基于遗传算法与经验误差最小化的 SVM 模型选择方法

周 欣, 许建华

(南京师范大学 计算机科学与技术学院, 江苏 南京 210097)

[摘要] 支持向量机(SVM)的推广能力依赖于核函数形式及核参数和惩罚因子的选取,即模型选择. 在分析参数对分类器识别精度的影响基础上,提出了基于遗传算法和经验误差最小化的支持向量机参数选择方法. 在 13个 UCI数据集上的实验表明了本文算法的正确性与有效性,且具有良好的推广性能.

[关键词] 支持向量机,核函数,核参数,经验误差,遗传算法

[中图分类号] TP 18 [文献标识码] A [文章编号] 1672-1292(2009)02-0065-07

SVM Model Selection Based on Genetic Algorithms and Empirical Error Minimization

Zhou Xin, Xu Jianhua

(School of Computer Sciences, Nanjing Normal University, Nanjing 210097, China)

Abstract The spreading capacity of support vector machine (SVM) depends largely on the selection of kernel function and its parameters, and penalty factor, that is model selection. Having analyzed the parameters' influence on the classifiers' recognition accuracy, we propose a new method for SVM model selection using genetic algorithm and empirical error minimization. The experiments on 13 different UCI benchmarks show its correctness, effectiveness and good spreading performance.

Key words support vector machine, kernel function, kernel parameter, empirical error, genetic algorithm

支持向量机(Support Vector Machine, SVM)是 Vapnik等学者在 1974年首先提出来的,经过上世纪 90年代众多学者的研究与发展,已经成为一类得到广泛应用的机器学习算法. 对于实际分类问题,支持向量机性能的优劣依赖于核函数形式及核参数和惩罚参数的选取. 确定 SVM的参数(如惩罚因子 C和 RBF宽度 θ),使得期望的测试误差最小,被称为 SVM的模型选择或者超参数选取. 如何根据训练样本集来选择合适的超参数,已经成为目前研究的一个热点.

最常用的模型选择方法是 *K*-折叠交叉验证法和留一法. 首先需根据人工经验确定一个近似的最优参数范围,然后在参数集上,多次重复分类器的设计与测试,遍历地搜索出最优参数,其共同缺点是计算量大. R¾tsch[1]等人依据 5-折叠交叉验证法,利用数据集前 5次剖分最优参数的平均值来设置 13个数据集的最佳参数. 由于留一法误差的计算比较困难,所以通常采用一些留一法误差上界作为模型选择的目标函数,如半径间隔界、张成界(span bound)和 VC维界等. Chapelle[2]等人用梯度下降法最小化半径间隔界和张成界来实现 SVM模型的自动选择; Keerthi[3]采用拟牛顿法对半径间隔界最小化来选择高斯核函数 SVM的超参数; Duan[4]等人用实验说明半径间隔界对 L2-SVM能找到较佳的参数,但是不适用于 L1-SVM. 有学者提出最小化经验误差来进行模型选择,在一个验证数据集上最小化经验误差估计来优化 SVM超参数. 经验误差[5]是由验证数据集上的后验概率计算得到的,即在 SVM训练之后,使用一个 Sigmoid函数将 SVM在验证数据集上的输出映射为相应的概率,该概率是一个后验概率,通过后验概率来计算验证数据

收稿日期: 2008-08-04
基金项目: 国家自然科学基金(60875001)资助项目.
通讯联系人: 许建华,教授,研究方向: 模式识别、机器学习、信号处理等. E-mail: xujianhua@njnu.edu.cn

集上的误差估计. Cherie[6,7]等人用拟牛顿法通过最小化经验误差来进行模型选择; Adankon[8]把惩罚参数放在核函数中对 L1-SVM 进行了重新定义,也用拟牛顿法最小化经验误差对 L1-SVM 进行模型选择,取得了较好的效果.

但是上述基于梯度的优化算法(如牛顿法、共轭梯度法等)都要求目标函数是可微分的,而且此类方法可能会陷于局部最优解. 正如 Keerthi[3]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. 然而,令人遗憾的是,除非作者对某一个问题具有很好的启发知识,否则,并不容易获得问题合适的初值.

