第 34 卷　第 15 期　　计　算　机　工　程　　2008 年 8 月
Vol.34　No.15　　Computer Engineering　　August 2008
• 人工智能及识别技术 •　文章编号：1000—3428(2008)15—0223—03　文献标识码：A　中图分类号：TP301

基于人工免疫的支持向量机模型选择算法

姚全珠，田　元
(西安理工大学计算机科学与工程学院，西安 710048)

摘　要：支持向量机中参数设置对训练支持向量机分类的精确度有不可忽视的影响。支持向量机参数的选取可看作参数的组合优化。免疫算法是一种有效的随机全局优化技术，它具有不易陷入局部最优解、解精度高、收敛速度快等优点。该文利用人工免疫算法进行支持向量机模型选择。该算法主要包括克隆选择、高频变异、受体编辑等操作。试验证明，该算法能够有效提高支持向量机分类的正确性。
关键词：支持向量机；模型选择；免疫算法

Model Selection Algorithm of SVM Based on Artificial Immune

YAO Quan-zhu, TIAN Yuan
(School of Computer Science and Engineering, Xi'an University of Technology, Xi'an 710048)

【Abstract】 The parameters setting for Support Vector Machine(SVM) in a training process impacts on the classification accuracy. The selection problem of SVM parameters is considered as a compound optimization problem. Immune algorithm is an efficient random global optimization technique. It has nice performances such as avoiding local optimum, high precision solution, and quick convergence. This paper proposes an immune algorithm applied to model selection of SVM. This algorithm includes clonal selection, hyper-mutation and receptor editing. Experimental results indicate that this method significantly improves the classification accuracy of SVM.
【Key words】 Support Vector Machine(SVM); model selection; immune algorithm

1　概述

支持向量机理论最早由 Vapnik[1]提出，该理论是一种基于统计学习理论中 VC 维理论和结构风险最小理论的通用学习方法。它可以解决小样本学习问题，而且对数据的维数、多变性不敏感，能够较好地进行模型选择并具有良好的推广能力。

目前已经在许多智能信息获取与处理领域都取得了成功的应用[2]。支持向量机的成功很大程度上依赖于核函数技术(kernel tricks)的成功应用，该技术推动了支持向量机处理传统数据的研究。支持向量机的模型选择问题就是给定一个核函数，通过调节核参数和惩罚因子 C 来提高支持向量机训练精度，同时降低错误率，因此支持向量机的参数选择直接影响着 SVM 的性能[3]。目前，参数的选取缺乏理论指导，大多数核函数中参数的选取还只能凭借先验知识，甚至通过猜的手段来确定。

本文受到文献[4]的启发，将支持向量机参数选取看作是函数最优化问题，采用人工免疫算法来求解目标函数的最优值，即找到核参数和惩罚因子 C 的最优值。通过实验证明，本文提出的算法可以为支持向量机找到较好的参数，同时具有训练速度快的优点。

2　支持向量机原理

支持向量机可以用于数据分类问题的处理。下面针对训练样本集：2 类线性和 2 类非线性分别加以讨论。

对于 2 类线性可分问题，已知：训练集包含 l 个样本点：

$$T=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_l,y_l)\}\in(\boldsymbol{x}\times y)^l$$

$$\boldsymbol{x}_i\in\boldsymbol{X}=R^n,y_i\in Y=\{1,-1\}$$

其中，$\boldsymbol{x}_i$ 是向量，其分量称为特征；Y 是输出。支持向量就是寻求一个平面 $\omega\cdot\boldsymbol{x}+b=0$，使得训练数据点距离这个分类面尽量得远。这种极大化“间隔”的思想导致求解对变量 ω 和 b 的最优问题：

$$\min_{w,b}=\frac{1}{2}\|\omega\|^2\qquad y_i\{(\omega\cdot\boldsymbol{x}_i)+b\}\geq 1,i=1,2,\cdots,n \tag{1}$$

