ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.19, No.7, July 2008, pp.1654−1665
DOI: 10.3724/SP.J.1001.2008.01654
© 2008 by *Journal of Software*. All rights reserved.
E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

面向在线图符识别的免疫克隆选择算法*

张莉莎[1,2], 孙正兴[1,2+]

[1](南京大学 计算机软件新技术国家重点实验室,江苏 南京 210093)

[2](南京大学 计算机科学与技术系,江苏 南京 210093)

An Immune Clonal Selection Algorithm for Online Symbol Recognition

ZHANG Li-Sha[1,2], SUN Zheng-Xing[1,2+]

[1](State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China)

[2](Department of Computer Science and Technology, Nanjing University, Nanjing 210093, China)

+ Corresponding author: E-mail: szx@nju.edu.cn, http://cs.nju.edu.cn/szx/

Zhang LS, Sun ZX. An immune clonal selection algorithm for online symbol recognition. *Journal of Software*, 2008,19(7):1654−1665. http://www.jos.org.cn/1000-9825/19/1654.htm

Abstract: In order to collect sufficiently many samples and keep their distinguishability in online sketchy symbol recognition, this paper proposes a detector-generation based clonal selection algorithm and an evaluation method. The algorithm generates detectors with an *r*-contiguous-bits unchanged rule (*r*-CBUR) and a *p*-receptor editing to search in a wide feature space and try to avoid local convergences. Hand-Written Chinese characters are selected as experimental samples, for which the influence of the training parameters is analyzed. The experimental results show the improvements of the training process and the classification results of sketchy symbol recognition.

Key words: artificial immune model; detector; online sketchy symbol; sample training; contiguous-bits unchanged rule; receptor editing; hyper-mutation; local convergence; clonal selection algorithm

摘 要: 针对在线手绘图符识别中样本训练存在的“收集足够数量模板或样本并保持其区分度”的难点,提出了一种适合于手绘图符识别的基于检测器生成的克隆选择算法及其评价方法.该算法采用 *r*-连续位不变规则和 *p*-受体编辑生成初始检测器,使算法具有更广泛的搜索空间并不致陷入局部收敛.以手写文字作为实验对象,评价该算法各参数对个体训练的影响,实验结果验证了该算法对手绘图符样本训练及分类的改进.

关键词: 人工免疫模型;检测器;在线手绘图符;样本训练;连续位不变规则;受体编辑;高频变异;局部收敛;克隆选择算法

中图法分类号: TP391 **文献标识码**: A

作为笔式用户界面(calligraphic user interface)[1,2]的核心技术,在线图符识别技术(online symbol recognition)已得到了卓有成效的研究[3],出现了诸如特征匹配[4]、图匹配[5]和可变模板匹配[6]等模板识别及诸

* Supported by the National Natural Science Foundation of China under Grant Nos.69903006, 60373065, 60721002 (国家自然科学基金); the National High-Tech Research and Development Plan of China under Grant No.2007AA01Z334 (国家高技术研究发展计划(863)); the Program for New Century Excellent Talents in University of China under Grant No.NCET-04-0460 (新世纪优秀人才资助计划)

Received 2006-10-09; Accepted 2007-01-04

© 中国科学院软件研究所 http://www.jos.org.cn

如支撑向量机(support vector machine,简称SVM)[7]、隐马尔可夫模型(hidden Markov model,简称HMM)[8]和贝叶斯网络(Bayesian network)[9]等机器学习识别两大类方法.但是,这些方法的有效性对预定义图符模板集或图符训练样本集具有较强的依赖性,而图符对象类型众多,且输入图符具有笔画随意性和信息模糊性[10].因此,在线手绘图符识别面临着"收集足够数量模板或样本并保持其区分度"的难题.

人工免疫系统(artificial immune system)在网络安全、模式识别、组合优化等多个领域已经取得了大量研究成果,尤以网络安全领域的研究成果最为突出[11−13].网络安全检测中也包含大量模式识别问题,但它们一般都先用否定选择算法进行自体耐受,再对成熟的检测细胞和记忆细胞使用克隆选择算法来生成能够识别变种抗原的抗体细胞[12,14−16].然而,作为对图符形状及用户手绘风格模式进行识别的在线图符识别问题面对众多的图符类别,可能导致自体耐受过程变得复杂和计算量庞大,其在随机生成免疫细胞群体和克隆变异过程中也可能陷入局部最优解,从而失去学习某些抗原结构的机会.

本文拟将免疫系统自动生成与抗体相近的细胞的能力应用到在线手绘图符识别中,采用将初始检测器生成方法与克隆选择算法相结合的方法,令其自主生成与绘图者所提供的初始图符样本相似的训练样本,实现从少量训练样本中自动获得适于在线图符识别的样本群体.实际上,在模式识别其他领域中,尤其是信息模糊性较大的识别也存在类似于手绘图符识别所面临的问题,因此,本文研究的算法在这些领域中同样适用.

1 相关工作

国内外学者对在线手绘图符识别技术进行了卓有成效的研究[3,17,18].已有手绘图符识别主要采用基于模板匹配的识别方法和机器学习方法,如广泛应用的动态时间牵引算法DTW(dynamic time warping)[19]——一种将先验形状知识和结构模板相结合的ASSM(active shape structural model)可变形状模板统计认证方法[6].机器学习方法以隐马尔可夫模型HMM应用最为广泛[8,20,21],其次为支撑向量机SVM [7]和贝叶斯网络[9].从方法的原理和实际应用效果来看,SVM是为区分两类问题而设计的,扩展到多类问题的能力十分有限;HMM方法基于状态转换概率的思想,较适合于描述时序性强的问题;在ASSM可变形状模板中,特征值描述模型的能力很强,弹性调整算法的识别率较高,但在手绘草图识别实验中等错误率(equal error rate,简称EER)值仍比较高[6].再者,由于图符对象涉及范围大,笔画构成具有更大的自由度,尤其是图形的语义更具有不确定性.同时,这些方法一般都要求相当数量的训练样本以保证样本训练后生成的手绘图符模板具有较高的区分度,而要求用户事先提供大量手绘样本会给用户带来较大的不便.因此,这些方法并不能很好地解决这一领域存在的"需要足够数量模板或样本并保持其区分度"的难题,即:(1) 识别所需收集样本数量较大;(2) 图符类别众多、模板区分度小.反观另一类新兴的进化模型——人工免疫模型(artificial immune model,简称AIM),以其良好的多样性、耐受性、免疫记忆、自学习、自适应等特点,已成为继神经网络、进化算法、遗传算法之后的又一大研究热点,克隆选择算法也成为研究最多的人工免疫算法之一.利用克隆选择的特性,一方面可以对初始样本进行训练,生成与已有样本相似、新的更完备的图符模板库.另一方面能够对外来抗原进行有针对性的克隆繁殖,生成的模板或样本具有更好的类别区分度,从而为较好地适应并解决手绘图符识别中"收集足够数量的样本并保持模板区分度"的难点提供了基础.