近年来,有些学者开始使用不需要导数的现代优化技术(如遗传算法、模拟退火、粒子群算法)来研究 SVM 模型选择技术: Zheng[9]等用遗传算法最小化半径间隔界来选择 SVM 的参数; Acevedo[10]等用模拟退火算法和改进的半径间隔界来调整 L1-SVM 的参数; Guo[11]用模拟退火法直接最小化测试样本集的分类误差进行模型选择. 但是,由于前面的方法都是以半径间隔界作为目标函数的,而实验证明[4]半径间隔界对 L1-SVM 作模型选择的效果并不理想.

本文提出用遗传算法直接对经验误差进行最小化,建立基于遗传算法与经验误差的 SVM 模型选择方法. 在 13个 UCI数据集上的实验说明,用遗传算法通过最小化经验误差来进行模型参数的自动选择,提高了分类的正确率,为解决支持向量机的参数选取问题提供了一条有效的途径.

1　SVM 的分类原理及参数的影响

两类支持向量机(SVM)的基本理论是: 给定训练样本 $\boldsymbol{x}_i \in \mathbf{R}^n$, $i = 1, \cdots, l$, 类别 $y_i = \{1, -1\}$, 通过一个非线性变换的映射函数 $z = \Phi(x)$ 把观测数据 x 从原始空间映射到高维空间中,目的是找到一个最优超平面 $w \cdot z + b = 0$ 把两类样本分开. 通常解决 SVM 的对偶问题:

$$\begin{aligned} &\max \quad \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j \Phi(\boldsymbol{x}_i) \cdot \Phi(\boldsymbol{x}_j), \\ &\text{s.t.} \quad \sum_{i=1}^{l} \alpha_i y_i = 0, \quad 0 \leqslant \alpha_i \leqslant C, \quad \forall\, i. \end{aligned} \tag{1}$$

式中, α_i 为每个样本对应的 Lagrange乘子,非零 Lagrange乘子对应的向量为支持向量; C 为惩罚因子,实现分类器的复杂度和推广误差之间的折衷; 内积 $\Phi(\boldsymbol{x}_i) \cdot \Phi(\boldsymbol{x}_j)$ 用核函数来表示, $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \Phi(\boldsymbol{x}_i) \cdot \Phi(\boldsymbol{x}_j)$, 本文采用径向基(RBF)核函数表示:

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\sigma^2}\right) = \exp\left(-\frac{\theta}{2}\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2\right). \tag{2}$$

式中, θ 为径向基核函数的参数. 求解对偶问题(1), 可以得出相应的非线性分类器:

$$f(\boldsymbol{x}_i) = \sum_{j=1}^{l} y_j \alpha_j k(\boldsymbol{x}_j, \boldsymbol{x}_i) + b, \quad i = 1, \cdots, l. \tag{3}$$

由于支持向量机的推广性取决于核函数的选择以及支持向量机超参数的选择,核函数参数和惩罚因子的选择对 SVM 的性能至关重要. 当采用 RBF核时, SVM中的参数主要有惩罚因子 C 及核参数 θ 等. 下面以 banana数据集为例,研究参数 C 和 θ 对分类器性能的影响:

(1) 惩罚因子 C　固定 $\theta = 1$, 参数 C 在 $(0, 400)$ 之间变化,绘出训练误差和测试误差随参数 C 的变化曲线,如图 1所示. 可以看出,当 C 取值较小时,训练误差和测试误差均较大,且随 C 的增大而减小,为欠学习现象; 当 C 的取值过大时,训练误差较小,而测试误差随 C 的增大而增大,为过学习现象.

(2) 核参数 θ　固定 $C = 150$, θ 在 $(0, 10)$ 之间变化,绘出训练误差和测试误差的变化曲线,如图 2所示. 当 θ 较小时,训练误差和测试误差都很大,为欠学习现象; 当 θ 较大时,训练误差小而测试误差大,为过学习现象.

