文章编号：1001-0920(2010)03-0367-04

一种基于 PSO 的 RBF-SVM 模型优化新方法

徐海龙，王晓丹，廖 勇，张宏达，蒋玉娇
（空军工程大学 导弹学院，陕西 三原 713800）

摘 要：针对使用径向基核函数的支持向量机，采用粒子群优化方法实现模型优化. 基于训练集中样本之间的最近平均距离和最远平均距离，给出参数 σ 的取值空间，从而减小了超参数搜索的范围，并采用对数刻度进一步提高粒子群优化方法的参数搜索效率. 与遗传算法和网格法的对比实验表明，所提出的方法收敛速度更快，得出的超参数更优.

关键词：模型优化；支持向量机；粒子群优化；搜索效率
中图分类号：TP181 **文献标识码**：A

New approach for optimizing model of RBF-SVM based on PSO

XU Hai-long, *WANG Xiao-dan*, *LIAO Yong*, *ZHANG Hong-da*, *JIANG Yu-jiao*
(Institute of Missile, Air Force Engineering University, Sanyuan 713800, China. Correspondent: XU Hai-long, E-mail: xhl_81329@163.com)

Abstract: For the radial basis function (RBF) kernel based support vector machines (SVM), particle swarm optimization (PSO) is employed to carry out the model optimization. The value space of the parameter σ is presented on the analysis of the mean shortest distance and mean furthest distance among the samples of the training set, thus the search region is reduced, and logarithmic scale is employed to further improve the search efficiency of PSO. Extensive experimental results on comparison with genetic algorithm and grid based approaches show that the proposed approach converges faster and produces better hyper-parameters.

Key words: Model optimization; Support vector machine; Particle swarm optimization; Search efficiency

1 引 言

支持向量机(SVM)模型包括核函数、超参数的设置，合理设置模型才能使 SVM 发挥最佳的泛化性能. 研究者已针对 SVM 模型的选择问题进行了大量的研究，可以归为两类：1) 根据影响分类器性能的各种因素构造优化目标函数，基于梯度方法并通过迭代法得出最优超参数[1-3]；2) 在超参数空间中用各种搜索方法，按照测试误差在搜索空间中进行参数优选，如网格搜索法及其改进算法[4]以及各种启发式搜索算法[5,6]. 第 1 类方法的优点在于不用训练 SVM 本身，速度较快，但容易陷入局部最优解，特别是若迭代初始值选择不当，则对最终结果影响更大[7]. 第 2 类方法直接以 SVM 的测试精度为指标，对每一组超参数都需要训练并测试一个 SVM. 采用网格法时，若网格密度较小，则网格点覆盖最优超参数的几率很小，将网格密度增加能提高找到最优参数组合的概率，但会导致算法复杂度迅速增高.

SVM 模型自动选择本质上为一个非线性非二次规划的复杂寻优问题，采用遗传算法(GA)能够有效解决此问题[5]，但是收敛速度较慢. 与 GA 相比，粒子群优化(PSO)算法具有执行速度快、受问题维数变化影响小等优点，已有研究者将 PSO 与 SVM 相结合用于特征子集的选择[9]. 对于任何搜索算法，设置合理的搜索范围是保证算法快速得到最优参数组合的重要条件，若范围太大，则会增加不必要的寻优操作；若范围太小，则可能不包括最优参数组合. 对此，首先根据样本的空间分布信息与核函数的意义进行参数空间预估计；然后在这个参数空间中采用对数刻度应用 PSO 方法实现参数寻优. 本文将此算法称为基于参数区间预估与对数刻度的 PSO (PAL-PSO). 与 GA 和网格法的对比实验表明了本文方法的有效性. 考虑 RBF 核函数的广泛应用，本文仅针对 RBF 核进行研究.

收稿日期：2009-03-24；**修回日期**：2009-06-11.
基金项目：国家自然科学基金项目(60975026)；陕西省自然科学研究计划项目(2007F19).
作者简介：徐海龙(1981—)，男，陕西韩城人，博士生，从事智能信息处理、支持向量机的研究；王晓丹(1966—)，女，陕西汉中人，教授，博士生导师，从事智能信息处理、雷达目标识别等研究.

2 基本概念

2.1 SVM

设训练样本集为 $(x_i,y_i),i=1,2,\cdots,l,l$ 为训练样本个数,$x_i\in R^d$ 为训练样本,$y_i\in\{1,-1\}$ 为输入样本 x_i 的类别,标准支持向量机(C-SVM)的优化形式为

$$\min J(w,b,\xi)=\frac{1}{2}\|w\|^2+C\sum_{i=1}^{l}\xi_i;$$
$$\text{s.t. } y_i[w^{\mathrm{T}}x_i+b]\geqslant 1-\xi_i,$$
$$\xi_i\geqslant 0,\ i=1,2,\cdots,l. \tag{1}$$

其中 $w\in R^d$ 为分类超平面的法向量. 通过 Lagrange 函数求偏导,回代到式(1),可得等价的对偶形式为

$$\max_{\alpha}L(\alpha)=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j\langle x_i,x_j\rangle;$$
$$\text{s.t. }\sum_{i=1}^{l}y_i\alpha_i=0,$$
$$0\leqslant\alpha_i\leqslant C,\ i=1,2,\cdots,l. \tag{2}$$

引入核函数映射 $K(x_i,x_j)=\langle\varphi(x_i),\varphi(x_j)\rangle$,则式(2)变为

$$\max_{\alpha}L(\alpha)=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j K(x_i,x_j).$$

对于 RBF 核,有 $K(x_i,x_j)=\mathrm{e}^{-\|x_i-x_j\|^2/2\sigma^2}$.