自1958年克隆选择模型面世至今,免疫算法在理论上日趋完善[22,23],在网络安全、模式识别、组合优化、控制等领域,免疫理论都有一定的应用实例,尤以网络安全领域的研究成果最为丰硕[11].国内相关研究较晚,主要集中在大规模网络入侵检测和控制等技术[12,13].免疫系统对病毒的识别和学习机制也适用于模式识别领域[24−26].Carter[27]提出了一种基于免疫和监督学习的模式识别和分类系统Immunos_81;Sathyanath[28]也提出了一种基于免疫的图像识别方法,通过否定选择算法和克隆选择算法进化专家库,用来识别不同家具的颜色和木料.此外,孙飞显等人[29]提出了基于人工免疫原理的中文姓名识别方法,实验表明,该方法比传统的统计和决策树识别具有更好的准确率和速度.但总的来说,在模式识别领域的研究仍相当有限.

人工免疫理论在网络安全中的应用包括利用免疫检测细胞进行模式识别,如对病毒的识别[13,15]、对非法访问的识别[16,30]、风险检测中网络风险评估[14,31]、网络入侵动态取证识别[12]等.手绘图符识别作为对图符形状和

© 中国科学院软件研究所 http://www.jos.org.cn

用户手绘风格进行识别的一项技术,在模型和算法上都与网络安全检测存在差异.在免疫模型定义上,网络安全与免疫系统所遇到的问题十分相似,都是针对“自体”和“非自体”的分类问题,而手绘图符识别面向的图符类别繁多,针对每一个手绘样本的“自体”(属于该类的样本)和“非自体”(不属于该类的样本)的定义都是不同的.在所用免疫算法上,由于网络安全检测对安全性要求很高,一般先进行自体耐受,如:Forrest等人在病毒检测中使用否定选择算法检测非法字符串[22],再对成熟的检测细胞和记忆细胞使用克隆选择算法生成能够识别变种抗原的抗体细胞[12,14,15,30].而手绘图符识别的目的主要是根据形状信息识别提呈样本所属的图符类别,在考虑用户适应性时还会根据用户手绘习惯和风格的信息辅助判别,除非是手写签名认证等涉及身份验证的交叉应用,一般并不涉及安全性问题,因此自体耐受并非必须的;再者,众多的图符类别将导致自体耐受过程变得复杂和计算量庞大.因此,标准的人工免疫算法和现存的一些针对特定领域设计的人工免疫算法并不直接适用于手绘图符识别,需要对算法本身进行改进.

据此,本文提出了一种基于检测器生成的克隆选择算法和相应的评价方法,利用人工免疫系统能够自动生成与已有样本细胞相似细胞的特性来生成相似样本,解决样本采集数量大的难点,利用人工免疫系统能对外来抗原进行完备识别并进行针对性克隆繁殖的特性来解决图符模板类别众多且区分度小的问题.

本文第 2 节介绍面向在线图符识别的克隆选择算法的设计和具体流程.第 3 节详细描述算法中的检测器生成方法,包括 *r*-连续位不变规则和 *p*-受体编辑的思想和步骤.第 4 节提出改进算法的评价参数和方法.第 5 节是实验,分析各评价参数在手写文字样本训练中的作用,评价算法的训练效果.

2 面向在线图符识别的克隆选择算法设计

标准克隆选择算法中的初始免疫细胞群体一般是随机生成的[13],与外部抗原结构极不相似.理论上,要经过足够的克隆、变异和亲和力测试迭代过程,才能获得识别抗原的抗体群体.这种随机初始群体带来的迭代时间问题可以通过导向性初始群体生成来优化,如采用经过自体耐受的成熟检测器或已有记忆细胞[12,14,15,30]作为初始训练细胞.但是,前者不适合众多类别的图符识别,后者的特征分布不如随机生成细胞广泛,而变异机制可能使算法陷入局部收敛,从而失去学习某些抗原结构的机会.因此,本文将初始检测器生成方法结合到克隆选择算法中,使初始细胞群体既接近抗原群体,又具备广泛的搜索空间.

改进的克隆选择算法过程如图 1 所示.

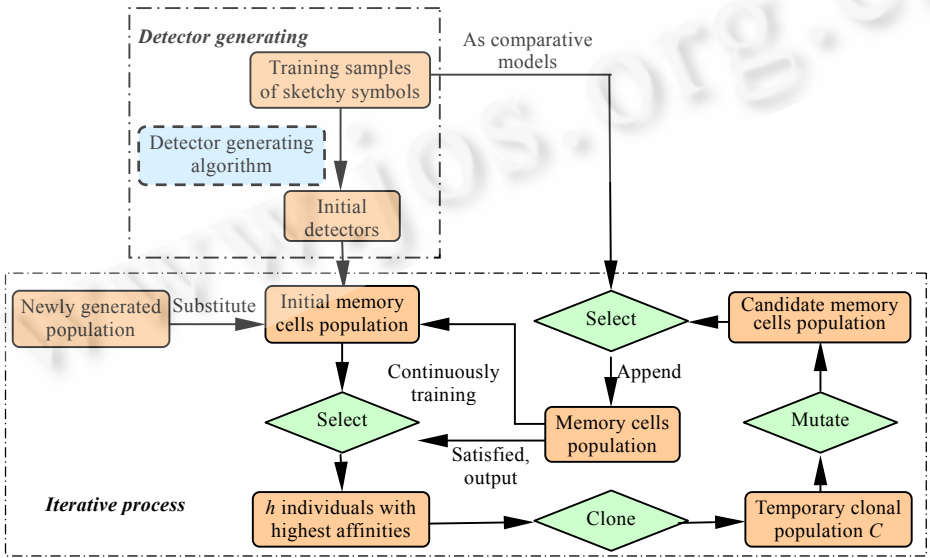

Fig.1 Flow chart of detector generation-based clonal selection algorithm

图 1 基于检测器生成的克隆选择算法流程

© 中国科学院软件研究所 http://www.jos.org.cn

在对N_{class}个类别各N_{sample}个训练样本$TS^{(i,j)}$采用检测器生成算法后,可得到$N_{class}\times N_{sample}$个初始检测器$D_{ini}^{(i,j)}$,并以此作为初始记忆细胞群体$R_{ini}^{(i,j)}$的一部分,通过选择、克隆、变异的迭代,生成更多具有用户图符特征的抗体集合,并将符合亲和力要求、能够匹配图符样本的抗体作为成熟记忆细胞,加入成熟记忆细胞集合$R_{mut}^{(i,j)}$.图中的检测器生成部分是本文改进的重点.然而,初始检测器$D_{ini}^{(i,j)}$是通过自主、可控的方式生成的,新个体的结构取决于训练样本和预设的编辑范围.初始检测器的个体数量有限,不足以检测出大范围特征值所表示的图符特征,而且也包含一些偏离训练样本较大的个体.为了获得更多的有效记忆细胞,需利用免疫细胞高频变异的特性来实现.因此,在初始检测器的基础上运用类似克隆选择算法的机制完成记忆细胞的进化.算法流程如下:

(1) 生成初始检测器集合($D_{ini}^{(i,j)}$)(具体方法见第 3 节).