为求解原始问题，根据最优化理论，可以转化为对偶问题来求解

$$\min_{\alpha}\ \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}y_iy_j\alpha_i\alpha_j(\boldsymbol{x}_i\cdot\boldsymbol{x}_j)=\sum_{j=1}^{l}\alpha_j$$

$$\text{s.t}\quad\sum_{i=1}^{l}y_i\alpha_i=0, \tag{2}$$

$$\alpha_i\geq 0\qquad i=1,2,\cdots,l$$

解上述问题后得到的分类规则函数是

$$f(\boldsymbol{x})=\text{sgn}\{(w\cdot\boldsymbol{x})+b\}=\text{sgn}\left\{\sum_{i=1}^{l}a_i\cdot y_i(\boldsymbol{x}_i\cdot\boldsymbol{x})+b\right\} \tag{3}$$

对于 2 类线性不可分问题，为第 i 个训练点 $(\boldsymbol{x}_i,y_i)$ 引入松弛变量(Slack Variable) $\xi_i\geq 0$，把约束条件放松到 $y_i\{(\omega\cdot\boldsymbol{x}_i)+b\}+\xi_i\geq 1,i=1,2,\cdots,n$(即“软化”约束条件)。显然 ξ_i 可以描述为训练集错划的程度。现在就有 2 个目标：希望超平面间隔最大，即($\frac{2}{\|\omega\|}$最大)；又希望训练集错划程度 $\sum_{i=1}^{l}\xi_i$ 尽可能小，所以引入惩罚参数 C。在实际使用时，可以选择 C 来修改这 2 个目标的权重。这样新的目标函数成为

删除的内容:

删除的内容:：

带格式的: 项目符号和编号

作者简介：姚全珠(1960－)，男，教授，主研方向：软件工程，数据挖掘；田　元，硕士
收稿日期：2007-09-07　**E-mail：**tyonline@sina.com

$$\min_{\omega,b,\xi} \quad \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{l}\xi_i$$
$$\text{s.t} \quad y_i\left((\omega\cdot \boldsymbol{x}_i)+b\right)+\xi_i \geq 1,\ i=1,2,\cdots,l \qquad (4)$$
$$\xi_i \geq 0, i=1,2,\cdots,l$$

其中，$\sum_{i=1}^{l}\xi_i$ 体现了经验风险；而 $\|w\|$ 则体现了分类表达能力。所以惩罚参数 C 实质上是对经验风险和表达能力匹配的一个裁决。当 $C\to\infty$ 时，线性不可分的原始问题退化为线性可分。

对于非线性分类，通过引入核函数，将原空间样本数据通过非线性变换映射到高维特征空间 $H:\Phi:R^d\to H$。在高维空间中求最优或广义最优分类面。

常用的核函数如下：

(1)多项式核函数

$$K(\boldsymbol{x},\boldsymbol{x}_i)=\left[(\boldsymbol{x}\cdot\boldsymbol{x}_i)+1\right]^d$$

(2)径向基核函数(Radial Basis Function, RBF)

$$K(\boldsymbol{x},\boldsymbol{x}_i)=\exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x}_i\|^2}{2\sigma^2}\right)$$

(3)Sigmoid 核函数

$$K(\boldsymbol{x},\boldsymbol{x}_i)=\tanh\left(b(\boldsymbol{x}\cdot\boldsymbol{x}_i)-c\right)$$

其中，b,c 为常数。这样，对于非线性分类问题，最终归结为一个二次规划问题：

$$\min_{\alpha} \quad \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}y_iy_j\alpha_i\alpha_jK(\boldsymbol{x}_i\cdot\boldsymbol{x}_j)-\sum_{j=1}^{l}\alpha_j = -e^{\mathrm{T}}\alpha+\frac{1}{2}\alpha^{\mathrm{T}}Q^{\mathrm{T}}\alpha$$
$$\text{s.t} \quad \sum_{i=1}^{l}y_i\alpha_i=0,$$
$$0\leq\alpha_i\leq C \quad i=1,2,\cdots,l$$