由图 1和图 2可以看出,惩罚因子 C 和高斯核函数参数 θ 对分类结果的影响都很大. 惩罚因子 C 可以实现分类器的复杂度和经验误差之间的折衷考虑. C 越大,分类器的复杂程度越大,经验误差下降; 当 C 达到一定值时,随着 C 的变化,分类器的经验误差将几乎不变. 而核参数 θ 主要影响样本数据在高维特征空间中分布的复杂程度,改变核参数 θ 则隐含改变映射函数,如何寻找一个合适核参数 θ 对于分类器的设计

来说至关重要.因此要想获得推广能力良好的 SVM 分类器,必须选择合适的超参数.只有选择合适的模型参数, SVM 的优越性才能更好的发挥出来.因此,本文利用遗传算法寻找支持向量机分类器的最佳模型参数.

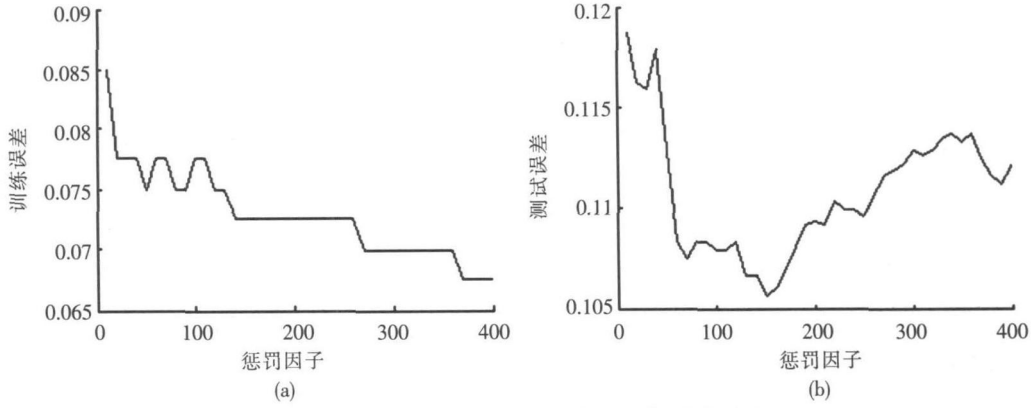


图 1　训练误差和测试误差与惩罚因子之间的关系

Fig.1　Relations of training error and testing error change with penalty factor

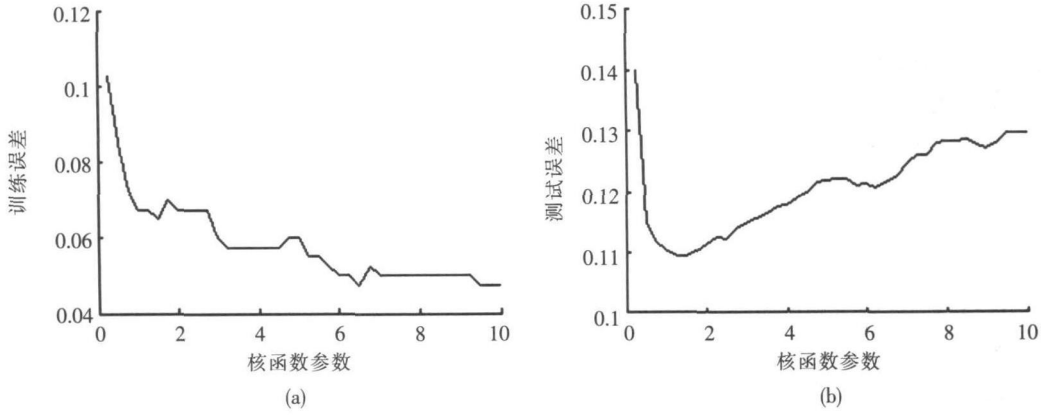


图 2　训练误差和测试误差与核参数之间的关系

Fig.2　Relations of training error and testing error change with kernel parameter

2　经验误差准则

基于最小化误差估计的思想,用一个独立的验证数据集来近似估计 SVM 的期望风险[2],则 SVM 的分类误差为:

$$T = \frac{1}{N}\sum_i \Psi(-y_i f(\boldsymbol{x}_i)). \tag{4}$$

式中, Ψ为一个阶梯函数, N 为验证数据集的大小.验证数据集越大,误差估计越可靠.