(2) 随机生成一定数量(r个)新的细胞群体(P_r).

(3) 构造初始记忆细胞集合($R_{ini}^{(i,j)}$),它由初始检测器($D_{ini}^{(i,j)}$)和新的细胞群体(P_r)组成,即$R_{ini}^{(i,j)}=D_{ini}^{(i,j)}+P_r$.

(4) 选择h个与抗原具有最高亲和力的个体(h视类别数N_{class}和每个类别训练样本数N_{sample}而定).

(5) 克隆这h个最好的个体,组成一个临时的克隆群体(C).与$TS^{(i,j)}$的亲和力越高,个体在克隆时的规模就越大.

(6) 把克隆群体提交到高频变异.亲和力越大,个体变异的概率和深度越小;反之,变异的概率越大,程度越深.从而产生一个候选记忆细胞群体(R^*).

(7) 对R^*进行再选择,将各检测器与$TS^{(i,j)}$进行亲和力计算,若亲和力大于阈值t,则将个体加入成熟记忆细胞集合$R_{mut}^{(i,j)}$.

(8) 生成 r'个新的个体取代$R_{mut}^{(i,j)}$中 d 个低亲和力的个体,保持群体多样性.

循环执行(4)~(8),直至满足收敛条件为止.

3 检测器生成算法

检测器生成算法采用r-连续位不变规则(r-contiguous-bits unchanged rule,简称r-CBUR),生成具有若干与TS相似的连续笔画段的初始检测器,再对TS作p-受体编辑($1\le p\le N_{stroke}$),生成与TS有p个不相似笔画的初始检测器,从而组成初始检测器集合D_{ini},目的是预先获得与训练样本存在一定差异的个体,使算法可以在更大范围内进行搜索,不致陷入局部极值而失去了生成其他有用模板的机会.为了便于描述算法,先给出笔画和样本的表示及受体编辑、高频变异、交叉及其相关操作的定义:

描述 1:$\forall i\in[1,N_{class}],j\in[1,N_{sample}]$,训练样本:$TS_{i,j}=IS_1\cup\ldots\cup IS_{N_{stroke}^{i,j}}$;其中,$IS_k$为该样本的第 k 个笔画.下面为了描述简略,用TS表示$TS_{i,j}$,且所有与i,j有关的参数变量都将省略i,j.

描述 2:

(1) 补笔画:$\overline{IS_k}=TS-IS_k$;为该样本除去第 k 个笔画的剩余笔画信息.

(2) 对IS_k进行受体编辑得到的笔画信息:$IS_k'=RE(IS_k)$.

(3) 对样本TS的笔画IS_k做受体编辑,得到的新样本表示为TS_k;同理,依次对样本TS的笔画$IS_{k_1},\ldots,IS_{k_p}$做受体编辑得到的新样本表示为$TS_{k_1,\ldots,k_p}$;$TS_k(l)$表示样本TS_k的第l个笔画.

定义 1. 受体编辑(receptor editing)操作,即对 TS 中某笔画 k 进行受体编辑:

$$RE(TS)=RE(IS_k)=\left(\bigcup\nolimits_q RE(F_{k,q})\cup RE(D_k)\right),\forall k\in[1,N_{stroke}],1\le q\le N_{point}^k.$$

$RE(IS_k)$表示对笔画k进行受体编辑,即对所有特征$F_{k,q}$和D_k的值实施特定幅度的变化.

定义 2. 高频变异(hyper-mutation)操作,即对 TS 中某笔画 k 进行高频变异:

$$HM(TS)=HM(IS_k)=\left(\bigcup\nolimits_q HM(F_{k,q})+HM(D_k)\right),\forall k\in[1,N_{stroke}],1\le q\le N_{point}^k.$$

$HM(IS_k)$表示对笔画k进行高频变异,即对所有特征$F_{k,q}$和D_k的值实施一定范围内的随机突变.

© 中国科学院软件研究所 http://www.jos.org.cn

定义 3. 对TS中任意$p(1\le p\le N_{stroke})$个笔画进行受体编辑,得到$\binom{p}{N_{stroke}}$个新个体的集合:

$$TS_{(p)}=\{TS_{k_1,\ldots,k_p}\}=\left\{\bigcup_{t=1}^{p}IS_{k_t}'\cup\overline{\bigcup_{t=1}^{p}IS_{k_t}}\right\},\ \forall t\in[1,p],k_t\in[1,N_{stroke}],\text{且}\forall t_1<t_2,k_{t_1}<k_{t_2},k_0=0.$$

我们称该过程为 p-受体编辑.

定义 4. 两个样本之间的交叉操作为$\forall k_1,k_2,k_1<k_2$, $TS_{k_1}\times TS_{k_2}=(TS_{k_1}-TS_{k_1}(k_2))\cup TS_{k_2}(k_2)$,交叉操作是对称的,即 $TS_{k_1}\times TS_{k_2}=TS_{k_2}\times TS_{k_1}$.

3.1 r-连续位不变规则(r-CBUR)

r-连续位不变规则即在受体编辑过程中,免疫个体保持存在$n(n>0)$个笔画段,每个笔画段具有$r_i(1\le i\le n)$个连续特征笔画不变(相似).手绘图符个体表示为

$$IND=\left(\{0\}_{t_1-1}\bigcup_{c=1}^{n}(\{0\}_{r_c}\{0\}_{t_{c+1}-t_c-r_c})\right),\ t_{n+1}=N_{stroke}+1,$$

其中,个体IND包含了$n(n>0)$个相似的连续笔画段,每一连续笔画段分别从位置t_c开始,且包含r_c个笔画.其中,1 表示该笔画与TS中对应笔画相似,0 则为不相似; $\{1\}_{r_c}$ 表示该个体从第t_c个笔画至第t_c+r_c-1 个笔画(连续r_c笔)与TS对应的连续笔画段相似.手绘图符模板集表示如下:

$$M_{r_1,\ldots,r_n}=\{M_{t_1,\ldots,t_n}\},\forall k\in[1,n],t_{k+1}-t_k>r_k,t_{n+1}=N_{stroke}+1.$$

模板集是所有包含至少n个连续笔画的模板,且模板前n个连续笔画段的长度分别是$r_1,\ldots,r_n$,模板表示为

$$M_{t_1,\ldots,t_n}=\left(\bigcup_{c=1}^{n}(\{0\}_{t_c-t_{c-1}-v_{c-1}}\{1\}_{v_c})\{*\}_{N_{stroke}-t_n-v_n+1}\right),\ \forall c\in[1,n],v_c\ge r_c,$$

其中,*取 0 或 1.一个模板代表了一组个体,它们从左到右依次包含至少n个连续笔画段,且在前n段中,每一段从位置t_c开始,包含v_c个连续笔画,即包含r_c(或以上)个笔画,在n个笔画段之后的取值是随意的.