3 自然免疫系统

生物免疫系统是一个高度进化、复杂的系统，它还具有学习、记忆和自适应的调节能力。该系统是由免疫分子、免疫细胞、免疫组织和免疫器官组成的复杂系统，其重要功能是通过产生抗体抗御病原体即抗原的侵害，以保证生命运行规律的稳定[5]。

人体免疫系统有先天性免疫系统和适应性免疫系统，先天性免疫是与生俱有的，有能力识别侵入体内的各种微生物；适应性免疫是后天形成的，也称获得性免疫，它由淋巴细胞中的 B 细胞和 T 细胞受病原体(抗原)的刺激、诱导后发生增殖、分化以及特异性免疫响应而形成的[4]。一种 B 细胞只产生一种特异性抗体，有的以分泌形式分布于血液和组织液中，有的结合在 B 细胞膜上作为抗原受体。这些受体有选择地结合抗原的特定部分，它们之间的匹配程度称为亲和度。亲和度强的抗体 B 细胞通过增殖和分化，不断地产生新的 B 细胞和新的抗体，部分亲和力弱的 B 细胞就会消亡淘汰。B 细胞在增殖、分化过程中，同时经历着超变异[4]。

病原体侵入肌体，引起先天性和获得性免疫反应。获得性免疫反应产生能够识别抗原的抗体，抗体的形状与抗原对应或部分对应，并捕获大量的抗原。抗体与抗原发生特异性结合后，通过中和、溶解和调理等作用，最终使抗原从体内清除[6]。

4 基于人工免疫的支持向量机模型选择算法

人工免疫算法的思想来源于生物免疫的原理。它是通过抽取和反映生物免疫系统，结合工程应用而人工模拟的一种计算模型。它模仿生物的免疫过程，具有良好的全局搜索能力和记忆功能。人工免疫系统中免疫算法已经用于机器学习、异常和故障诊断、机器仿真、网络入侵检测、参数优化、工业设计等领域，并表现出较卓越的性能和效率。

关于支持向量机模型选择，国内外学者提出了很多种核函数参数选取的方法[7]。最简单的方法就是，对参数进行组合，然后进行网格搜索，最后把精度最高的一个组合作为最优参数，可以看出这是一个非常耗时的过程而且精确度不高。文献[7]提出的基于梯度的方法，虽然可以有效地进行参数选择，但是这种方法对一些核函数求导困难、通用性差。因此，本文提出了基于人工免疫的支持向量机模型选择算法。

4.1 算法描述

本文将抗原作为目标函数，然后随机产生一组抗体作为目标函数的解，并计算抗原和抗体之间的亲和度，将亲和度作为可行解与最优解的逼近程度。接下来对抗体进行高频变异和受体编辑生成下一代抗体群进行优化，直至满足终止条件，算法结束。

在详细介绍该算法之前，先介绍该算法用到的一些定义和算法。

定义 1 支持向量机中的参数称为抗体。该算法中的抗体以实数进行编码。以 RBF 核函数来举例说明，此时支持向量机中有 2 个参数，那么抗体编码由 2 部分组成，分别存放 C,σ。如果核函数有 2 个参数，则抗体编码由 3 部分构成，分别存放核参数的 2 个参数和惩罚参数 C，其他有更多参数的情况依此类推。

定义 2 可行解对问题的满足程度称为亲和度。该算法用分类器与参考模型的方均误差(MSE)，即 $f^2=\frac{1}{n}\sum_{i=1}^{n}(y-f(x))^2$ 作为抗体和抗原的亲和度。