将 SVM 中的后验概率估计 $P(y=1\mid \boldsymbol{x})$ 作为误差估计,文献[12]中 Platt 给出了一个 Sigmoid 函数来近似估计后验概率 $P(y=1\mid \boldsymbol{x})$:

$$P(y=1\mid \boldsymbol{x}) = \frac{1}{1+\exp(Af(\boldsymbol{x})+B)}, \tag{5}$$

式中, $f(\boldsymbol{x})$ 是 SVM 的输出, y 是数据样本 $\boldsymbol{x}$ 的可能目标值, A 和 B 为参数.

这里要考虑参数 A 和 B 的选择.为了得到最佳的参数 A 和 B,可以用牛顿优化算法在一个独立的验证数据集上来最小化下面的问题:

$$\begin{aligned} &\min F(A,B) = -\sum_i t_i \log(P_i) + (1-t_i)\log(1-P_i), \\ &\text{s.t.}\quad t_i = \frac{(y_i+1)}{2}. \end{aligned} \tag{6}$$

通过验证数据点被最小化, 可得到 A 和 B 的最优值. 其中 $P_i = \frac{1}{1+\exp(Af_i + B)}$ 是所估计的后验概率. 定义 $t_i = \frac{(y_i + 1)}{2}$, 当输入向量 $\boldsymbol{x}_i$ 属于正类样本时, $t_i = 1$; 当输入向量 $\boldsymbol{x}_i$ 属于负类样本时, $t_i = 0$.

对于观察所给数据样本 $\boldsymbol{x}_i$ 的目标值的误差估计 (即错分概率估计) 可以表示为:

$$E_i = P(y_i \neq z_i) = |t_i - P_i| = P_i^{1-t_i}(1-P_i)^{t_i}. \tag{7}$$

式中, $z_i = \mathrm{sign}(f_i)$, $f_i = f(\boldsymbol{x}_i)$ 是相应的 SVM 输出值, P_i 是被估计的后验概率. 对于大小为 N 的数据集, 平均误差估计如下:

$$E = \frac{1}{N}\sum_{i=1}^{N}E_i = \frac{1}{N}\sum_{i=1}^{N}P_i^{1-t_i}(1-P_i)^{t_i}. \tag{8}$$

可以通过最小化平均误差估计来优化 SVM 超参数.

3　基于 *GA* 的 *SVM* 模型选择

遗传算法是模拟生物进化过程中的自然选择和遗传变异的一种随机优化方法, 结合了适者生存和随机信息交换的思想, 通过自然选择、交换、变异等作用机制, 实现种群的进化. 在寻优过程中, GA 直接以目标函数值作为搜索信息, 在解空间随机产生多个起始点并同时开始搜索, 由适应度函数来指导搜索方向, 它具有很强的全局搜索能力, 能够在复杂搜索空间快速寻求全局最优解.

3 1　适应度函数

根据本文的具体情况, 设计的适应度函数为

$$\mathrm{Fit}(C, \sigma^2) = \max - \frac{1}{N}\sum_{i=1}^{N}E_i. \tag{9}$$

式中, N 为验证数据集的样本个数; E_i 为第 i 个样本后验错分概率估计值, 由于 $E_i \in [0, 1]$, 所以算法中 max取值为 1. 可见当 SVM 在验证数据集上的经验误差越小, 对应该组参数的染色体适应度值越大.

3 2　选择操作

本文采用基于排序的适应度分派原则. 首先按照适应度值对种群内的个体进行排序, 然后按下式确定第 i 个个体被选择的概率 P_i: 首先利用按比例的适应度分配法来确定种群中每个个体被选择的概率. 若种群大小为 M, 第 i 个个体的适应度为 f_i, 则其被选中的概率 P_i 表示为:

$$P_i = \frac{f_i}{\sum_{i=1}^{M}f_i}. \tag{10}$$

适应度越高, 被选中的概率越大. 然后以概率 P_i 采取轮盘赌方法选择作为父代的个体进行下一步的交叉、变异.

