

文章编号:1001-9081(2009)06-1582-03

## 面向多类别模式分类问题的新型阴性选择算法

刘殊

(广东体育职业技术学院 信息资源中心, 广州 510663)  
(qinwj2007@126.com)

**摘要:** 针对阴性选择算法缺乏高效的分类器生成机制和“过拟合”抑制机制的缺陷, 提出了一种面向多类别模式分类的阴性选择算法 CS-NSA。通过引入克隆选择机制, 根据分类器的分类效果和刺激度对其进行自适应学习; 针对多类别模式分类的“过拟合”问题, 引入了检测器集合的修剪机制, 增强了检测器的分类推广能力。对比实验结果证明: 与著名的人工免疫分类器 AIRS 相比, CS-NSA 体现出更高的正确识别率。

**关键词:** 阴性选择算法; 克隆选择; 模式分类

中图分类号: TP301 文献标志码:A

### Novel negative selection algorithm for multi-class pattern classification problems

LIU Shu

(Information Resource Center, Guangdong Vocational Institute of Sport, Guangzhou Guangdong 510663, China)

**Abstract:** A negative selection algorithm for multi-class pattern classification problems named CS-NSA was proposed. The algorithm used clonal selection mechanism to implement self-adaptive learning of detectors and adopted detector trimming mechanism to tackle the over-fitting problem in multi-class classification. This mechanism enhanced the generalization capability of the detectors. The results of comparative experiments show that the proposed algorithm exhibits higher classifying accuracy than that of AIRS, a famous artificial immune classifier.

**Key words:** negative selection algorithm; clonal selection; pattern classification

## 0 引言

近年来, 人工免疫系统 (Artificial Immune Systems, AIS) 的研究成果逐渐体现出更为理论化的趋势<sup>[1]</sup>, 特别是对于其核心组成部分——免疫算法而言, 传统的算法模型在鲁棒性、自适应学习等方面已很难适应复杂工程领域的实际需求<sup>[2]</sup>。在此情况下, 借鉴生物免疫机理, 根据工程问题的实际要求设计面向具体问题<sup>[3]</sup>的新型免疫算法已成为了研究者们拓展 AIS 应用范围的有效手段。

阴性选择算法 (Negative Selection Algorithm)<sup>[4]</sup> 是一种被广泛研究的免疫算法, 很多 AIS 都是建立在阴性选择算法的框架之上<sup>[1]</sup>。其核心思想是通过模拟生物免疫系统中 T 细胞的成熟机制, 对 AIS 中异常数据检测器进行训练。图 1 描绘了阴性选择算法的基本框架。在训练阶段, 随机生成大量初始检测器, 并逐一与系统定义的自体数据集进行匹配。若检测器能够识别某条自体数据, 则被删除。那些与所有自体数据均不能匹配的检测器成为有效检测器, 并被存入检测器集合。在检测阶段, 未知类别数据与有效检测器进行匹配。如果至少有一个检测器能够识别该数据, 则该数据被归类为非自体; 否则归为自体。

根据前述的研究思路, 近年来研究者们对阴性选择算法的理论模型进行了诸多有效的改进。文献[5]采用蒙特卡罗法设计了一种随机化的阴性选择算法模型。文献[6]提出了一种可变状态检测器阴性选择算法 V-detector。文献[7]深入分析了阴性选择算法相比于阳性选择算法所具有的优势, 指

出阴性选择算法在运算效率和鲁棒性之间取得了良好的平衡。文献[8]设计了一种检测器长度可变的阴性选择算法, 在一定程度上解决了“漏洞”区域问题, 并减少了检测器的冗余。然而, 众多的改进措施并不能实质性地改善阴性选择算法在可扩展能力和检测器的形态空间覆盖能力上的固有缺陷。此外, 阴性选择算法目前尚未能有效地拓展为通用的多类别模式分类模型, 这是基于以下原因: 首先, 阴性选择算法采用的是随机生成检测器的模式, 从而缺乏计算效率; 其次, 阴性选择算法没有良好的过拟合 (overfitting) 抑制机制, 从而缺乏抗噪能力以及推广能力; 第三, 阴性选择算法采用了不完全信息机制, 其检测器只包含一个类别的分类信息, 这对多类模式分类问题来讲是效率低下的。