为了简化算法描述,本文给出如下定义:

定义 5.

(1) 标志参数 $Tag_{k_1,\ldots,k_p}=c(0\le c\le n)$,表示个体 $TS_{k_1,\ldots,k_p}$ 从第k_1到第k_p个笔画已经包含了c个连续笔画段,长度分别为$r_1,\ldots,r_c$.如果 $Tag_{k_1,\ldots,k_{p-1}}=n$,说明 $TS_{k_1,\ldots,k_{p-1}}$ 已达到包含n个连续相似笔画段的要求.

(2) 令 $R=N_{stroke}-\sum_{q=c+1}^{n}r_q-(n-c)$,该参数用于判定当前个体中是否还存在可进行受体编辑的笔画.

(3) 始末笔画位置的计算:

$A=k_{p-1}+1$, $B=k_{p-1}+r_{c+1}+1$, $C=N_{stroke}-\sum_{q=c+1}^{n}(r_q+1)+1$, $D=\min\left\{N_{stroke},(N_{stroke}-\sum_{q=c+2}^{n}(r_q+1)+1)\right\}$.

对于给定的一组$k_1,\ldots,k_p$,计算 $TS_{k_1,\ldots,k_p}=TS_{k_1,\ldots,k_{p-1}}\times TS_{k_p}$ 的算法如下:

Step 0.预置$Tag_0=n$,表示TS是符合连续笔画段数目要求的样本.

Step 1.$\forall k\in[1,r_1]$,TS_k的前k个笔画未形成连续的r_1个相似笔画,令$Tag_k=0$;反之,$\forall k\in[r_1+1,N_{stroke}]$,$TS_k$的前$k$个笔画已包含至少$r_1$个连续的相似笔画,令$Tag_k=1$;

$\forall k$,$Tag_k=n$,将TS_k归入模板M_1.

循环执行 Step 2~Step 3:

Step 2.

(1) 若 $Tag_{k_1,\ldots,k_{p-1}}=n$,则$k_p$的取值范围是$[k_{p-1}+1,N_{stroke}]$,令 $Tag_{k_1,\ldots,k_p}=n$,并将新个体 $TS_{k_1,\ldots,k_p}$ 归入 $TS_{k_1,\ldots,k_{p-1}}$ 所属的模板中.

(2) 若 $Tag_{k_1,\ldots,k_{p-1}}=c(0\le c\le n-1)$,分两种情况讨论:

(2.1) 若$k_{p-1}\le R$,

© 中国科学院软件研究所 http://www.jos.org.cn

如果 $k_{p-1}+r_{c+1}>N_{stroke}-\sum_{q=c+1}^{n}(r_q+1)$,那么:

(2.1.1) k_p的取值范围是$[B,D]$, $Tag_{k_1,\ldots,k_p}=c+1$;

(2.1.2) k_p的取值范围是$[A,C]$, $Tag_{k_1,\ldots,k_p}=c$;

否则,k_p的取值范围是$[A,D]$:若$k_p>k_{p-1}+r_{c+1}$, $Tag_{k_1,\ldots,k_p}=c+1$;否则, $Tag_{k_1,\ldots,k_p}=c$.

(2.2) 若$k_{p-1}>R$,则说明k_{p-1}之后没有可编辑的位置,应舍弃 $TS_{k_1,\ldots,k_{p-1}}$.

Step 3.若 $Tag_{k_1,\ldots,k_p}=n$,

$\forall q\in[1,n], m_q\in[1,\ldots,p], \exists k_{m_1},\ldots,k_{m_n}$,且为最小的一组值, s.t. $\forall k_{m_q}$, $k_{m_q}-k_{m_q-1}>r_q$.

令 $t_q=k_{m_q-1}+1(q\in[1,n])$,将 $TS_{k_1,\ldots,k_p}$ 归属于模板 $M_{t_1,\ldots,t_n}$.

通过上述算法可以得到与TS至少有n个相似的连续笔画段的个体细胞,作为手绘图符模板生成过程中的一部分检测器群体.由算法描述可知,相对于某个手绘图符个体(该文字个体笔画数为n_s)所生成的个体数目N计算如下:

(1) n=1 时,个体中至少存在一段包含 r 个(或以上)连续笔画的情况:

$$\begin{cases} N_1(n_s,r)=2^{n_s-r}+\sum_{k=1}^{r}N_1(n_s-k,r), & n_s\ge r \\ N_1(n_s,r)=0, & n_s<r \end{cases}.$$

(2) n≥2 时,个体中至少存在n个连续笔画段,每个笔画段分别包含$r_c(1\le c\le n)$(或以上)个连续笔画的情况:

$$\begin{cases} N_n(n_s,r_1,\ldots,r_n)=\sum_{k=1}^{r_1}N_n(n_s-k,r_1,\ldots,r_n)+\sum_{k=r_1+1}^{n_s-\sum_{l=2}^{n}r_l}N_{n-1}(n_s-k,r_2,\ldots,r_n)+\ldots+\sum_{k=\sum_{l=1}^{n-1}r_l+1}^{n_s-r_n}N_1(n_s-k,r_n), & n_s\ge\sum_{c=1}^{n}(r_c+1)-1 \\ N_n(n_s,r_1,\ldots,r_n)=0, & n_s<\sum_{c=1}^{n}(r_c+1)-1 \end{cases}.$$

模板集大小的计算方法与计算个体数目是相似的:

当 n=1 时,手绘图符模板中至少存在一段包含 r 个(或以上)连续笔画的情况:

$$\begin{cases} MN_1(n_s,r)=1+\sum_{k=1}^{r}MN_1(n_s-k,r), & n_s\ge r \\ MN_1(n_s,r)=0, & n_s<r \end{cases}.$$

当 n≥2 时,计算公式同上.

3.2 p-受体编辑

受体编辑层次如图 2 所示,对于某一个特定的 p,本文利用(p−1)-受体编辑的结果,迭代地进行 p-受体编辑,从而避免了每次对 TS 进行 p-受体编辑而需要重新进行一次$\binom{p}{N_{stroke}}$的笔画组合操作.

根据定义 4,通过迭代的交叉操作可以得到原始样本进行受体编辑后差异程度不同的所有新个体,该类个体可以在任意位置的笔画上接受编辑,而无须满足 r-连续相似笔画段的要求.