3 3　交叉操作

采用线性组合的交叉操作方式. 例如以某一概率对某两个染色体 x_1、x_2 进行交叉操作时, 可以采用如下方式:

$$\begin{aligned} x_1 &= \alpha x_1 + (1-\alpha)x_2, \\ x_2 &= (1-\alpha)x_1 + \alpha x_2, \end{aligned} \tag{11}$$

式中, α 为 0, 1之间随机数.

3 4　变异操作

在将变异的染色体中随机选择一个变异位 j, 把它设置为一个归一化的随机数 $U(a_i, b_i)$. a_i、b_i 为对应该变异位的上下限:

$$x_j = \begin{cases} U(a_i, b_i) & \text{if } i = j, \\ x_i & \text{otherwise.} \end{cases} \tag{12}$$

用遗传算法通过最小化经验误差来调整 SVM 超参数的算法步骤如下:

(1) 初始化种群: 在给定的范围内 ($C \in [0, 10]$, $\theta \in [0, 1]$) 随机产生初始种群 $X(0) = \{x_1(0),$

$x_2(0)$, …, $x_n(0)$} 作为父代种群, 令 $k = 0$;

(2) 计算种群 $X(k)$ 中每一个个体的适应度函数值:

①对每组固定的参数 C 和 θ, 用训练样本集训练 SVM 分类模型;

②用前面得到的 SVM 模型对验证集进行预测分类, 得到一组预测输出值 f_i;

③由前面的预测输出值 f_i, 用牛顿优化过程来计算 A 和 B 的最佳值;

④计算后验概率 $P_i = \dfrac{1}{1 + \exp(A \cdot f_i + B)}$, 得到误差估计 $E_i = P_i^{1-t_i}(1 - P_i)^{t_i}$ (即错分概率估计);

⑤用平均误差估计作为适应度函数;

(3) 判别停止准则是否满足, 若满足则转第 7步;

(4) $k = k + 1$;

(5) 应用选择算子在 $X(k-1)$ 中选择 $X(k)$;

(6) 对 $X(k)$ 进行交叉、变异操作后转第 2步;

(7) 输出最佳的惩罚系数 C 和核参数 θ.

循环结束时, 最佳个体对应的参数即为最佳超参数. 用得到的最佳参数再调用 SVM 训练算法, 对整个训练样本集进行训练, 用训练得出的模型对测试样本集进行测试, 得出测试集的分类正确率及分类误差.

4　实验及分析

将 13个由 Ratsch 收集的 UCI基准数据集[13]: Banana, Breast cancer, Diabetes, Flare-Solar, German, Heart, Image, Ringnorm, Splice, Thyroid, Titanic, Twonorm, Waveform, 每个数据集包含了 100(或 20)个不同的剖分. 本文仅对前 5组和前 10组进行模型选择, 然后取平均值. 表 1给出了 13个数据集的样本特征维数、训练和测试样本数目及每个数据集的剖分次数.

表 1　UCI 13个数据集的数据特征

Table 1　General information about 13 UCI datasets

数据集	训练集大小	测试集大小	样本维数	剖分次数
Banana	400	4 900	2	100
Breast cancer	200	77	9	100
Diabetes	468	300	8	100
Flare- Solar	666	400	9	100
German	700	300	20	100
Heart	170	100	13	100
Image	1 300	1 010	18	20
Ringnorm	400	7 000	20	100
Splice	1 000	2 175	60	20
Thyroid	140	75	5	100
Titanic	150	2 051	3	100
Twonorm	400	7 000	20	100
Waveform	400	4 600	21	100

下面以在 Breast cancer数据集上的实验结果为例, 来说明此算法找到的是全局最优解. 本算法中的各控制参数设置为: 种群大小为 100, 最大进化代数为 60, 交叉概率为 0.8, 变异概率为 0.1. 算法采用浮点数编码方式, 以避免二进制编码方式在遗传操作时进行反复译码、编码的操作; 此外, 可以克服二进制字符串的有限长度的影响, 从而提高进化算法的性能和求解精度. 通过初步的实验说明, 参数 θ 在 [0, 1] 之间, 而 C 的值在 [0, 10] 之间实验能取得较好的效果, 算法同时对参数 C 和 θ 进行优化.

首先用本文的算法分别对 Breast cancer数据集中的两个剖分做模型选择, 来验证本文提出算法的有效性. 图 3中的两个图分别为 SVM 在其中两个 Breast cancer剖分上用网格扫描法得到的验证集的分类误差, 网格扫描时 C 的步长为 0.2, θ 的步长为 0.01. 对应两个剖分用本文算法模型选择的结果分别为 $C = 2.076\,7$, $\theta = 0.075\,4$ 和 $C = 0.429\,3$, $\theta = 0.130\,1$, 在图中用“★”来标识. 从图上可以看出模型选择的结果基本上对应于用网格扫描法得到的分类错误率的最低点, 即我们找到了验证集上对应的最佳参数.

为了将本算法和 5-折叠法、半径间隔界及张成界做模型选择方法作比较, 首先对 5个数据集 Breast

cancer, Diabetes, Heart, Thyroid, Titanic来做测试. 用文献[1]和[2]中同样的方法进行实验, 选取每个数据集的前 5个训练集, 随机抽取 1/3的数据作为验证集, 余下的 2/3数据作为模型选择的训练集, 用本文的模型选择算法来优化惩罚因子 C和核参数 θ. 先对每个数据集的前 5个和前 10个剖分做模型选择, 分别计算选择出的 5组和 10组参数的平均数作为最终的参数, 然后最终的参数对训练数据集的 100个剖分分别进行训练, 再对测试数据集测试得到测试集的分类误差, 最后求得 100个剖分平均值和标准差, 实验结果如表 2所示.

表 2　用不同方法选择 SVM超参数得到的测试误差及对应模型选择的参数(C和θ)

Table 2　Testing error and the model parameters (C and θ) by different algorithms for SVM model selection

数据集	5-折叠法[1]	半径间隔界[2]	张成界[2]	Mathias方法[8]	GA_SVM 前 5组平均	GA_SVM 前 10组平均
Breast cancer	26.04±4.7 [15.19, 50]	26.84±4.71	25.59±4.18	25.79±4.21 [1.080, 0.100]	25.55±4.39 [0.5129, 0.0907]	25.62±4.49 [0.6539, 0.0583]
Diabetes	23.53±1.73 [-, 20]	23.25±1.70	23.19±1.67	23.25±1.82 [0.500, 0.06]	23.35±1.63 [1.6131, 0.0339]	23.32±1.61 [1.4576, 0.0357]
Heart	15.95±3.26 [3.162, 120]	15.92±3.18	16.13±3.11	15.98±3.32 [0.666, 0.010]	15.57±3.18 [0.6456, 0.0185]	15.48±3.33 [0.4284, 0.0209]
Thyroid	4.80±2.19 [10, 3]	4.62±2.03	4.56±1.97	4.64±2.13 [10.00, 0.183]	4.77±2.28 [0.5632, 0.2704]	4.57±2.17 [0.5576, 0.3268]
Titanic	22.42±1.02 [100 000, 2]	22.88±1.23	22.5±0.88	22.93±1.17 [1.100, 0.1197]	22.59±1.01 [3.2746, 0.3241]	22.57±0.89 [2.8841, 0.2266]

由表 2的实验结果可以看出, 本文的结果同 5-折叠交叉验证法的方法很相似. 文献[8]中用拟牛顿法来最小化经验误差, 每次迭代过程中, 都需要对目标函数进行求导. 而本文的算法直接对目标函数最小化, 从结果可以看出, 测试集的识别率比拟牛顿法有所提高. 即基于遗传算法与经验误差的模型选择方法比 5-折叠验证法、半径间隔法等具有更好的分类性能.

由于半径间隔法、张成界和 Mathias方法没有给出在余下 8个数据集上的测试结果, 故用本文的方法只和 5-折叠交叉验证法的测试结果相比较. 由表 2的结果可以看出, 对数据集的前 10个剖分比用前 5个剖分做模型选择的分类误差低, 因此以下实验对 UCI数据库中的余下的 8个数据集分别选取前 10个剖分做模型选择, 得出 10组参数的平均数作为最终的参数. 然后用上面同样的方法, 对训练数据集的 100个剖分分别进行训练, 再对测试数据集测试得到测试集的分类误差. 表 3中给出了本文的方法与 5-折叠交叉验证法在 13个数据集上的比较结果.

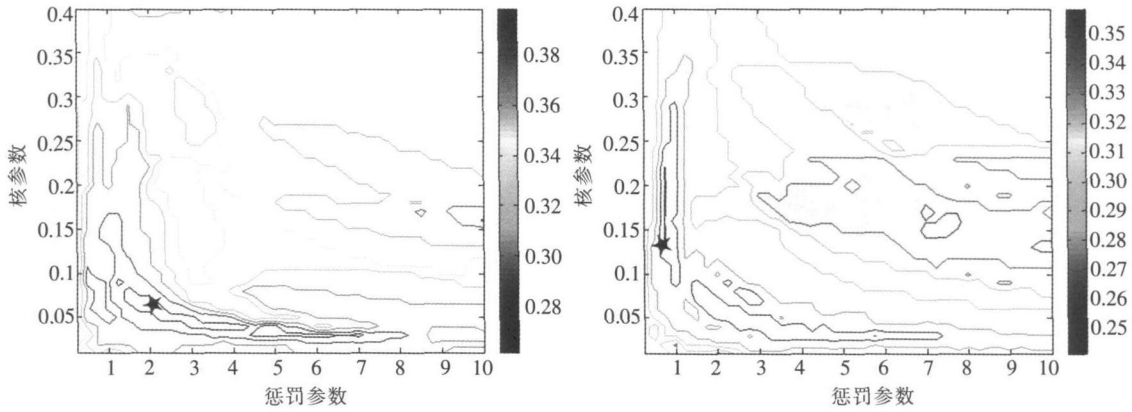
表 3　本文方法与交叉验证法的结果在 13个数据集上的测试误差及对应参数(C和θ)

Table 3　Testing error and the model parameter (C and θ) by GA-SVM and 5-fold cross validation on 13 UCI datasets

数据集	5-折叠法	GA_SVM(前 10组平均)
Banana	11.53±0.66 [316.2, 1]	11.34±0.49 [3.2708, 0.3273]
Breast cancer	26.04±4.7 [15.19, 50]	25.62±4.49 [0.6539, 0.0583]
Diabetes	23.53±1.73 [-, 20]	23.32±1.61 [1.4576, 0.0357]
Flare-Solar	32.43±1.82 [1.023, 3]	32.45±1.80 [1.4496, 0.0433]
German	23.61±2.07 [3.162, 55]	23.60±2.23 [1.2585, 0.0337]
Heart	15.95±3.26 [3.162, 120]	15.48±3.33 [0.4284, 0.0209]
Image	2.96±0.60 [500, 30]	3.56±0.67 [3.2369, 0.3273]
Ringnorm	1.66±0.12 [1000000000, 10]	1.47±0.08 [0.2231, 0.0769]
Splice	10.88±0.66 [1000, 70]	10.83±0.67 [2.5279, 0.0155]
Thyroid	4.80±2.19 [10, 3]	4.57±2.17 [0.5576, 0.3268]
Titanic	22.42±1.02 [100000, 2]	22.57±0.89 [2.8841, 0.2266]
Twonorm	2.96±0.23 [3.162, 40]	2.43±0.13 [0.1763, 0.0404]
Waveform	9.88±0.43 [1, 20]	9.86±0.48 [0.6228, 0.0424]

由表 3的实验结果可以看出, 本文的方法在 13个数据集上的结果优于 5-折叠交叉验证法的结果. 为了同用遗传算法最小化半径间隔界来选择 SVM的参数相比较, 本文给出文献[9]中用遗传算法对 heart数据集实现参数的自动选择结果, 如表 4所示.

由表 4可以看出本文算法测试集的分类正确率比用遗传算法最小化半径间隔界模型选择有所提高, 即本文直接最小化验证数据集的经验误差的算法能给出较好的结果.