4.1.1 克隆选择算法

首先将抗体集中的抗体按亲和度由大到小排列，然后按下式计算克隆集规模，根据克隆集规模选择抗体到克隆集中。

$$Nc=\sum_{i=1}^{N}\left\lfloor\frac{\alpha N}{i}\right\rfloor \qquad (5)$$

其中，Nc 为克隆的规模；α 为克隆系数，该参数用来控制克隆的规模；$\lfloor\ \rfloor$ 操作符用来表示取整数。从式(5)可以看出，亲和度越高的抗体，克隆的规模也就越大，算法在很大程度上使高亲和度抗体得以更好地保存和发展。

4.1.2 高频变异算法

对克隆集中每个抗体进行高频变异，得到变异集。变异的算子作用到抗体的每一个分量：

$$x_i^{'}=x_i+\gamma_i\cdot N(0,1)$$
$$\gamma_i=(1/\beta)\cdot\exp(-f^2) \qquad (6)$$

其中，$x_i^{'}$ 是变异体；$N(0,1)$是均值为 0、方差为 1 的独立正态随机变量；β 为变异控制系数；f^2 为抗体和抗原的亲和度。

高频变异机制是为了使免疫应答能快速地成熟并增加抗体的多样性。如果对克隆集中的抗体进行相同频率的变异可能产生不良的变化，这种不良的变化可能抵消掉有利的变异，从而引起更加糟糕的抗体。式(6)的高频变异算法按抗体的亲和度进行，低亲和度的抗体可能会经历更深度的变异，高亲和度抗体的变异可能被抑制，这样有利于抗体向最优解收敛。

4.1.3 受体编辑算法

去掉亲和度低的抗体，计算剩余抗体的数量 Ns 并标记这些抗体为记忆抗体，随机从记忆集选择抗体，替换进入下一代演化的抗体，即随机产生 $d\%\times Ns$ 个抗体进入下一代的演化抗体。

带格式的：项目符号和编号

删除的内容：这

带格式的：项目符号和编号

带格式的：项目符号和编号

带格式的：项目符号和编号

删除的内容：),

删除的内容：,

删除的内容：,

删除的内容：应

带格式的：项目符号和编号

高频变异是对某个抗体的周围进行搜索，以找到更高亲和度的抗体，但这容易导致局部极值。受体编辑允许在宽范围内进行搜索，虽然有可能导致新找到抗体是在亲和度更低的区域，但是这样的跳跃可能跳到一个更佳的位置，再通过高频变异，达到全局最优。因此，受体编辑提供了取消局部极值的能力。

4.2 基于人工免疫的支持向量机模型选择算法

由以上描述可总结出算法的主要步骤如下：

Step1 初始化，主要是将目标函数(SVM)设置为抗原，其最优参数设置为抗体。

Step2 随机产生初始抗体群 Ab1，抗体群中的抗体视为待优化问题的 N 个候选解。

Step3 计算抗体和抗原之间的亲合度。

Step4 根据克隆选择算法从 Ab1 中选择抗体加入到克隆集 Ab2 中。

Step5 对克隆集中的抗体用高频变异算法进行变异，产生抗体群 Ab3。

Step6 从 Ab1 中淘汰亲和度低的抗体，然后加入到抗体集 Ab4 中。

Step7 从 Ab3 中选出高亲和度的抗体组成 Ab5，淘汰 Ab5 中相似的抗体加入到 Ab4 中，然后由受体编辑算法生成新一代的抗体群体。

Step8 优化过程完成则继续，否则跳到 Step3。

Step9 输出优化结果，结束。

5 实验

为了验证基于人工免疫的支持向量机模型选择算法的有效性，从 UCI 机器学习知识库[8]中选取了 5 组分类数据集进行实验。

实际应用中核函数的种类有很多，研究表明，当缺少过程的先验知识时，选择高斯核函数比选择其他核函数好[9]。因此，多数应用研究都采用高斯核函数。考虑到这个原因，本次试验选取了径向基核函数作为实验目标。本文对所有学习精度的估计，均采用 k 折交叉验证，取 k=5。本算法中的其他参数为：抗体规模集 N=50；克隆系数 α=0.1；变异控制系数 β= 30。