2.2 粒子群算法

Kennedy 等[8] 受鸟群捕食行为的启发,提出了 PSO 算法. PSO 算法描述为:在 D 维目标搜索空间中,有 m 个粒子组成一个群体,每个粒子由 D 维向量表示,其状态由 3 个向量表示. $\boldsymbol{x}_i=(\boldsymbol{x}_{i1},\cdots,\boldsymbol{x}_{iD})^{\mathrm{T}},i=1,2,\cdots,m$,表示在搜索空间中的当前位置,将 $\boldsymbol{x}_i$ 代入给定目标函数衡量该位置的优劣;$v_i=(v_{i1},\cdots,v_{iD})$ 表示第 i 个粒子的飞翔速度;$p_i(p_{i1},\cdots,p_{iD})^{\mathrm{T}}$ 表示第 i 个粒子迄今为止搜索到的最优位置. 另外,用 $p_g=(p_{g1},\cdots,p_{gD})^{\mathrm{T}}$ 表示整个粒子群迄今为止搜索到的最优位置.

PSO 对粒子进行如下操作:

$$v_{id}^{n+1}=wv_{id}^{n}+c_1r_1^{n}(p_{id}^{n}-x_{id}^{n})+c_2r_2^{n}(p_{gd}^{n}-x_{id}^{n}), \tag{3}$$
$$\boldsymbol{x}_{id}^{n+1}=\boldsymbol{x}_{id}^{n}+v_{id}^{n+1}. \tag{4}$$

其中:c_1 和 c_2 为加速系数,w 为惯性系数,r_1 和 r_2 为 $[0,1]$ 间的随机数. 为确保算法的稳定性和收敛速度,采用文献[10]的参考系数设置:$w=0.7279$,$c_1=c_2=1.49445$.

3 基于 PAL-PSO 的 SVM 模型优化

3.1 参数区间预估计

基于 RBF 核的 SVM 需要优化的参数包括因子 C 和径向基半径 σ. 核函数也可以理解为样本对其附近样本的覆盖,所以 σ 的选择应当参考样本间的距离. 考虑所有训练样本

$$k_1d_{\min}=\sigma_{lb}<\sigma<\sigma_{ub}=k_2d_{\max}. \tag{5}$$

其中:$d_{\min}=\frac{1}{l}\sum_{i=1}^{l}d_i^{\min}$ 和 $d_{\max}=\frac{1}{l}\sum_{i=1}^{l}d_i^{\max}$ 分别表示训练样本的平均最近距离和平均最远距离;$d_i^{\min}=\min\limits_{j,j\neq i}d_{ij}$ 和 $d_i^{\max}=\max\limits_{j,j\neq i}d_{ij}$ 表示每个样本的最近样本距和最远样本距,这里 $d_{ij}=d(x_i,x_j)=\|x_i-x_j\|_2$ 为欧式距离. 下面讨论因子 k_1 和 k_2 的设置.

(1) 确定下界 σ_{lb}

概率论中将出现概率小于 0.003 的事件称作小概率事件. 借用这个概念,认为小于这个数字的核函数值说明当前样本对其相邻样本的影响很微小,此时,该 SV 难以覆盖距离其最近的样本,换言之,该 SV 所覆盖的只有它自己. 所以径向基核函数的半径不宜小于此值,否则易造成过适应. 令 $K(x_i,x_j)$ 为 0.003,可得 $\sigma=0.2934d_{\min}$.

(2) 确定上界 σ_{ub}

向量应覆盖距离其较近的向量,即对距离其较近的向量影响大,对于距离其较远的向量影响小. 较大的 σ 覆盖范围大,但是至少该核函数对距离最远的向量能够体现出这种影响的衰减. 设定衰减系数 r,则有 $\mathrm{e}^{-d_{\min}^2/2\sigma^2}-\mathrm{e}^{-d_{\max}^2/2\sigma^2}<r$. 令 $r=0.003$,并假设 $d_{\min}\ll d_{\max}$,则 $\sigma_{ub}=12.9d_{\max}$.

根据上面的分析,本文近似取 $k_1=0.3,k_2=13$,代入式(5),即可确定 $\sigma\in(\sigma_{lb},\sigma_{ub})$. 参数 C 的范围不易确定,按照经验预定为 $C\in[C_{lb},C_{ub}]=[1,5000]$.