(a) C=2.076 7,θ=0.075 4　　(b) C=0.429 3,θ=0.130 1

图 3　网格扫描法得到的验证集分类错误率

Fig.3　Classification error rate on validation set of grip scan method

5　结　语

本文在分析了 SVM 超参数对其性能的影响及性能估计后,把遗传算法与支持向量机算法相结合,提出了用遗传算法通过最小化经验误差来实现高斯核函数 SVM 的模型自动选择.用该方法选择出的参数与用网格扫描法得到的最佳参数基本吻合,通过 13 个 UCI 数据集的实验表明基于遗传算法与最小化经验误差的模型选择方法可以实现 SVM 模型自动选择.与传统的基于梯度的算法相比,不需要对目标函数进行求导,直接对目标函数最小化;与交叉验证法相比,测试集的识别率有所提高.本文提出的基于遗传算法与经验误差最小化的 SVM 模型选择方法是一种较好的模型选择方法,这种方法也可以应用于其他类型支持向量机的模型选择,具有一定的推广价值.

表 4　文献用遗传算法实现参数选择与本文方法在 heart 数据集上测试结果的比较结果

Table 4　The comparison results of testing error from GA-SVM and results of [9] on heart clatasets

方法	参数 C	参数 θ	测试集分类正确率
文[9]的方法	1	2	84.01%
本文方法	0.428 38	0.020 88	84.52%

[参考文献] (References)

[1] Ratsch G, Onoda T, Muller K R. Soft margins for AdaBoost [J]. Machine Learning, 2001, 42: 287-320.

[2] Chapelle O, Vapnik V, Bousquet O, et al. Choosing multiple parameters for support vector machines [J]. Machine Learning, 2002, 46: 131-159.

[3] Keerthi S S. Efficient tuning of SVM hyperparameters using radius margin bound and iterative algorithms [J]. IEEE Transactions on Neural Networks, 2002, 13: 1 225-1 229.

[4] Duan K, Keerthi S S, Poo A N. Evaluation of simple performance measures for tuning SVM hyperparameters [J]. Neurocomputing, 2003, 51: 41-59.

[5] Ayat N E, Cheriet M, Suen C Y. Optimization of the SVM kernels using an empirical error minimization scheme [C]//Lee S W, Verri A. Pattern Recognition with Support Vector Machines. Berlin Heidelberg: Springer, 2002, 2388: 354-369.

[6] Adankon M M, Cheriet M, Ayat N E. Optimizing resources in model selection for support vector machines [C]//2005 International Joint Conference on Neural Networks. Canada, Montreal, 2005: 925-930.

[7] Ayat N E, Cheriet M, Suen C Y. Automatic model selection for the optimization of the SVM kernels [J]. Pattern Recognition, 2005, 38: 1 733-1 745.

[8] Adankon M M, Cheriet M. New formulation of SVM for model selection [C]//2006 International Joint Conference on Neural Networks. Canada, Vancouver: IEEE Press, 2006: 1 900-1 907.

[9] Zheng C H, Li C J. Automatic parameters selection for SVM based on GA [C]//5th World Congress on Intelligent Control and Automation. Hangzhou, China: IEEE Press, 2004: 1 869-1 872.

[10] Javier A, Saturnino M, Philip S. Tuning L1-SVM hyper-parameters with modified radius margin bounds and simulated annealing [C]//Computational and Ambient Intelligence. Berlin Heidelberg: Springer-Verlag, 2007, 4507: 284-291.

[11] Guo X C, Liang Y C, Wu C G, et al. PSO-based hyper-parameters selection for LS-SVM classifiers [C]//Neural Information Processing. Hongkong, China: IEEE Press, 2006, 4233: 1 138-1 147.

[12] Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods [C]//Bartlett P J, Scholkopf B, Schuurmans D. Advances in large margin classifiers. Cambridge MA: MIT Press, 1999: 67-74.

[13] Ratsch G. Benchmark data sets[EB/OL]. http://ida.first.fhg.de/projects/bench/benchmarks.htm. 1999/2003-7.

[责任编辑:严海琳]