本次试验对网格搜索算法(GS)、人工免疫的支持向量机模型选择算法(IAS)进行测试，2 种算法均采用标准 C++实现，使用 Microsoft Visual C++6.0，缺省编译器优化选项进行编译。系统平台为 2.66 GHz Pentium 4 处理器，Windows_2000 标准版，256 MB RAM。

通过表 1 中的实验数据可以看出，对于同一个测试数据集，基于人工免疫的支持向量机模型选择算法的学习精度均高于网格搜索算法。从表 2 可以看出，前者的寻优速度也均高于后者的速度。因此可以得出结论，基于人工免疫的支持向量机模型选择算法可以有效改进学习性能，提高学习精度。

表 1 2 种算法的检测精度 (%)

数据集	GS/	IAS
German	77.705 91	82.790 8
heart	84.305 7	96.135 7
Australian	86.215 3	90.413 2
Diabetes	78.121 2	83.151 1
Vehicle	82.391 6	84.017 2

表 2 2 种算法的学习时间 s

数据集	GS	IAS
German	995.06	131.66
heart	30.11	5.73
Australian	212.93	44.38
Diabetes	693.30	85.17
Vehicle	432.28	60.84

6 结束语

本文研究了自然免疫系统以及人工免疫系统。借用免疫算法的优点，提出基于人工免疫的支持向量机模型选择算法。通过实验结果证明，这种算法能够快速地收敛到全局最优解，从而明显提高 SVM 的模型的学习精度，缩短学习时间。

参考文献

[1] Vapnik V N. The Nature of Statistical Learning Theory[M]. New York, USA: Springer, 1995: 23-60.

[2] Sanchez A D. Advanced Support Vector Machines and Kernel Methods[J]. Neurocomputing, 2003, 55(1): 5-20.

[3] Müller K R, Mika S, Ratsch G, et al. An Introduction to Kernel-based Learning Algorithms[J]. IEEE Transactions on Neural Networks, 2001, 12(2): 181-202.

[4] 罗印升，李人厚，张 雷，等. 人工免疫算法在函数优化中的应用[J]. 西安交通大学学报, 2003, 7(8): 840-843.

[5] 杨延彬. 免疫学及检验[M]. 北京：人民卫生出版社, 1999: 1-65.

[6] Zheng Hong, Zhang Jingxin, Nahavandi S. Learning to Detect Objects by Artificial Immune Approaches[J]. Future Generation Computer Systems, 2004, 20(7): 1197-1208.

[7] Chapelle O, Vapnik V. Choosing Multiple Parameters for Support Vector Machines[J]. Machine Learning, 2002, 46(1): 131-159.

[8] Murphy M. UCI-Benchmark Repository of Artificial and Real Data Sets[EB/OL]. (1995-09-15). http://www.ics.uci.edu/～mlearn/.

[9] Smola A J. Learning with Kernels[D]. Berlin, Germany: Technical University of Berlin, 1998.

（上接第 207 页）

[3] Maes S, Tuyls K, Bram V, et al. Credit Card Fraud Detection Using Bayesian and Neural Networks[C]//Proc. of the 1st International NAISO Congress on Neuro Fuzzy Technologies. Havana, Cuba: [s. n.], 2002.

[4] Flach P. The Many Faces of ROC Analysis in Machine Learning[C]//Proc. of ICML'04. Valencia, Spain: [s. n.], 2004: 485-487.

[5] Han Jiawei, Kamber M. 数据挖掘概念与技术[M]. 范 明, 译. 北京：机械工业出版社, 2006: 213-214.

[6] Jeske D R, Gokhale D V. Generating Synthetic Data from Marginal Fitting for Testing the Efficacy of Data Mining Tools[J]. International Journal of Production Research, 2006, 44(14): 2711-2730.

带格式的：项目符号和编号

删除的内容: 表

带格式的：项目符号和编号

删除的内容: ,

删除的内容: 果

带格式的：项目符号和编号

删除的内容: