第23卷第2期
2006年4月

控制理论与应用
Control Theory & Applications

Vol. 23 No. 2
Apr. 2006

文章编号：1000－8152(2006)02－0187－06

基于启发式遗传算法的SVM模型自动选择

郑春红[1,2]，焦李成[2]，丁爱玲[2]

（1. 西安电子科技大学 电子工程学院，陕西 西安 710071；2. 西安电子科技大学 智能信息处理研究所，陕西 西安 710071）

摘要：支撑矢量机(SVM)模型的自动选择是其实际应用的关键. 常用的基于穷举搜索的留一法(LOO)很繁杂且效率很低. 到目前为止，大多数的算法并不能有效地实现模型自动选择. 本文利用实值编码的启发式遗传算法实现基于高斯核函数的SVM模型自动选择. 在重点分析了SVM超参数对其性能的影响和两种SVM性能估计的基础上，确定了合适的遗传算法适应度函数. 人造数据及实际数据的仿真结果表明了所提方法的可行性和高效性.

关键词：支撑矢量机(SVM)；模型选择；模型自动选择；遗传算法

中图分类号：TP18，TP391　　**文献标识码**：A

Automatic model selection for support vector machines using heuristic genetic algorithm

ZHENG Chun-hong[1,2], JIAO Li-cheng[2], DING Ai-ling[2]

(1. School of Electronic Engineering, Xidian University, Xi'an Shaanxi 710071, China;
2. Institute of Intelligent Information Processing, Xidian University, Xi'an Shaanxi 710071, China)

Abstract: Motivated by the facts that automatic model selection for support vector machine (SVM) is an important issue to make it practically useful, and the commonly-used leave-one-out (LOO) method is complex and time consuming, we proposed an effective strategy for automatic model selection for SVM with Gauss kernel by using a heuristic real-coded genetic algorithm (GA). Based on the extensive analysis of the effects of the hyper-parameters on the generalization performance and two estimates of SVM, the appropriate fitness function for GA operation is determined. Simulations are performed on both artificial data and real data to demonstrate the effectiveness and efficiency of the proposed approach. The significance of the proposed method is its easy implementation and better performances in comparison with the commonly used loo method.

Key words: support vector machine; model selection; automatic model selection; genetic algorithm

1 引言(Introduction)

支撑矢量机(support vector machine，简称SVM)在分类、回归等领域得到了广泛的应用[1]，它是由AT&T Bell实验室的Vapnik在1992年提出的基于统计学习理论且推广能力非常好的一种新型小样本学习机[2,3]. 正如大多数学习机算法，SVM中的模型选择问题在解决过匹配和欠匹配的折中问题上也是一个关键所在，特别是在小样本学习中，内嵌超参数的SVM如果参数选择不当，会导致系统性能恶化. 对任何一个分类问题，确定SVM的超参数使得期望的测试误差最小被称为SVM的模型选择[4~9]. SVM的模型选择本质上是一个非线性非二次规划非常复杂的全局优化问题. 大多采用交叉验证实现模型选择，较为有效的交叉验证方法是“留一法”(leave-one-out，简称LOO)[5]. “留一法”参数选择，首先需根据人工经验确定一个近似的最优参数集范围，然后在该参数集上用“留一法”遍历搜索最优参数. 该方法的计算量非常庞大，仅在一个参数点上就要求训练$l-1$次(l为样本数).

近年来，为了加速SVM的训练，利用SVM推广能力的相关界进行模型自动选择的方法纷纷被提出[6~9]. Lee与Lin[6]研究了LOO测试误差率和SVM分解方法的终止准则之间的关系，通过引入分解方法的一个非常松散的终止准则实现SVM模型

收稿日期：2005－02－01；收修改稿日期：2005－06－22.
基金项目：国家自然科学基金资助项目(60133010,60372047)；西安电子科技大学博士点基金资助项目；西安电子科技大学青年工作站项目资助项目.

的自动选择. 事实上,该方法是充分利用全部训练样本,采取穷举搜索的方法最小化期望 LOO 测试误差率得到最优超参数. 当所给样本有限时,这一方法显出了它的局限性. Chapelle 等[7]应用梯度下降算法实现 SVM 模型的自动选择. Keerthi[8]采用拟牛顿法进行高斯核函数 SVM 模型自动选择. Gold 与 Sollich[9]应用混合 Monte-Carlo 算法实现 SVM 模型的自动选择. 上述基于梯度的数值方法可能会陷于局部最优解;而且,正如 Keerthi 所指出的,如果迭代的初值选择不当,就更不会获得令人满意的模型参数[8]. 然而,十分令人遗憾的是,除非作者对某一个问题具有很好的经验和十足的知识,否则,并不容易获得问题合适的初值.

遗传算法(genetic algorithm, 简称 GA)是一类借鉴生物界自然选择和自然遗传机制的随机搜索算法,较以往传统的搜索算法具有使用方便、鲁棒性强、便于并行处理等特点,广泛应用于各种领域[10~12]. 本文采用 GA 进行 SVM 模型参数的自动选择. 在重点分析了 SVM 超参数对其性能的影响和两种性能估计的基础上,给出了 GA 的适应度函数,利用实值编码的启发式遗传算法实现了 SVM 的模型自动选择. 对人造数据和实际数据的试验结果,证明了所提方法的可行性和高效性.

2 SVM 模型选择(Model selection for SVMs)

2.1 支撑矢量机(SVMs)

本文以支撑矢量分类器展开. 已知一组训练样本 $X = \{(x_i,y_i)\,|\,x_i \in \mathbb{R}^n, y_i = \pm 1, i = 1,\cdots,l\}$, $\mathbb{R}^n$ 表示输入空间. 支撑矢量机通过一个非线性映射,可以将输入矢量 x 映射到一个高维特征空间,并构造一个最优超平面(即最大间隔超平面),这个最优的超平面满足如下约束条件

$$y_i((w \cdot x_i) - b) \geqslant 1 - \xi_i,\ i = 1,\cdots,l, \quad (1)$$

最小化函数

$$\Phi(w,\xi) = \frac{1}{2}(w \cdot w) + C\Big(\sum_{i=1}^{l}\xi_i\Big). \quad (2)$$

其中: $\xi_i \geqslant 0$ 是对误差的度量, C 是惩罚系数, 在决策函数的复杂性以及训练样本的误分数之间起一个折中作用.

利用对偶规则可以将上述问题转化为易于求解的对偶问题

$$\alpha^* = \underset{\alpha}{\operatorname{argmin}} \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j K(x_i,x_j) - \sum_{k=1}^{l}\alpha_k. \quad (3)$$

约束条件为

$$\begin{cases} 0 \leqslant \alpha_i \leqslant C,\ i = 1,\cdots,l, \\ \sum_{j=1}^{l}\alpha_j y_j = 0. \end{cases} \quad (4)$$

其中: $K(x_i,x) = \varphi(x_i)\cdot\varphi(x)$ 为核函数,实现输入空间到高维空间的映射;对应于拉格朗日乘子 α 不为零(记为 α^*)的样本称为支撑矢量. SVM 的间隔 $\gamma = 1/\|w\|$ 可由下式计算:

$$\gamma = \Big(\sum_{i,j\in SV} y_i y_j \alpha_i^* \alpha_j^* K(x_i,x_j)\Big)^{-1/2}. \quad (5)$$

注意对偶形式下并不需要计算 ξ, 而 C 是要选择的模型参数之一.

上述的核函数可以根据 Mercer's 定理确定. 通常采用高斯函数作为核函数,可表示为[3]

$$K(x_i,x_j) = \exp\Big(-\frac{\|x_i - x_j\|_2}{2\sigma^2}\Big). \quad (6)$$

其中 σ 为核函数参数.

SVM 推广能力的优劣可以通过其间隔和支撑矢量的数目等泛化性能指标来进行评价. 一般来说,具有最大间隔和最小支撑矢量数目的支撑矢量机具有最好的推广能力, 对测试集的测试误差最小. 图 1 与图 2 分别示出了随机产生的含 50 个样本点两类数据的 SVM 的间隔 γ 及支撑矢量 N_{sv} 数随模型参数 σ 与 C 的变化.

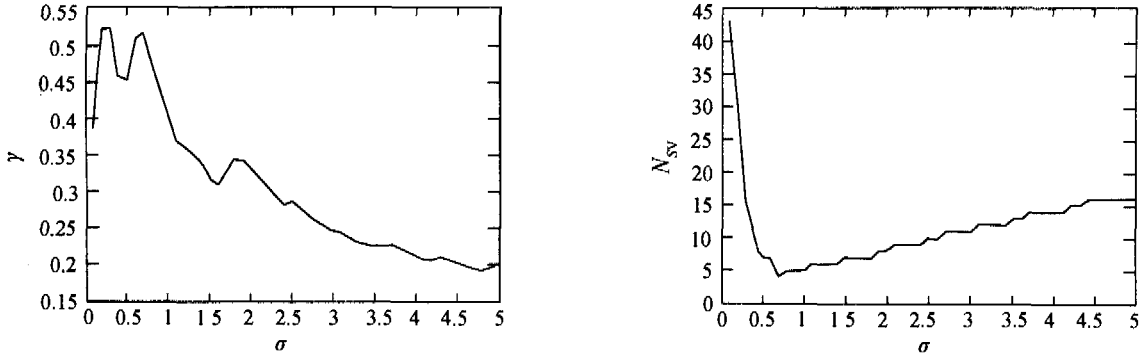


图 1 SVM 间隔 γ 与支撑矢量数 N_{sv} 随核参数 σ 的变化(C = 10)

Fig. 1 Variation of SVM margin and number of SVs with σ (C = 10)

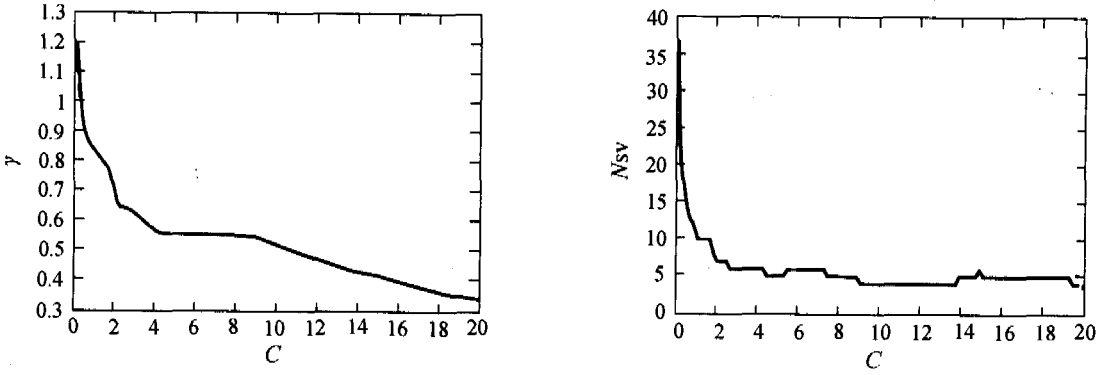


图2　SVM间隔γ与支撑矢量数N_{SV}随惩罚系数C的变化($\sigma = 0.7$)

Fig. 2　Variation of SVM margin and number of SVs with $C(\sigma = 0.7)$

由图1和图2可以看出,模型参数σ与C对间隔和支撑矢量数有较大的影响,要想同时达到最大间隔及最小支撑矢量数是比较困难的.因此,需要恰当地确定SVM模型参数才能使得所构造的SVM对测试样本的误差率最小.

直接选择参数而使SVM的实际误差最小是很难实现的.最常用的方法是LOO交叉验证法.LOO交叉验证法针对每一组可能的参数需要进行$l-1$次优化,构造SVM,计算测试集误差,得到一个平均误差;在此基础上遍历所有可能的参数,最终得到一组使得测试集平均误差最小的参数[5].当有足够的数据可用时,通过交叉验证也许可以得到适当的参数,但这将是一个非常繁杂且效率很低的过程.对于小样本数集,可以通过建立估计或相关界来实现模型选择.

至此,高斯核函数SVM模型选择问题可描述为:如何合理地确定$\theta = (C,\sigma)$使得所获得的SVM具有最好的泛化性能,即对测试集具有最佳的识别率.

2.2　性能指标(Performance estimates)

SVM模型选择通常都是通过最小化某种泛化性能指标来实现的.下面,简单介绍两种常用的SVM泛化性能指标.

2.2.1　直观估计(Intuitional estimate)

直观估计就是在求解SVM最大间隔超平面(即最大化间隔γ)的同时,使得支撑矢量数目最少.

根据SVM在高维空间中的推广定理,l个样本被最大间隔超平面分开的测试误差概率的期望值(对训练集)为[3]

$$\mathrm{E}p_{\mathrm{error}} \leqslant \mathrm{E}\min\left(\frac{m}{l},\frac{[R^2/\gamma^2]}{l},\frac{n}{l}\right). \tag{7}$$

其中:m为支撑矢量的个数;R是将数据包含其中的最小超球半径,γ是间隔值,n是输入空间的维数.

在SVM算法中,重点考虑的是式(7)的前两项,即数据压缩的期望是大的(这意味着支撑矢量的个数要小),同时间隔的期望是大的.因此,如下的测试误差概率界(即直观估计)作为实现模型选择的性能指标

$$T = \frac{m}{l} + \frac{1}{\gamma}. \tag{8}$$

2.2.2　半径-间隔界(Radius-margin bound estimate)

根据上述的推广定理[3],作者知道SVM泛化性能决定于$\mathrm{E}\{R^2/\gamma^2\}$,而不是仅仅依靠其间隔$\gamma$.Chapelle等[7]根据LOO的误差数提出下面的测试误差界:

$$T = \frac{1}{l}\frac{R^2}{\gamma^2}. \tag{9}$$

其中R和γ分别为上述定义的半径和间隔.

R^2的最小化可在$\|x_i - x^*\|^2 \leqslant R^2$的约束条件下,使目标函数$R^2 - \sum\beta_i(R^2 - (x_i - x^*)^2)$最小化来求解[2,3].其中:$x^*$为待确定的球心位置向量,$x^* = \sum_{i=1}^{l}\beta_i x_i$.经推导,$R^2$可通过下述的二次优化的最优解来获得[2,3,8]

$$\begin{cases}\max \sum_{i=1}^{l}\beta_i K(x_i,x_i) - \sum_{i,j=1}^{l}\beta_i\beta_j K(x_i,x_j) \\ \text{s.t.} \begin{cases} 0 \leqslant \beta_i,\ i = 1,\cdots,l, \\ \sum_{i=1}^{l}\beta_i = 1. \end{cases}\end{cases} \tag{10}$$

为了简化计算,本文采用和Keerthi[8]相同的近似,假设所有的数据均包含在一个单位超球体中,即$R = 1.0$.即使R^2/γ^2不能用来估计测试误差,文献[7]中的几个试验也表明R^2/γ^2可以作为SVM模型选择的一个较好的性能指标.

3　基于遗传算法的SVM模型自动选择(GA-based automatic model selection for SVMs)

SVM模型自动选择本质上为一个非线性非二

次规划的非常复杂的寻优问题. GA 是由美国 Holland 教授在 1975 年提出的,是一种模拟达尔文遗传选择和自然生物进化过程的随机搜索算法. 其主要特点是不依赖于梯度信息的隐含并行随机群体搜索,能在搜索过程中自动获取和积累有关搜索空间的知识,并能自适应地控制搜索进程,从而得到全局最优解. 正是由于 GA 不太需要基于问题的知识,因此较以往传统搜索算法具有使用方便、鲁棒性强、便于并行处理等特点.

应用 GA 来解决 SVM 模型选择这一问题,必须考虑的如下问题:1)参数编码;2)初始种群的产生;3)适应度函数的确定;4)遗传操作.

3.1 实值编码策略(Real-coded scheme)

参数编码是实现 GA 的关键,采用何种方式编码是由问题本身的性质所决定的. 对于 SVM 模型选择问题,根据上述分析可以知道,其本身是一个约束优化问题. 对于约束优化问题,实值编码比二进制编码具有更高的搜索效率[11, 12]. 因此,采用实值编码策略来实现 SVM 模型参数的编码. 对于具体的高斯核函数 SVM 模型自动选择,每个染色体中的基因由 2 位十进制浮点数组成,代表一个待确定的参数 $\theta = (C,\sigma)$,将这两个浮点数级联起来就形成一条 4 位十进制浮点数构成的染色体.

3.2 初始种群的产生(Produce initial population)

采用实值编码,解空间与染色体空间重合. 考虑到种群数目过大不仅增加 GA 运算的时间,而且会使种群形态过于分散,从而使算法的收敛困难;所以我们选择种群规模 pop_size 的大小为训练样本的 30%. 在解空间中随机产生初始种群,并使其均匀分布于解空间.

3.3 适应度函数的确定(Fitness function)

GA 是对适应度函数的最大化寻优,而 SVM 模型选择是对各种泛化性能指标的最小化优化,因此,需将最小化泛化指标转化为最大化的适应度函数

$$\text{fit} = 1/(T + 0.01). \tag{11}$$

其中 T 分别为式(8)与(9)所示的直观估计与半径-间隔界.

3.4 遗传操作算子(Genetic operator)

遗传操作是实现寻优的关键,包括选择、交叉和变异操作算子. 为了提高算法的效率,采用启发式搜索策略,实现模型参数的寻优. 启发式遗传算法根据种群进化情况,动态地调整遗传算子,维持种群的多样性,并克服早熟及加快搜索速度. 在进化的前几代,采用基本的 GA,随着进化的不断进行,引入最佳个体保存法以使最优解不被交叉和变异操作所破坏,并使得算法的收敛速度加快;同时采用自适应变异概率,使得算法既避免了早熟,又使得算法具有了较佳的局部搜索能力.

1) 选择. 选择算子一般由两部分组成:复制选择算子和生存选择算子. 复制选择是指为了进行交叉操作对母体(种群)中个体进行的选择,生存选择是从进行了交叉操作之后的群体中进行的选择. 若个体 α_i 的适应度函数 $\text{fit}(a_i)$,则选中 a_i 为下一代个体的概率为

$$P(a_i) = \frac{\text{fit}(a_i)}{\sum_{j=1}^{\text{pop\_size}} \text{fit}(a_j)} \times \text{pop\_size}. \tag{12}$$

显然适应度高的个体,繁殖下一代的数目较多;而适应度较小的个体,繁殖的数目较少,甚至被淘汰.

2) 交叉. 采用算术交叉算子[11, 12] $P' = aP_1 + (1 - a)P_2$[11,12],其中,P' 为交叉后的新一代个体;P_1, P_2 为参与交叉操作的父代个体;a 为(0,1)上的常数,通常 $a = 0.5$.

3) 变异. GA 引入变异时要考虑两个问题,一是如何在初期取较大的变异算子而维持种群的多样性,防止出现早熟现象;一是当算法已接近最优解邻域时,如何使变异算子减小,确保其局部随机搜索能力,加速向最优解收敛. 本文采用如下的自适应变异概率解决上述两个问题.

$$P_m = \frac{\exp(-1.5 \times t/2)}{\text{pop\_size} \times \sqrt{L}}. \tag{13}$$

其中:t 是进化代数,L 是染色体长度.

4 仿真试验结果(Testing results)

为了最有效的利用遗传算法进行 SVM 模型自动选择,设计了如下两个仿真试验. 试验一主要是根据上面的分析,确定哪一个泛化性能指标更好;试验二对现实数据利用所提的 GA 方法实现 SVM 模型自动选择,并与 Lee 和 Lin 基于 LOO 原理提出的 LOOM 方法[6]进行比较,验证所提方法的可行性与高效性.

随机产生一组均匀分布的二维数据,大小为 1000 个样本,500 个为正样本,500 个为负样本. 随机选出 25 个正样本及 25 个负样本作为训练集,用全部数据作为测试样本. 根据试验法的经验,确定模型参数空间分别为 $C \in [1 \quad 30]$,$\sigma \in [0.1 \quad 5.0]$. 分别采用直观估计和半径-间隔界作为 SVM 模型自动选择的泛化性能指标,基于 GA 的 SVM 模型自动选择算法收敛过程分别如图 3 和 4 所示. 表 1 给出了两种指标下,采用 GA 所获得的模型参数及 SVM 的性能. 图中:f 代表适应度函数,t 代表进化代数,I 代表最佳个体.

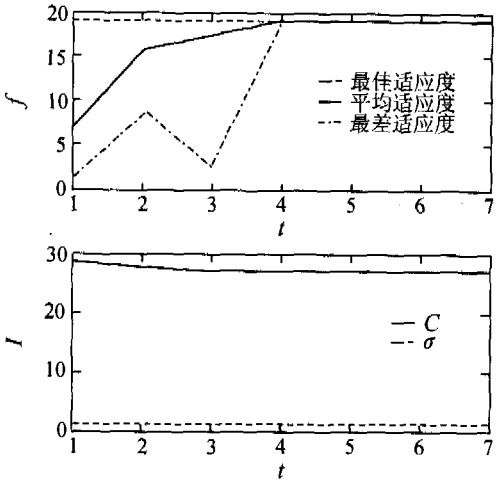


图3　直观估计泛化指标GA的收敛过程

Fig. 3　Evolution of GA with intuitional estimate

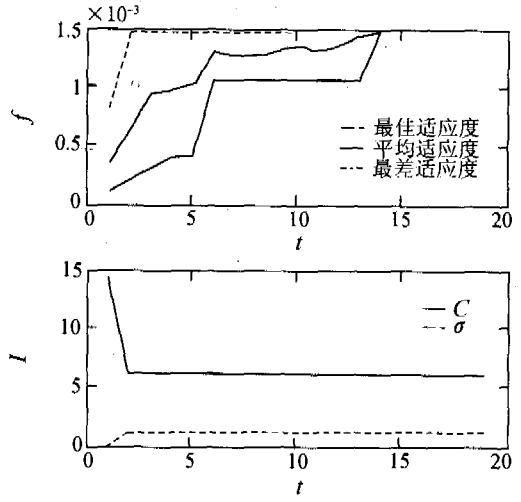


图4　半径-间隔界泛化指标GA的收敛过程

Fig. 4　Evolution of GA with Radius-margin bound estimate

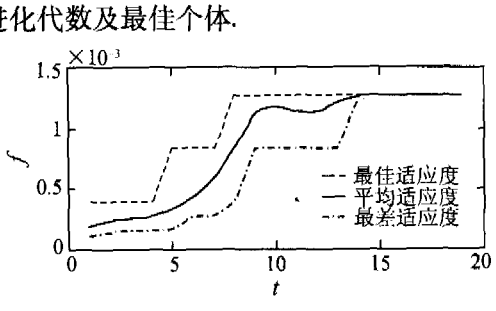
从表1可以看出,采用半径-间隔界作为泛化性能指标进行模型自动选择,所获得的SVM在满足最大间隔的条件下,支撑矢量数也较少,测试集误差也比直观估计小.所以,选取半径-间隔界作为在实际数据集上进行SVM模型自动选择的泛化性能指标.

表1　不同泛化性能指标下,GA所获得的SVM模型参数及性能

Table 1　SVM model parameters and performance by GA with various estimates

	模型参数		性能		
	C	σ	支撑矢量数	最大间隔	测试集误差率
直观估计	27	1	2	0.458590	1.0%
半径-间隔界	6	1	5	0.544090	0.8%

采用基准数据库中的Heart数集进行与常用的LOOM方法[6]进行比较试验.Heart数据是一个13维的高维数集,共有294个样本,其中负样本维188个.随机选取20%的数据作为训练集,全部数据作为测试集.根据经验,SVM模型参数的搜索空间为$C\in[1\quad 30]$,$\sigma\in[0.1\quad 5.0]$.*GA*模型选择参数收敛过程如图5所示.图中:f,t和I分别代表适应度函数、进化代数及最佳个体.



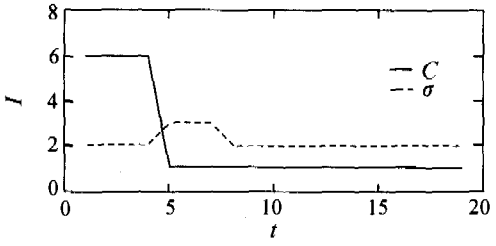


图5　Heart数集的GA模型选择参数收敛过程

Fig. 5　Model selection evolution of GA for Heart data

表2给出了与LOOM方法[6]的比较结果.LOOM方法必须采用全部数据用来求解模型参数,才能获得比较理想的结果.而在实际应用中,常常并没有大量的样本来进行模型训练,只能通过小样本(一般小于全部数据的20%)来实现模型选择,从而对大量的数据进行测试.因而,仍采用LOOM方法,但仅使用20%样本进行模型选择的结果也同时在表2中列出.

由表2可以看出,基于GA的SVM模型自动选择方法获得的SVM比标准的LOOM模型自动选择方法,对测试集的识别率有所提高.若考虑小样本机器学习的实际情况,仅采用20%的数据来进行模型选择,所提的基于GA的SVM模型自动选择方法所获得的SVM对测试集的识别率则有较大的提高.这充分证明了本文所提方法的可行性与高效性.这主要是因为LOOM方法是根据LOO法的基本思想,通过穷举搜索使得训练样本误差最小,当训练样本数据较少时(取全部数据的20%),虽然能保证训练样本集的识别率较高,但是对测试集的识别率较低.

表2 GA,LOOM所得到的Heart数集模型参数及测试集正确识别率

Table 2 Comparison results of GA with LOOM for the Heart data

模型选择算法	训练样本数	模型参数		测试集识别率
		C	σ	
GA	59	1	2	84.01%
LOOM *	294	128	0.000250	83.70%
LOOM	59	512	0.002	80.00%

注:LOOM * 的模型选择采用全部数据来实现.

5 结论(Conclusion)

在分析了SVM超参数对其性能的影响及性能估计后,提出了实值编码的启发式遗传算法(GA)实现高斯核函数SVM的模型自动选择.通过试验确定了半径-间隔界作为SVM模型自动选择的泛化性能指标.人造数据和Heart数集仿真试验结果表明,基于半径-间隔界的GA模型选择方法可以有效的实现SVM模型自动选择.其性能与Lee和Lin的LOOM方法[6]相比,测试集识别率有所提高.若考虑小样本机器学习的实际情况,与LOOM方法[6]相比,所获得的SVM对测试集的识别率有很大的提高.本文所提的基于GA的高斯核SVM模型自动选择方法,为SVM的实际应用打下了坚实的基础.

参考文献(References):

[1] http://www.clopinet.com/isabelle/projects/SVM/applist.html [OL/DB].

[2] VAPEIK V. *Statistical Learning Theory*[M]. Berlin: Springer, 1998.

[3] VAPNIK V. *The Nature of Statistical Learning Theory*[M]. New York: Spring-Verlag, 1995.

[4] CHAPELLE O, Vapnik V. Model selection for support vector machines [M]//SOLLA S A, LEEN T K, Müller K R. *Advances in Neural Information Processing Systems*. Solla Cambridge, MA: MIT Press, 2000:230 - 236.

[5] TSUDA K, RATSCH G, MIKA S, et al. Learning to Predict the Leave-One-Out Error of Kernel Based Classifiers[M]. *Lecture Notes in Computer Science*, 2001, 2130: 331 - 338.

[6] LEE J H, LIN C J. *Automatic model selection for support vector machines, technical report* [R]. Taipei, Taiwan: Department of Computer Science and Information Engineering, National Taiwan University, 2000.

[7] CHAPELLE O, VAPNIK V, BOUSQUET O, et al. Choosing multiple parameters for support vector machines[J]. *Machine Learning*, 2002, 46 (1 - 3): 131 - 159.

[8] KEERSTHI S S. Efficient tuning of SVM hyperparameters using radius/margin bound and iterative algorithms[J]. *IEEE Trans on Neural Networks*, 2002, 13 (5): 1225 - 1229.

[9] GOLD C, SOLLICH P. Model selection for support vector machine classification [J]. *Neurocomputing*, 2003, 55 (1/2): 221 - 249.

[10] 陈国良, 王煦法, 庄镇泉, 等. 遗传算法及其应用[M]. 北京:人民邮电出版社, 1996.
(CHEN Guoliang, WANG Xufa, ZHUANG Zhenquan, et al. *Genetic Algorithm and Its Application* [M]. Beijing: Posts & Telecom Press, 1996.)

[11] MICHALEWICZ Z. *Genetic Algorithms + Data Structures = Evolution Programs* [M]. 2nd ed. Berlin: Springer-Verlag, 1994.

[12] SU Y X, DUAN B Y, ZHENG C H. Genetic design of kinematically optimal fine tuning Stewart platform for large spherical radio telescope [J]. *Mechatronics*, 2001, 11 (7): 821 - 835.

作者简介:

郑春红 (1969—),女,副教授,博士,感兴趣的研究领域为智能进化计算与机器学习,E-mail:chzheng@xidian.edu.cn;

焦李成 (1959—),男,西安电子科技大学电子工程学院院长,教授,博士生导师,主要从事非线性科学和智能信号处理以及神经网络与大规模并行处理等领域的研究,E-mail:lchjiao@mail.xidian.edu;

丁爱玲 (1967—),女,长安大学副教授,西安电子科技大学在职博士研究生,研究兴趣为智能信号处理与机器学习,E-mail:cd-dingal@sohu.com.