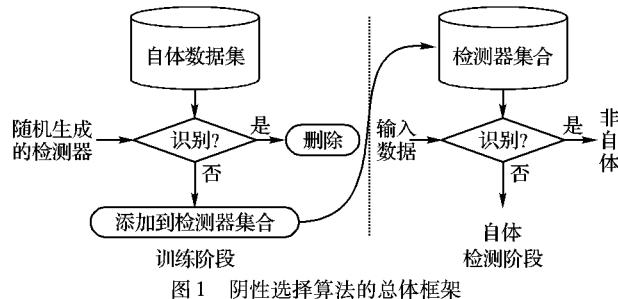


图 1 阴性选择算法的总体框架

本文面向多类别模式分类问题设计了一种新型的阴性选择算法 CS-NSA (Clonal Selection based Negative Selection Algorithm)。该算法通过在阴性选择算法的检测器生成环节引入克隆选择机制, 对检测器的生长进行启发式自适应学习。此

收稿日期: 2008-12-12; 修回日期: 2009-02-20。

作者简介: 刘殊(1968-), 男, 吉林榆树人, 高级讲师, 硕士, 主要研究方向: 网络、数据库。

外,针对模式分类的“过拟合”问题,还引入了检测器集合的修剪机制,增强了检测器的分类推广能力。另外,每一个检测器包含了除“自体”抗原以外的其他“非自体”类别抗原的信息,增强了其解决多类分类问题的计算效率。本文将CS-NSA与著名的人工免疫分类器AIRS<sup>[9]</sup>进行对比实验,实验结果证明CS-NSA具有更高的正确识别率和更强的抗噪声能力。

## 1 新型阴性选择算法CS-NSA

CS-NSA的基本思想是通过在阴性选择的框架下引入克隆选择机制,并在学习过程中结合过拟合抑制机制对分类器进行有效的训练。经过训练得到的分类器被称为记忆分类器,可用于对未知类别数据的分类。

我们首先对CS-NSA中的数据结构进行概括。算法中采用 $d$ 代表用于训练分类器的数据, $D$ 代表所有训练数据的集合;数据的维度用 $L$ 表示,总的类别数目用 $C$ 表示。 $d.v$ 和 $d.c$ 分别代表训练数据的向量和类别( $c = 1, 2, \dots, C$ ); $d.u$ 代表训练数据向量的第 $i$ 个分量值( $i = 1, 2, \dots, L$ )。 $d.u \in \{0, 1\}$ 用以表示训练数据是否被某个分类器所识别。 $D = D_1 \cup D_2 \cup \dots \cup D_C$ ,其中 $D_i \cap D_{j \neq i} = \emptyset$ 且 $d \in D_i \Leftrightarrow d.c = i$ 。算法中采用 $cl$ 表示分类器, $CL$ 表示全部分类器的集合。 $cl.v$ 和 $cl.c$ 分别代表分类器的向量和类别; $cl.r$ 代表分类器的识别半径。 $cl.sti_1$ 代表分类器的刺激值; $cl.sti_2$ 代表分类器的二级刺激值; $cl.active = cl.sti_1 + cl.sti_2$ 则代表分类器的总体活跃度。 $CL = CL_1 \cup CL_2 \cup \dots \cup CL_C$ ,其中 $CL_i \cap CL_{j \neq i} = \emptyset$ 且 $cl \in CL_i \Leftrightarrow cl.c = i$ 。此外,算法采用 $mcl$ 和 $MCL$ 分别代表记忆分类器和所有记忆分类器的集合。 $mcl$ 和 $MCL$ 的其他属性与上述的 $cl$ 以及 $CL$ 类似。

### 1.1 CS-NSA的分类器学习算法

CS-NSA分类器学习算法的目标在于训练有效的记忆分类器集合。具体来讲,算法依次生成每一个类别的记忆分类器集合;对记忆分类器集合 $MCL_i$ 而言, $D_i$ 将作为自体数据集,而 $D - D_i = \bigcup_{j \neq i} D_j$ 将作为非自体数据集,两类数据集共同参与学习过程。算法1描述了CS-NSA学习算法的总体流程。

#### 算法1 CS-NSA学习算法

```

For i = 1 to C
  MCLi ← ∅;
  For all d ∈ D - Di
    d.u ← 0;
  For all d ∈ D - Di
    If d.u == 0;
      cl ← generateMCL(i, d);
      MCLi ← MCLi ∪ cl;
    For all d ∈ D - Di
      If [ ∑k=1L (clk - dk)2 ]1/2 < cl.r
        d.u ← 1;

```

算法1中,generateMCL()被反复执行,直到记忆分类器集合能够识别所有的非自体数据。该算子融入了克隆选择机制,使所产生的记忆分类器具有更强的分类性能。算法2描述了generateMCL()的流程。

#### 算法2 算子generateMCL(i, d)

```

cl.c ← i; cl.v ← d.v;
cl.r ← min ∑e ∈ D_i, k=1L (clk - ek)2 / max ∑e ∈ D_i √ ∑k=1L (clk - ek)2; (1)

```

```

stimulation(i, cl); winner ← cl;
Do max_clone_times iterations
  clone ← cl;
  Do {
    mutate(clone);
    clone.r ← min ∑e ∈ D_i, k=1L (clonek - ek)2 / max ∑e ∈ D_i √ ∑k=1L (clonek - ek)2;
  } While [ ∑k=1L (clonek - dk)2 ]1/2 < clone.r
  stimulation(i, clone);
  If clone.sti1 > winner.sti1 winner ← clone;
  Else if clone.sti1 == winner.sti1
    If clone.sti2 > winner.sti2 winner ← clone;
    Else if clone.sti2 == winner.sti2 AND clone.r > winner.r
      winner ← clone;
  Return winner;

```

算子generateMCL()对父代分类器进行克隆,产生max\_clone\_times个克隆体。克隆体经历变异算子mutate()和激励算子stimulation(),并在此基础上根据刺激度相互竞争,最终的唯一获胜者“winner”作为算子的输出。变异算子mutate()采用了高斯变异。算子stimulation(i, cl)的流程如算法3所示。

#### 算法3 算子stimulation(i, cl)

```

For all e ∈ D - Di
  If √ ∑k=1L (clk - ek)2 < cl.r
    If e.u = 0 cl.sti1 ← cl.sti1 + 1;
    Else cl.sti2 ← cl.sti2 + 1;
    cl.active ← cl.sti1 + cl.sti2;

```

我们采用式(1)计算分类器的识别半径,其目的在于避免分类器识别自体数据,这是因为:

$$\frac{\min_{e \in D_i, k=1}^L (cl_k - e_k)^2}{\max_{e \in D_i} \sqrt{\sum_{k=1}^L (cl_k - e_k)^2}} = \rho \times \frac{\min_{e \in D_i} \sqrt{\sum_{k=1}^L (cl_k - e_k)^2}}{\max_{e \in D_i} \sqrt{\sum_{k=1}^L (cl_k - e_k)^2}} \quad (3)$$

其中:

$$\rho = \frac{\min_{e \in D_i} \sqrt{\sum_{k=1}^L (cl_k - e_k)^2}}{\max_{e \in D_i} \sqrt{\sum_{k=1}^L (cl_k - e_k)^2}} \leq 1 \quad (4)$$

从而有:

$$cl.r \leq \sqrt{\sum_{k=1}^L (cl_k - e_k)^2}, \forall e \in D_i \quad (5)$$

式(5)表明了分类器 $cl$ 避免了对自身数据的误识别。

### 1.2 CS-NSA的多类别分类算法

学习算法结束后,所得到的记忆分类器集合 $MCL_1, MCL_2, \dots, MCL_C$ 将被用于对未知类别数据的分类。在面对未知类别数据 $d$ 时,若 $MCL_i$ 中至少有一个分类器能够识别 $d$ ,则称 $MCL_i$ 识别 $d$ ;否则,称 $MCL_i$ 不能识别 $d$ 。根据阴性选择机制,分为三种情况:

- 1) 只有一类分类器 $MCL_i$ 不能识别 $d$ 。此时 $d$ 被归类为第*i*类。
- 2) 有不止一类分类器不能识别 $d$ ,设这些类别分类器的集合为 $MCL\_1$ 。此时,从 $MCL\_1$ 中选出与 $d$ 具有最大距离的

分类器  $\varphi$ ,  $d$  被归类为类别  $\varphi.c$ 。

3) 所有的分类器都能够识别  $d$ 。此时,从所有分类器中选出与  $d$  具有最小距离的分类器  $\psi$ ,  $d$  被归类为类别  $\psi.c$ 。

### 1.3 CS-NSA 的过拟合抑制机制

多类分类问题的过拟合现象往往是由噪声数据引起的。注意到在训练中识别噪声数据的分类器  $mcl_{noise}$  具有以下两个特点:其一,  $mcl_{noise}.r$  较小;其二,  $mcl_{noise}.active$  较小, 这是因为噪声数据所占的比例往往较小。事实上, 在基于 IRIS 数据集(见本文仿真实验环节的相关论述)的实验中, 我们发现如果消除原始数据集中的噪声数据, 使得数据集成为线性可分, 则所得到的记忆分类器的活跃度分布较为平均, 且多样性保持较好, 分类器最大活跃度达到 29; 而如果直接对原始含噪声数据进行学习, 则得到的记忆分类器的活跃度分布较为偏倚, 有大约 48% 的记忆分类器的活跃度小于 4, 这些分类器大多属于噪声分类器, 且有效分类器的多样性较差, 最大活跃度仅为 22。图 2 描述了在基于 IRIS 数据集的实验中得到的记忆分类器活跃度分布图。根据以上分析, 为了消除 CS-NSA 学习算法的过拟合效应, 我们设计了分类器集合的修剪机制。我们引入分类器活跃度阈值  $\sigma$ 。对于记忆分类器  $mcl$ , 若  $mcl.active \leq \sigma$ , 则将  $mcl$  从记忆分类器集合中删去。采用该机制, 不仅可以消除过拟合现象, 而且有利于控制分类器规模, 降低计算复杂度。

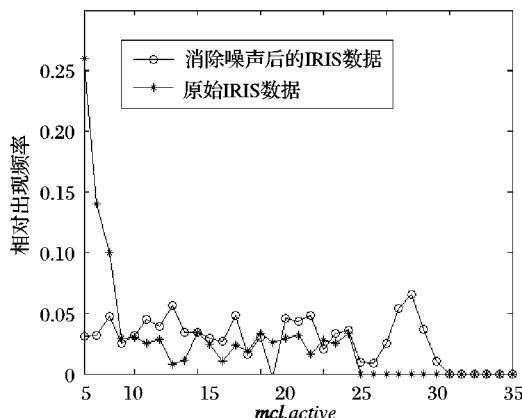


图 2 记忆分类器活跃度分布示意图(基于 IRIS 数据集)

## 2 仿真实验及结果分析

我们对 CS-NSA 的分类性能进行对比验证, 研究了分类器活跃度阈值  $\sigma$  以及克隆规模  $max\_clone\_times$  的取值对算法性能的影响, 并将 CS-NSA 其与著名的人工免疫分类器 AIRS<sup>[9]</sup>进行了性能对比。实验平台为 Intel Core Duo 1.66 GHz + 1.5 GB RAM + Windows XP, 采用 Java 编码。实验采用了 UCI 机器学习数据库中的 IRIS 数据集。该数据集共包含 150 条数据, 数据维度为 4, 所有数据分为 3 个类别, 每个类别均包含 50 条数据, 两两类别之间均线性不可分。在实验之前, 所有数据维度被归一化到区间 [0, 1]。

分类器活跃度阈值  $\sigma$  控制着算法对噪声数据的敏感度。我们针对 11 个不同的  $\sigma$  取值进行了实验, 以检验  $\sigma$  对分类效果的影响。 $\sigma$  的取值区间设定为 [0, 10], 取值步长为 1。对每一个  $\sigma$  取值进行 30 次独立实验, 每次实验采用 10 折交叉验证方法计算正确分类率, 并对 30 次实验的结果进行统计平均得到算法最终的正确分类率。注意到当  $\sigma$  取值为 0 的时候, 算法等同于没有实施过拟合抑制机制(见 1.3 节的相关分析)。图 3 对实验结果进行了描述, 其中图 3(a) 是在训练样本集上获

得的实验数据;图 3(b) 是从测试样本集上获得的实验数据。从图中可以看出, 在  $\sigma$  从 0 增大到 4 这一区间内, 不论在训练样本上还是在测试样本上, 算法的平均正确分类率均随着  $\sigma$  的增大而升高。其中训练样本平均正确分类率从 92.2% 增加到 97.2%, 测试样本平均正确分类率从 90.9% 增加到 96.5%。这与我们在 1.3 节所分析的分类器活跃度分布情况是相吻合的。然而, 当  $\sigma$  进一步增大时, 两类正确分类率均明显下降, 当  $\sigma$  增大到 10 时, 两类正确分类率分别降至 89.1% 和 87.2%。综合以上分析, 我们认为  $\sigma$  的最优取值为 4。

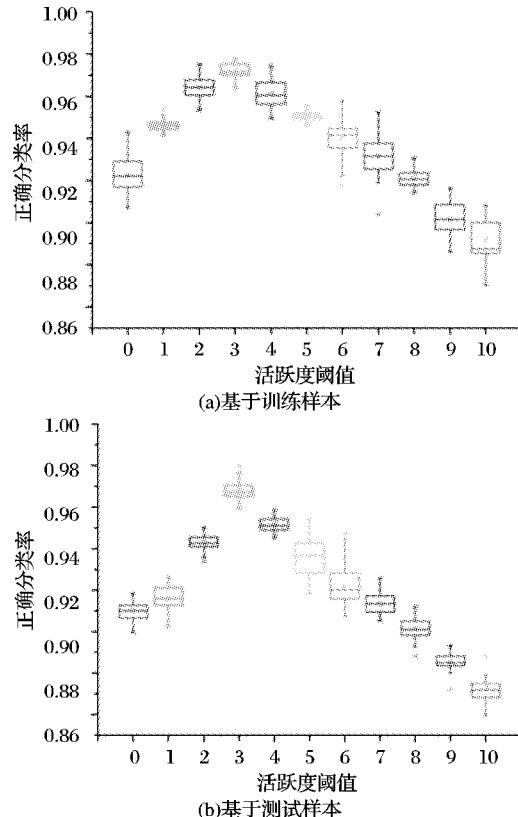


图 3 在不同的活跃度阈值取值情况下记忆分类器的平均正确分类率变化趋势

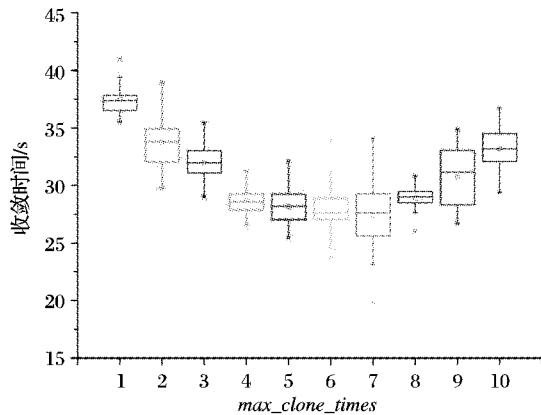


图 4 在不同的  $max\_clone\_times$  取值情况下算法收敛速度变化趋势

作为 CS-NSA 的另一个重要参数,  $max\_clone\_times$  对算法主要影响并不是体现在记忆分类器的正确分类率之上, 而是表现在算法的收敛速度之上。我们针对 10 个不同的  $max\_clone\_times$  取值进行了实验, 以检验其对分类效果的影响。取值区间设定为 [1, 10], 取值步长为 1。对每一个取值进行 30 次独立实验, 并对 30 次实验的结果进行统计平均得

(下转第 1589 页)

的推荐技术和协同过滤推荐技术,实现隐式评分,有效解决了稀疏问题和“冷开始”问题,提高了推荐质量;支持客商双方对推荐结果的讨价还价,促成实现推荐目的。但在实验中发现做出实时推荐依然存在不足,在今后的研究中,作者将试图采用分布式等手段,研究高性能算法,提高实时推荐和谈判的效果;根据“用户接受模型”等理论研究成果对用户偏好进行进一步研究,以更好支持推荐和谈判协商。



图3 议价界面

#### 参考文献:

- [1] KWAK M, CHO D S. Collaborative filtering with automatic rating for recommendation[C]// Proceedings of ISIE 2001. New York: Industrial Electronics, 2001: 625 – 628.
- [2] YANO E, SUEYOSHI E, SHINOHARA I, et al. Development of a recommendation system with multiple subjective evaluation process models[ C]// Proceedings of the 2003 International Conference on CyberWorlds. Washington, DC: IEEEComputerSociety, 2003: 344 – 351.
- [3] YU XIAO-GAO, JIAN YIN. A new clustering algorithm based on KNN and DENCLUE[ C]// Proceedings of ICMLC. Washington, DC: IEEE Press, 2005: 2033 – 2038.
- [4] 邓爱林. 电子商务推荐系统关键技术研究[D]. 上海: 复旦大学, 2003.
- [5] NICHOLS D M. Implicit rating and filtering [ EB/OL]. [ 2008 – 10 – 10 ]. <http://www.ercim.org/publication/ws-proceedings/DE-LOSS/nichols.pdf>.
- [6] SRINIVASA N, MEDASANI S. Active fuzzy clustering for collaborative filtering[ C]// Proceedings of 2004 IEEE International Conference on Fuzzy Systems. Washington, DC: IEEE Press, 2004: 1697 – 1702.
- [7] LEE W P. Towards Agent-based decision making in the electronic marketplace: Interactive recommendation and automated negotiation [ J]. Expert Systems with Applications, 2004, 27(4): 665 – 679.
- [8] YU XIAO - GAO , YU XIAO - PENG . A new k - nearest neighbor searching algorithm based on angular similarity[ C]//Proceedings of the 7th International Conference on Machine Learning and Cybernetics. Washington, DC: IEEE Press, 2008: 1779 – 1784.
- [9] 杨子晨, 孟波, 熊德林, 等. 谈判支持系统研究综述[ J]. 系统管理学报, 2002, 11(2): 100 – 108.
- [10] 邓爱林, 朱扬勇. 基于项目聚类的协同过滤推荐算法[ J]. 小型微型计算机系统, 2004, 25(9): 1665 – 1670.

(上接第 1584 页)

到算法最终的收敛速度。图 4 对实验结果进行了描述。随着  $max\_clone\_times$  从 1 增大到 6, 算法的平均收敛时间逐步缩减, 从 36.9 s 减小到 27.6 s。这得益于克隆选择机制中克隆规模的增大对分类器的局部优化带来的正面影响, 即克隆体越多, 越能够通过变异和自适应学习得到更为优秀的检测器, 从而算法只需要更少的时间来覆盖所有的非自体数据(见算法 1 和算法 2)。但是当  $max\_clone\_times$  进一步增大时, 算法的收敛时间却反而增加。这是因为过多的克隆数量将导致算法的克隆选择环节过于复杂, 使得时间过度消耗在克隆、变异以及对刺激度的比较之上。

我们采用相同的 IRIS 数据集以及相同的实验设置, 对著名的人工免疫分类器 AIRS 进行了实验, AIRS 的参数采用了文献[11]中的最优参数配置。在 30 次独立实验中, AIRS 的训练样本平均正确分类率为 97.0%, 测试样本平均正确分类率为 96.2%。而在  $\sigma$  取最优值 4 的情况下, CS-NSA 的这两类正确分类率达到了 97.2% 和 96.5% (见图 3), 这表明 CS-NSA 体现出较 AIRS 更为优秀的分类效果。

#### 3 结语

本文提出了一种新型的阴性选择算法 CS-NSA。首先, 在 CS-NSA 的学习算法中融合了克隆选择的思想, 从而使分类器具备了更强的分类性能。其次, CS-NSA 将阴性选择算法有效地拓展到多类别模式分类问题, 这对扩展阴性选择算法的应用具有积极意义。第三, CS-NSA 引入了分类器集合的过拟合抑制机制, 有效地降低了噪声数据对分类结果推广能力的负面影响。

#### 参考文献:

- [1] DASGUPTA D. Advances in artificial immune systems [ J]. IEEE Computational Intelligence Magazine, 2006, 1(4): 40 – 49.
- [2] HART E, TIMMIS J. Application areas of AIS: the past, the present and the future [ EB/OL]. [ 2008 – 10 – 10 ]. <http://www-users.cs.york.ac.uk/jtimmis/utm/Papers/JASCfinal.pdf>.
- [3] FREITAS A, TIMMIS J. Revisiting the foundations of artificial immune systems for data mining [ J]. IEEE Transactions on Evolutionary Computation, 2007, 11(4): 521 – 540.
- [4] FORREST S , PERELSON A S , ALLEN L , et al . Self-nonself-discrimination in a computer[ C]// Proceedings of IEEE Symposium on Research in Security and Privacy. Oakland, CA: IEEE Press, 1994: 202 – 212.
- [5] GONZALEZI F, DASGUPTA D, NINO1 L F. A randomized real-valued negative selection algorithm[ C]// ICARIS 2003: Proceedings of the 2nd International Conference on Artificial Immune Systems, LNCS 2787. Berlin: Springer-Verlag, 2003: 261 – 272.
- [6] ZHOU JI , DASGUPTA D . Real-valued negative selection algorithm with variable-sized detectors [ C] // GECCO 2004: Proceedings of Genetic and Evolutionary Computation Conference, LNCS 3102. Berlin: Springer-Verlag, 2004: 287 – 298.
- [7] ZHOU JI , DASGUPTA D . Applicability issues of the real - valued negative selection algorithms[ C]// GECCO 2006: Proceedings of 8th Annual Conference on Genetic and Evolutionary Computation. New York: ACM Press, 2006: 111 – 118.
- [8] 何申, 罗文坚, 王熙法. 一种检测器长度可变的非选择算法 [ J]. 软件学报, 2007, 18(6): 1361 – 1368.
- [9] WATKINS A , TIMMIS J . Artificial immune recognition systems (AIRS): An immune-inspired supervised algorithm [ J]. Genetic Programming and Evolvable Machines, 2004, 5(3): 291 – 317.