具体算法描述如下:

(1) 第 0 步.原始样本 TS;

(2) 第 1 步.依次分别对样本TS中的各笔画进行受体编辑,得到N_{stroke}个新个体:

$$TS_1=IS_1'\cup\overline{IS_1},\ldots,\ TS_k=IS_k'\cup\overline{IS_k},\ldots,\ TS_{N_{stroke}}=IS_{N_{stroke}}'\cup\overline{IS_{N_{stroke}}},$$

其中,每个个体TS_k与原个体TS只在第k个笔画上存在差别;

(3) 第p步:$\forall k_1,\ldots,k_{p-1}\in[1,N_{stroke}]$,且$k_1<k_2<\ldots<k_{p-1}$,取所有$k_p>k_{p-1}$,根据 $TS_{k_1,\ldots,k_p}=TS_{k_1,\ldots,k_{p-1}}\times TS_{k_p}$ 来计算,

© 中国科学院软件研究所 http://www.jos.org.cn

可得($N_{stroke}-k_{p-1}$)个关于该特定($k_1,\ldots,k_{p-1}$)参数组的新个体;遍历第($p-1$)步所得的所有参数组,便可得到$\binom{p}{N_{stroke}}$个新个体,其中每个个体与 TS 在 p 个笔画上存在差别;

(4) 第 n 步:可得所有笔画上都有差异的 1 个新个体 $TS_{k_1,\ldots,k_{N_{stroke}}}$.

(5) 总共可获得 $\sum_{p=1}^{N_{stroke}}\binom{p}{N_{stroke}}=2^n-1$ 个新个体,加上原样本共有 2^n 个模板.

实际上,不需要事先获得所有新个体.经p-受体编辑的$TS_{(p)}$中的个体与原始样本TS的亲和力比($p-1$)-受体编辑$TS_{(p-1)}$的个体要小.在p较大时,$TS_{(p)}$的新个体与TS的亲和力很小,在假设TS能够很好地表示手绘图符特点的前提下,若$TS_{(p)}$中新个体客观上能够潜在地表达手绘图符的另一方面的特征,则这种特征应能在该类别图符其他训练样本中反映出来,即可能存在以下两种情况:(1) 其他训练样本进行受体编辑时会产生这些新个体;(2) 由其他训练样本生成的初始检测器通过高频变异可能产生与这些个体很相近的个体.因此,由某个样本TS生成初始检测器时只需要选择较小的参数p作受体编辑即可.

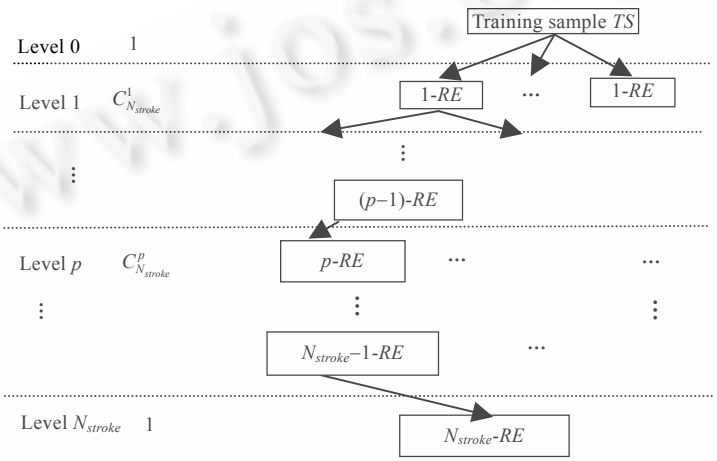

Fig.2 Hierarchical structure of receptor editing on sample TS

图 2 样本 TS 受体编辑的层次图

4 算法评价

4.1 算法训练的参数

受体编辑的幅度和个体变异的范围都会影响算法的训练结果,本文对这些参数的定义如下:

受体编辑幅度(receptor editing range,简称 RER):生成初始检测器时,为了保证检测器不致陷入局部收敛,个体在进行受体编辑时各个特征值须达到一定变化幅度.实验中,该幅度设置为特征数值的可变百分比.

个体变异范围(individual mutation scope,简称 IMS):记忆细胞成熟过程中,为了避免变异个体偏离太大而遭淘汰的情况,需控制个体特征位高频变异的范围.实验中,该范围的最大值设置为特征数值的可变百分比.

本文对手写图符的特征计算总体均值 f_mean 和总体标准差 f_dev,根据它们的值决定投射到二维平面上的个体之间的距离.在评价实验结果时,本文引入了个体映射域和群体映射域的概念.

定义 6. 在个体所投射的二维平面上,特定的距离参数ε下,个体映射域(individual mapping region,简称 IMR)为以该个体为圆心、ε为半径的圆,圆的面积称为个体映射面积(individual mapping area,简称 IMA);群体映射域(population mapping region,简称 PMR)为以该群体中所有个体为中心、ε为半径的圆的并,其面积称为群体映射面积(population mapping area,简称 PMA).

© 中国科学院软件研究所 http://www.jos.org.cn

个体映射域和群体映射域具有以下特点:(1) 理论上,ε越大,个体映射域和群体映射域越大,即 IMA 和 PMA 相应越大;(2) 个体映射域较大时,与该个体在特征上相近且以该个体为代表的个体数目较多,即该个体对应的图符样本代表的特征范围较大,反之亦然;(3) 群体映射域较大时,将有更多的边缘个体被归入该群体,则该群体对应的图符类别包容的特征范围也较大,反之亦然;(4) 当ε大于一定阈值时,个体间映射域会重叠,群体映射域随着ε增大而增大的趋势将逐渐变得不明显.群体映射域的重叠区域意味着图符类别的特征空间之间的交叠,将导致错误的图符分类;反之,ε太小也会导致较多样本无法被识别的情况.

定义 7. 群体训练前后的映射面积变化比例记为$P_{T/I}$,表示训练后的群体映射面积与训练前的群体映射面积的比值.

4.2 算法分类的评价参数

评价样本训练结果的优劣除了直接考察其对样本的分类情况,还可以通过分析多个类别图符的训练群体之间的重叠情况来实现.

定义 8. 群体映射域的重叠比例记为$OL_{\varepsilon}(\{P_i\})(1\le i\le n)$,表示对于特定的$\varepsilon$,$n$个训练群体$\{P_i\}$的映射域重叠部分(两个或以上类别的重叠)的面积与所有群体映射总面积的比值.群体训练前和训练后的映射域重叠比例分别记为OL_I和OL_T.群体映射域的重叠比例越小,分类错误的可能性越小;反之,则越大.

5 实验与评价

实验样本为 5 个用户的手写名字(汉字),作为 5 个不同类别的图符$P_i(i=1,2,3,4,5)$,每个用户以自己习惯的手写方式提供 50 个手写名字样本($N_{class}=5$,$N_{sample}=50$).本文设计了 3 个实验,分别从受体编辑和个体变异参数、单类别图符训练、多类别图符训练考察改进的免疫算法对手写文字样本训练和分类的影响.

5.1 训练参数

本实验的目的在于考察受体编辑幅度和个体变异范围对某个类别名字样本的训练过程的影响.本实验从图符类别P_5中随机选择了 10 个样本作为原始训练样本.

5.1.1 受体编辑幅度(RER)

图 3、图 4 描述了在个体变异范围$IMS=0.05$时,面对 5 种不同的受体编辑幅度RER和不同的距离参数ε,该类别名字的个体训练后群体映射面积的取值及其变化趋势.如图 3 所示,在特定的距离参数ε下,编辑幅度RER越大,训练后的样本群体映射面积PMA越大;随着ε的加大,训练后PMA先增长继而逐渐趋于平缓.如图 4 所示,训练前后面积比例$P_{T/I}$大于 1,说明训练后样本群体映射面积确实比训练前要大,且$P_{T/I}$随着RER的增大而增大;曲线的下降随着ε的增大而趋于平缓,说明ε足够大时对面积比例$P_{T/I}$的影响很小.

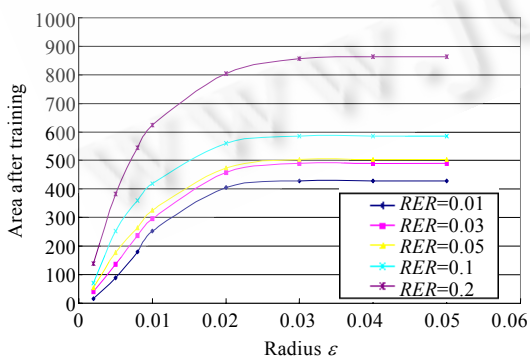

Fig.3 The influence of RER on PMA after training

图 3 RER 对训练后群体映射面积的影响

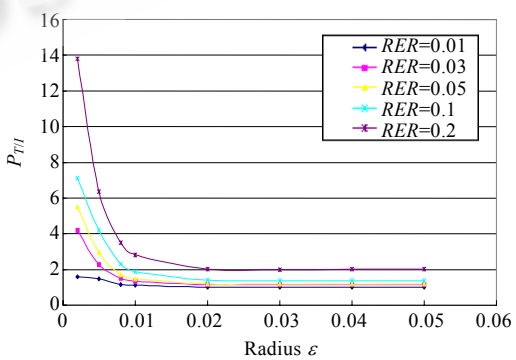

Fig.4 The influence of RER on $P_{T/I}$

图 4 RER 对训练前后群体映射面积比例的影响

© 中国科学院软件研究所 http://www.jos.org.cn

5.1.2 个体变异范围(IMS)

图 5、图 6 描述了在受体编辑幅度*RER*=0.05 时,面对 5 种不同的个体变异范围IMS和不同的距离参数ε,该类别名字的个体训练后群体映射面积的取值及其变化趋势.如图 5 所示,曲线十分接近,说明在特定的距离参数ε下,不同的个体变异范围对样本训练后的群体映射面积影响不大;与参数RER相似,随着ε的增大,训练后群体映射面积先显著增大继而逐渐趋于平缓.如图 6 所示,$P_{T/I}$大于 1,但IMS的不同取值对$P_{T/I}$的影响很小.实验结果表明:由于个体高频变异采用的是偶然性很大的特征位突变的形式,对改变群体映射域的贡献不明显,这也是本文在克隆选择算法的检测器生成中引入受体编辑操作的一个原因.

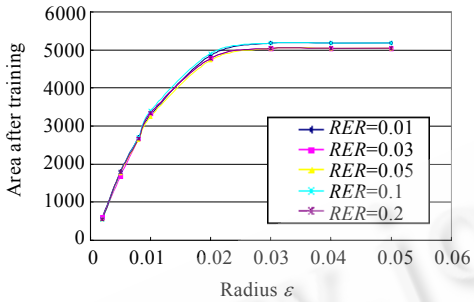

Fig.5 The influence of IMS on PMA after training

图 5 IMS 对训练后的群体映射面积的影响

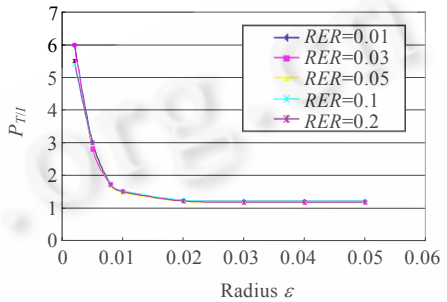

Fig.6 The influence of IMS on $P_{T/I}$

图 6 IMS 对训练前后群体映射面积比例的影响

5.2 单类别图符训练

同一类别的名字是由同一用户提供的习惯一致、具有相似几何特征和书写特征的图符,因而单类别图符的多组样本训练后所得个体对该图符具有相似的描述能力.本实验对名字P_5的 5 组(每组 10 个)原始样本分别进行训练,得到特征空间上训练后的群体映射图,如图 7 所示.图中同一形状标记的点为同一组样本训练后的个体集,密度较大点集的划分是由某一原始样本训练所得的个体集.由图 7 可知:(1) 5 组样本的特征散点比较集中,说明同类样本几何形状和用户手写习惯是一致的;(2) 训练时生成的个体与原始个体在特征空间上相近;(3) 某些点离点集中心及其他个体较远,说明个别样本之间仍可能存在较大差异(由于输入设备等偶然性因素).图 7 中第 5 组样本对应的个体集合相对稀疏,在特征空间上的跨度较大,直观上说,容易与其他类别交叠而造成错误分类.然而,人工免疫方法并非通过超平面或曲面来界定类别的,本文的算法就是使训练后的群体映射域围绕在有代表性的原始个体周围.图 8 是P_5的 5 组样本在*RER*=0.05,*IMS*=0.05 时的$P_{T/I}$.当$\varepsilon\geq$0.005 时,曲线十分接近,且在$\varepsilon\geq$0.01 时曲线趋于平缓,取值都在区间[1,2]浮动,说明稀疏的样本训练后其群体映射域增大的比例并不明显大于密集的样本,即训练算法生成的模板具有一定的稳定性,表现为对个体差异的容忍度及群体映射面积的稳定性,允许同类图符的不同样本之间存在一定程度的差异.

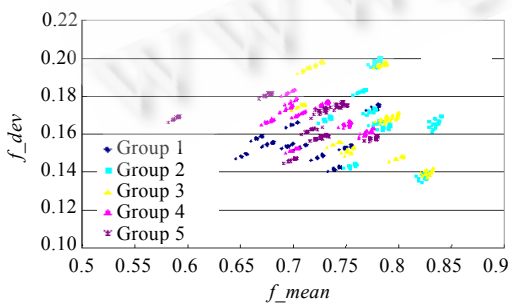

Fig.7 Pop. mapping of the 5-group samples of one-class

图 7 单类别图符的 5 组样本的群体映射图

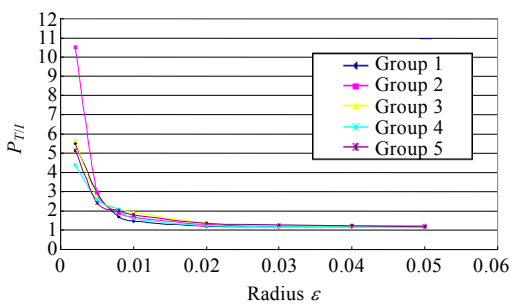

Fig.8 $P_{T/I}$ of the 5-group samples of one-class

图 8 单类别图符的 5 组样本训练前后的映射面积比例

© 中国科学院软件研究所 http://www.jos.org.cn

5.3 多类别图符训练

本实验从 5 类名字中各随机选择 10 个样本进行训练,再计算它们的群体映射域的重叠比例.如图 9 所示,重叠比例$OL(\{P_i\})(1\le i\le 5)$在训练前后变化不大,都随着距离参数ε的增长呈现明显的上升趋势.如图 10 所示,4 组曲线分别代表名字P_5与名字$P_i(i=1,2,3,4)$的重叠情况,实曲线和虚曲线分别表示样本训练前和训练后的重叠比例OL_1和OL_T.

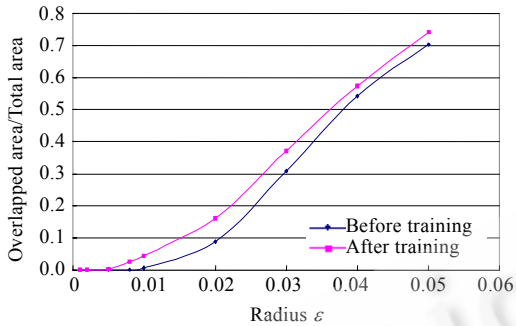

Fig.9 Overlap of the 5-class symbols

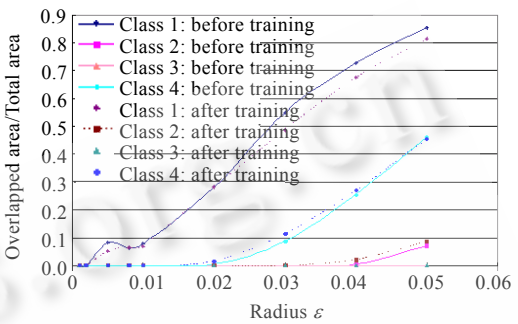

Fig.10 Overlap of the PMR between P_5 and the other 4 classes

图 9 5 类图符训练群体的重叠

图 10 P_5训练群体映射域与另 4 类名字的重叠

由图 10 可知:

(1) P_5与 4 类名字的重叠比例相差很大,事实上,两个群体可能重叠区域的位置主要是由原始训练样本本身的特征决定的.

(2) 每一对曲线中的两条曲线比较接近,说明训练前后OL_1和OL_T相差不太大.这说明:虽然经过训练后所得群体比原始个体集庞大得多,群体映射面积也显著增大(如图 3 所示),为单个类别的名字提供了更广大的特征空间进行识别,但同时并未增加名字分类错误的可能性.

(3) 可能出现$OL_1>OL_T$的情况,如P_5与P_1的重叠曲线(最上面一对曲线),这是因为两类名字中经过受体编辑和高频变异产生的新个体在特征空间上朝着重叠的反方向进化而偏离了两个群体可能重叠的区域,从而降低了重叠面积、增大了总面积,这是我们最期望的一种训练结果.

(4) 当$\varepsilon\le 0.01$时,重叠比例OL_1和OL_T都小于 0.1.若假设个体分布均匀,则两类名字的分类错误率小于 10%.

实际上,个体的分布是不均匀的,特征上具有代表性的个体在训练中将优先被选择、克隆和变异,保留下来的新个体也是比较有代表性的个体.在特征选择合理的情况下,经过训练后OL_T应小于OL_1,实际分类错误率也应小于重叠比例OL_T.实验表明,在一定的训练参数下,该算法有利于降低图符分类错误的可能性.

6 总 结

人工免疫理论中的克隆选择算法是研究最多的免疫算法之一,利用其克隆选择的特性可以较好地适应并解决图符识别中"收集足够数量的样本并保持模板区分度"的难点.本文以汉字手写文字为实验对象,针对在线手绘图符识别中两大难点,提出了一种面向图符识别的基于检测器生成的克隆选择算法,并分析该算法的训练效果.实验结果表明,改进的克隆选择算法可以从少量训练样本获得大规模样本群体,所得模板具有良好的稳定性,并有利于降低图符分类错误的可能性,为人工免疫原理在手绘图符识别以及信息模糊性较大,交互方式自由、随意的其他模式识别领域中的应用提供了很好的参考.

References:

[1] Landay JA, Myers BA. Sketching interfaces: Toward more human interface design. IEEE Computer, 2001,34(3):56−64.

© 中国科学院软件研究所 http://www.jos.org.cn

[2] Li Y, Guan ZW, Dai GZ. Research on development tools for pen-based user interfaces. Journal of Software, 2003,14(3):392−400 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/14/392.htm

[3] Sun ZX, Feng GH, Zhou RH. Techniques for sketch-based user interface: Review and research. Journal of Computer-Aided Design & Computer Graphics, 2005,17(9):1889−1899 (in Chinese with English abstract).

[4] Calhoun C, Stahovich TF, Kurtoglu T, Kara LB. Recognizing multi-stroke symbols. In: Proc. of the AAAI Spring Symp. on Sketch Understanding. Palo Alto, 2002. 15−23. http://www.aaai.org/Press/Reports/Symposia/Spring/ss-02-08.html

[5] Xu XG, Sun ZX, Peng BB, Jin XY, Liu WY. An online composite graphics recognition approach based on matching of spatial relation graphs. Int'l Journal of Document Analysis and Recognition, 2004,7(1):44−55.

[6] Al-Zubi S, Broemme A, Toennies K. Using an active shape structural model for biometric sketch recognition. In: Michaelis B, Krell G, eds. Proc. of the DAGM 2003. LNCS 2781, Beilin, Heidelberg: Springer-Verlag, 2003. 187−195.

[7] Sun ZX, Zhang LS, Tang EY. An incremental learning algorithm based on SVM for online sketchy shape recognition. In: Wang LP, Chen K, Soon OY, eds. Advance in natural Computing. LNCS 3610, Beilin, Heidelberg: Springer-Verlag, 2005. 655−659.

[8] Sun ZX, Jiang W, Sun JY. Adaptive online multi-stroke sketch recognition based on hidden Markov model. In: Daniel YS, Liu ZQ, Wang XZ, *et al*., eds. Advances in Machine Learning and Cybernetics. LNCS 3784, Beilin, Heidelberg: Springer-Verlag, 2005. 948−957.

[9] Sun ZX, Zhang LS, Zhang B. Online composite sketchy shape recognition based on Bayesian networks. In: Jiao LC, Wang LP, Gao XB, *et al*., eds. Proc. of the Natural Computing. LNCS 4222, Beilin, Heidelberg: Springer-Verlag, 2006. 506−515.

[10] Sun ZX, Liu WY, Peng BB, Zhang B, Sun JY. User adaptation for online sketchy shape recognition. In: Jolep L, Kwon YB, eds. Graphics Recognition: Recent Advances and Perspectives. LNCS 3088, Beilin, Heidelberg: Springer-Verlag, 2004. 303−314.

[11] Burnet FM. The Clonal Selection Theory of Acquired Immunity. Cambridge: Cambridge University Press, 1959.

[12] Ding JL, Liu XQ, Li T, Yang S, Yang P. Dynamic computer forensics based on artificial immune system against network intrusion. Journal of Sichuan University (Engineering Science Edition), 2004,36(5):108−111 (in Chinese with English abstract).

[13] Li T. Computer Immunology. Beijing: Publishing House of Electronics Industry, 2004 (in Chinese).

[14] Li T. An immunity based network security risk estimation. Science in China (Series F), 2005,48(5):557−578.

[15] Okamoto T, Ishida Y. A distributed approach against computer viruses inspired by the immune system. IEICE Trans. on Communications, 2000,83(5):908−915.

[16] Kim JW, Bentley P. The human immune system and network intrusion detection. In: Proc. of the 7th European Congress on Intelligent Techniques and Soft Computing (EUFIT'99). Aachen, 1999.

[17] Plamondon R, Srihari NS. On-Line and off-line handwriting recognition: A comprehensive survey. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2000,22(1):63−84.

[18] Liu CL, Jaeger S, Nakagawa M. Online recognition of Chinese characters: The state-of-the-art. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2004,26(2):198−213.

[19] Jain AK, Griess FD, Connell SD. On-Line signature verification. Pattern Recognition, 2002,35(12):2963−2972.

[20] Günter S, Bunk H. Handwritten word recognition using classifier ensembles generated from multiple prototypes. Int'l Journal of Pattern Recognition and Artificial Intelligence, 2004,18(5):957−974.

[21] Fierrez-Aguilar J, Nanni L, Lopez-Penalba J, Ortega-Garcia J, Maltoni D. An on-line signature verification system based on fusion of local and global information. In: Kanade T, Jain A, Ratha NT, eds. Proc. of the Audio- and Video-based Biometric Person Authentication. LNCS 3546, Beilin, Heidelberg: Springer-Verlag, 2005. 523−532.

[22] Forrest S, Perelson AS, Allen L, Cherukuri R. Self-Nonself discrimination in a computer. In: Proc. of the IEEE Symp. on Security and Privacy. Oakland: IEEE Computer Society Press, 1994. 202−212.

[23] Bentley K. Immune memory in the dynamic clonal selection algorithm. In: Proc. of the 1st Int'l Conf. on Artificial Immune System (ICARIS). Canterbury, 2002. 57−65.

[24] de Castro LN, Timmis J. Artificial immune systems: A novel approach to pattern recognition. In: Proc. of the Artificial Neural Networks in Pattern Recognition. Paisley, 2002. 67−84.

© 中国科学院软件研究所 http://www.jos.org.cn

[25] Tarakanov AO, Skormin VA. Pattern recognition by immunocomputing. In: Proc. of the Special Sessions on Artificial Immune Systems in Congress on Evolutionary Computation, IEEE World Congress on Computational Intelligence. Hawaii: IEEE Computer Society Press, 2002. 938−943.

[26] White J, Garrett SM. Improved pattern recognition with artificial clonal selection. In: Goos G, Hartmanis J, van Leeuwen J, eds. Proc. of the Artificial Immune Systems. LNCS 2787, Beilin, Heidelberg: Springer-Verlag, 2003. 181−193.

[27] Carter JH. The immune system as a model for pattern recognition and classification. Journal of the American Medical Informatics Association, 2000,7(3):28−41.

[28] Sathyanath S, Sahin F. An AIS approach to a color image classification problem in a real time industrial application. In: Proc. of the IEEE Int'l Conf. on System, Man and Cybernetics. Arizona: IEEE Computer Society Press, 2001. 2285−2290.

[29] Sun FX, Li T, Jiang YP, Wang TF, Ni JC, Gong X. A novel method of chinese name recognition based on artificial immune system. Journal of Sichuan University (Engineering Science Edition), 2006,38(1):98−102 (in Chinese with English abstract).

[30] Liang KX, Li T, Liu Y, Chen Y. A new model of intrusion detection based on artificial immune theory. Computer Engineering and Applications, 2005,41(2):129−132.

[31] Wang YF, Li T, Hu XQ, Song C. A real-time method of risk evaluation based on artificial immune system for network security. Acta Electronica Sinica, 2005,33(5):945−949 (in Chinese with English abstract).

附中文参考文献:

[2] 栗阳,关志伟,戴国忠.笔式用户界面开发工具研究.软件学报,2003,14(3):392−400. http://www.jos.org.cn/1000-9825/14/392.htm

[3] 孙正兴,冯桂焕,周若鸿.基于手绘草图的人机交互技术研究进展.计算机辅助设计与图形学学报,2005,17(9):1889−1899.

[12] 丁菊玲,刘晓洁,李涛,仰石,杨频.基于人工免疫的网络入侵动态取证.四川大学学报(工程科学版),2004,36(5):108−111.

[13] 李涛.计算机免疫学.北京:电子工业出版社,2004.

[29] 孙飞显,李涛,蒋亚平,王铁方,倪建成,龚勋.基于人工免疫原理的中文姓名识别方法.四川大学学报(工程科学版),2006,38(1):98−102.

[30] 梁可心,李涛,刘勇,陈桓.一种基于人工免疫理论的新型入侵检测模型.计算机工程与应用,2005,41(2):129−132.

[31] 王益丰,李涛,胡晓勤,宋程.一种基于人工免疫的网络安全实时风险检测方法.电子学报,2005,33(5):945−949.

张莉莎(1979－),女,广东普宁人,博士生,主要研究领域为生物计算技术,智能人机交互.

孙正兴(1964－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为多媒体计算,计算机视觉,智能人机交互.

© 中国科学院软件研究所 http://www.jos.org.cn