DSEC　(s) Clustering results of GAC　(t) Clustering results of KM

(V) 四种算法对数据集 Sticks 的典型聚类结果

(V) The typical implementation results on Sticks obtained from the four algorithms



(u) AIR 的聚类结果　(v) DSEC 的聚类结果　(w) GAC 的聚类结果　(x) KM 的聚类结果

(u) Clustering results of AIR　(v) Clustering results of DSEC　(w) Clustering results of GAC　(x) Clustering results of KM

(VI) 四种算法对数据集 Three-circles 的典型聚类结果

(VI) The typical implementation results on Three-circles obtained from the four algorithms

图 3　四种算法对人工数据集的典型聚类结果

Fig. 3　The typical implementation results on the artificial data sets obtained from AIR, DSEC, GAC and KM

表 1　四种算法在求解人工数据聚类问题时的性能比较

Table 1　Results of AIR, DSEC, GAC, and KM on artificial datasets

	Adjusted rand index			
数据集	AIR	DSEC	GAC	KM
Line-blobs	1	1	0.399	0.409
Size5	0.922	0.970	0.924	0.920
Spiral	1	1	0.034	0.033
Square4	0.962	0.835	0.937	0.816
Sticks	1	1	0.440	0.504
Three-circles	1	1	0.033	0.044

我们对每一个数据集独立运行 30 次, 各算法在求解以上 6 个问题时得到的 Adjusted Rand Index 的平均值如表 1 所示. 为了更直观地显示四种算法的聚类结果以及 6 个数据集的空间分布情况, 我们在图 3 中展示了四种算法的典型聚类结果.

从表 1 中的统计数据以及图 3 中的典型聚类结果可以明显看出, 对流形结构明显、非球形分布的 Line-blobs、Spiral、Sticks 和 Three-circles 四个问题, 以流形距离作为相似性度量的两种算法 AIR 和 DSEC 能够准确地进行类别划分, 而以欧氏距离作为相似性度量的 GAC 和 KM 对这四个数据的聚类效果非常差. 对流形结构不明显、呈球状分布的 Size5 和 Square4 两个问题, 四种算法均没有获得完全准确的类别划分, 但是 AIR 的聚类正确率要高于其他三种算法. 这充分说明了新的基于流形距离的相似性度量对复杂结构的数据聚类问题是非常有效的. 而基于人工免疫响应模型的 AIR 的全局搜索能力要好于基于遗传算法的 DSEC 和 GAC 算法以及基于梯度下降的 K- 均值算法.

3.2　手写体数字识别实验

本节我们选择了 USPS 数据集作为测试数据, 将新算法应用于手写体数字识别中. USPS 数据集是由 9 298 个 16×16 维灰度图像构成, 其中包含 7 291 个训练样本, 2 007 个测试样本. 由于 USPS 手写体数字集取自美国 Buffalo 日常信件的信封, 所以相对于另一个手写体数据集 NIST 而言识别起来更加困难. 实验取全部测试样本作为聚类数据集, 从中挑选三组较难识别的 $\{0,8\}$、$\{3,5,8\}$、$\{3,8,9\}$ 和两组相对容易识别的 $\{1,2,3,4\}$、$\{0,2,4,8\}$ 共五组数字集合进行识别, 实验中各参数设置与第 3.1 节相同, 对每个数字集合独立运行 30 次, 取 Adjusted

Rand Index 的平均值进行比较, 四种算法的聚类结果如表 2 所示.

表 2 四种算法求解 USPS 手写体数字识别问题的结果

Table 2 Results of AIR, DSEC, GAC, and KM on the USPS handwritten digit datasets

	Adjusted rand index			
数据集	AIR	DSEC	GAC	KM
$\{0,8\}$	0.912	0.907	0.821	0.6951
$\{3,5,8\}$	0.826	0.776	0.569	0.414
$\{3,8,9\}$	0.871	0.669	0.693	0.514
$\{1,2,3,4\}$	0.957	0.875	0.736	0.646
$\{0,2,4,8\}$	0.964	0.912	0.783	0.673

从表 2 中可以明显看出, 无论对三组较难识别的 $\{0,8\}$、$\{3,5,8\}$、$\{3,8,9\}$ 数据集, 还是对两组相对容易识别的 $\{1,2,3,4\}$、$\{0,2,4,8\}$ 数据集, AIR 均获得了最好的结果, DSEC 次之, KM 最差. AIR 在五个数字集合上的平均识别率最差仍达到了 0.826, 最高达到了 0.964, 因此, AIR 在实际应用问题中同样具有良好的性能.

3.3 鲁棒性分析

为了考察四种算法的鲁棒性, 我们采用文献 [19] 中的鲁棒性分析方法对四种算法在求解以上 11 个问题时的鲁棒性进行比较. 具体地, 算法 m 在某一特定数据集上的相对性能用该算法获得的 Adjusted Rand Index 的值 R_m 与所有算法在求解该问题时得到的最大的 Adjusted Rand Index 值的比值来衡量, 即

$$b_m = \frac{R_m}{\max\limits_k R_k} \tag{10}$$

因此, 在某个数据集上表现最好的算法 m^* 的相对性能 $b_{m^*} = 1$, 而其他算法的相对性能 $b_m \leq 1$. b_m 值越大, 则算法 m 在所有算法中的相对性能越好. 因此, 算法 m 在所有数据集上的 b_m 值的总和可以用来客观评价算法的鲁棒性, 总和越大鲁棒性越好. 图 4 为四种算法的鲁棒性比较结果, 每个算法对应的柱状图顶部所标数值为对应算法在所有 11 个问题上的 b_m 值的总和.

从图 4 中可以看出, AIR 获得了最高的总和值, 达到了 11. DSEC 也获得了比较满意的值, 达到了 10.4083. 而采用欧氏距离作为相似性度量的 GAC 和 KM 总和值却明显小于采用流形距离作为相似性度量的 AIR 和 DSEC. 这充分说明了新的基于流形距离的相似性度量对无监督分类和识别问题具有很好的鲁棒性. 实际上, AIR 的 b_m 值对测试的 11 个问题均为 1. AIR 对不同结构的数据聚类问题以及手写体识别问题均表现出了很好的性能, 因此, AIR 在所有比较的四种算法中具有最好的鲁棒性.

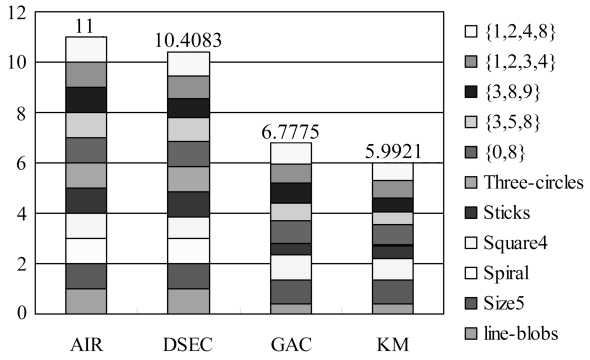


图 4 四种算法的鲁棒性比较

Fig. 4 Comparison of robustness of the four algorithms

4 结论

本文提出了一种基于人工免疫响应模型的聚类算法, 设计了一种新的针对聚类问题的个体编码方法, 从而将聚类问题建模为一个组合优化问题, 减小了问题的搜索空间, 并成功设计了一种能体现样本全局一致性的相似性度量. 在人工数据集聚类问题和手写体数字识别问题上的仿真实验表明, 新算法具有较好的聚类和识别性能.

然而, 这种基于流形距离的相似性度量是在聚类性能和计算复杂度之间的一个折中. 由于要用图论中的最短路径来计算流形距离, 因此其计算复杂度要明显高于欧氏距离的计算复杂度, 从而导致了 AIR 和 DSEC 的计算复杂度要高于 GAC 和 KM. 降低这种流形距离计算复杂度的一个有效途径就是采用一种线性复杂度的求解最短路径的快速算法或者近似算法, 这将是我们下一步的重要工作.

References

1 de Castro L N, Timmis J. *Artificial Immune Systems: A New Computational Intelligence Approach*. Berlin: Springer-Verlag, 2002

2 Jiao Li-Cheng, Du Hai-Feng, Liu Fang, Gong Mao-Guo. *Immunological Computation for Optimization, Learning and Recognition*. Beijing: Science Press, 2006
(焦李成, 杜海峰, 刘芳, 公茂果. 免疫优化计算、学习与识别. 北京: 科学出版社, 2006)

3 de Castro L N, Von Zuben F J. Learning and optimization using the clonal selection principle. *IEEE Transactions on Evolutionary Computation*, 2002, **6**(3): 239−251

4 Dasgupta D. *Artificial Immune Systems and Their Applications*. Berlin: Springer-Verlag, 1999

5 Jiao L C, Wang L. A novel genetic algorithm based on immunity. *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans*, 2000, **30**(5): 552−561

6 Gong M G, Jiao L C, Liu F, Du H F. The quaternion model of artificial immune response. In: Proceedings of the 4th International Conference on Artificial Immune Systems. Banff, Alberta, Canada: 2005. 207−219

7 Farmer J D, Packard N H, Perelson A S. The immune system, adaptation, and machine learning. *Physica D*, 1986, **2**(1-3): 187−204

8 MacQueen J B. Some methods for classification and analysis of multivariate observations. In: Proceedings of the 5th Symposium on Mathematical Statistics and Probability. Berkeley: University of California Press, 1967. 281−297

9 Hall L O, Ozyurt I B, Bezdek J C. Clustering with a genetically optimized approach. *IEEE Transactions on Evolutionary Computation*, 1999, **3**(2): 103−112

10 Maulik U, Bandyopadhyay S. Genetic algorithm-based clustering technique. *Pattern Recognition*, 2000, **33**(9): 1455−1465

11 Pan H, Zhu J, Han D. Genetic algorithms applied to multiclass clustering for gene expression data. *Genomics, Proteomics and Bioinformatics*, 2003, **1**(4): 279−287

12 Handl J, Knowles J. An evolutionary approach to multiobjective clustering. *IEEE Transactions on Evolutionary Computation*, 2007, **11**(1): 56−76

13 Su M C, Chou C H. A modified version of the K-means algorithm with a distance based on cluster symmetry. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2001, **23**(6): 674−680

14 Charalampidis D. A modified K-means algorithm for circular invariant clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005, **27**(12): 1856−1865

15 Gong M G, Jiao L C, Wang L, Bo L F. Density-sensitive evolutionary clustering. In: Proceedings of the 11th Pacific-Asia Conference on Knowledge Discovery and Data Mining. Nanjing, China: 2007. 507−514

16 Zhou D, Bousquet O, La T N, Weston J, Scholkopf B. Learning with local and global consistency. *Advances in Neural Information Processing Systems 16*. Cambridge, USA: MIT Press, 2004. 321−328

17 Gong Mao-Guo, Du Hai-Feng, Jiao Li-Cheng. Optimal approximation of linear systems by artificial immune response. *Science in China Series F: Information Sciences*, 2006, **49**(1): 63−79

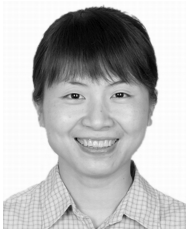
18 Gong Mao-Guo, Jiao Li-Cheng, Du Hai-Feng, Ma Wen-Ping. A novel evolutionary strategy based on artificial immune response for constrained optimizations. *Chinese Journal of Computers*, 2007, **30**(1): 37−47
(公茂果, 焦李成, 杜海峰, 马文萍. 用于约束优化的人工免疫响应进化策略. 计算机学报, 2007, **30**(1): 37−47)

19 Geng X, Zhan D C, Zhou Z H. Supervised nonlinear dimensionality reduction for visualization and classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 2005, **35**(6): 1098−1107

公茂果 西安电子科技大学讲师. 主要研究方向为人工免疫系统、进化计算、数据挖掘、网络安全及工程优化. 本文通信作者. E-mail: gong@ieee.org
(**GONG Mao-Guo** Lecturer at Innovative Research Team of the Ministry of Education of China, Xidan University. His research interest covers computational intelligence and hybrid intelligent systems, artificial immune systems, evolutionary computation, data mining, and optimization. Corresponding author of this paper.)

焦李成 分别于 1982 年、1984 年和 1990 年在上海交通大学、西安交通大学获学士、硕士、博士学位. 现任西安电子科技大学特聘教授. 主要研究方向为自然计算、信号和图像处理、智能信息处理. E-mail: lchjiao@mail.xidian.edu.cn
(**JIAO Li-Cheng** Professor at Innovative Research Team of the Ministry of Education of China, Xidan University. He received his bachelor degree from Shanghai Jiao Tong University, in 1982, master and Ph. D. degrees from Xi′an Jiaotong University, in 1984 and 1990, respectively. His research interest covers natural computation, signal and image processing, and intelligent information processing.)

马文萍 西安电子科技大学博士研究生, 主要研究方向为人工免疫系统、数据挖掘和图像处理.
Email: wpma@mail.xidian.edu.cn
(**MA Wen-Ping** Ph. D. candidate at Institute of Intelligent Information Processing, Xidian University. Her research interest covers artificial immune systems, data mining, and image processing.)



张向荣 西安电子科技大学讲师. 主要研究方向为模式识别、机器学习、图像处理方面及自然计算.
Email: xrzhang@mail.xidian.edu.cn
(**ZHANG Xiang-Rong** Lecturer at Institute of Intelligent Information Processing, Xidian University. Her research interest covers pattern recognition, machine learning, image processing, and natural computation.)