假设训练样本规模为 n,则计算所有训练样本之间的距离复杂度为 $O(n^2)$,对于每个样本计算其 $d_{\min}$ 和 $d_{\max}$ 的复杂度分别为 $O(n)$,因此参数区间预估计的时间复杂度分别为 $O(n^2)$. 相比整个参数优化过程,参数区间预估的时间耗费占的比例很小,若采用随机采样部分训练样本进行数区间预估,能进一步压缩其时间耗费.

3.2 对数刻度

对于SVM模型选择问题,分类器性能受超参数的影响有如下特点:在最优参数附近,对于同样大小的参数变化,若最优参数值较大,则目标函数受到的影响较小;若最优参数为一个较小的值,则目标函数将受到很大影响. 为综合考虑算法的开拓能力和开掘能力[4],对参数 σ 和 C 采用对数刻度.

参数区间设置为

$$\sigma'=\lg(\sigma)\in[\lg(\sigma_{lb}),\lg(\sigma_{ub})]=[\sigma'_{lb},\sigma'_{ub}], \tag{6}$$
$$C'=C\in[\lg(C_{lb}),\lg(C_{ub})]=[C'_{ub},C'_{lb}]. \tag{7}$$

3.3 PAL-PSO 优化问题设置

(1) 编码

采用连续值编码，超参数 σ 和 v 分别用一个实数表示. 种群规模为 m，$x_i=(x_{i1},x_{i2})$，$i=1,2,\cdots,m$.

(2) 初始化

随机生成粒子种群和粒子速度，设置粒子最大速度为

$$v_1^{ub}=0.2(\sigma'_{ub}-\sigma'_{lb}),\ v_2^{ub}=0.2(C'_{ub}-C'_{lb}). \tag{8}$$

(3) 目标函数

最小化测试分类误差

$$\min_{\sigma,v}\sum_{i=1}^{l_2} I(\varphi(x_i)\neq y_i). \tag{9}$$

其中：$\varphi(\cdot)$ 表示由给定参数与训练样本得出的 SVM 分类器，$x_i\in \mathrm{tst}X$ 为测试样本，y_i 为测试样本类别.

3.4 算法复杂度分析

在参数寻优问题中，对于每一组参数都需要训练一个 SVM. 采用 PSO 算法和 GA 算法优化 SVM 超参数时，计算每个个体的适应度都需要训练并测试一个 SVM. 若训练集为 $D_1=\{x_i,y_i\}_{i=1}^{l}$，$x_i\in R^d$，测试集为 $D_2=\{x_i,y_i\}_{i=1}^{l_2}$，则得出的支持向量数为 n_s，非线性 SVM 的训练和测试时间复杂度可记为 $O(dl^{\alpha})$ 和 $O(dl_2,n_s)$，通常有 $\alpha>2$. 而 PSO 和 GA 算法更新每个个体的复杂度分别为 $O(n_v)$ 和 $O(n_c)$，n_v 和 n_c 分别为 PSO 粒子维数和 GA 的编码长度. 显然有

$$\lim_{l\to\infty}\frac{\max(n_v,n_c)}{O(dl^{\alpha})}=0.$$

因此若迭代次数为 T，则整个算法的时间复杂度为 $O(Tn_pd(l^{\alpha}+l_2n_s))$，可见对于 PSO 和 GA，当种群规模 n_p 确定时，算法的时间效率是由迭代次数决定的.

4 实验与分析

为验证 PAL-PSO 算法对于优化 SVM 模型的有效性，用 GA 和网格法作为对比算法，采用 4 组 UCI 数据进行分类性能对比分析. 实验机器配置为 2048 M 内存，Intel Core2 Duo CPU(2.0 G)，运行环境为 Windows XP，算法用 Matlab2008b 编程实现.

4.1 实验设置

根据第 3 部分的描述确定 PAL-PSO 算法参数搜索范围和编码方式，进化代数为 $T=20$，种群规模为 $n_p=20$.

为了对比，利用 GA[5] 时，根据经验[4] 设置 $\sigma\in[0.1,100]$，采用 28 位二进制编码，其中参数 σ 占 16 位，C 占 12 位，进化代数与种群规模同 PAL-PSO；利用网格法时，参数范围同 GA 算法，采用 20×20 的等分参数网格，即每一种算法运行一次都需要训练 400 个 SVM.

实验用到 4 个 UCI 数据集，数据集设置见表 1. PAL-PSO 和 GA 在 4 个数据集上分别进行 3 次实验. 因为前 3 个数据集中训练集与测试集都是确定的，网格法的结果也是确定的，所以网格法对前 3 个数据集只需进行一次实验. 对于 yeast 数据集，采用网格法进行了 3 次实验. 因为这里用到的 yeast 数据集是先采用随机采样，然后去除其中的重复样本，再将样本集中前 200 个作为训练样本，所以每次实验中的训练集与测试集是不同的，并且测试集的规模也不相同.

表 1　实验数据设置

Data Sets	类别处理		数据集规模	
	positive	negative	Train set	Test set
heart disease	label = 0	other	150	153
hepatitis	label = 1	other	100	54
balance	label = 1	other	200	327
yeast	label = 1,2	label = 3,4	200	250-300

4.2 实验结果与分析

下面根据最终优化得到的 SVM 测试误差与搜索算法的时间花费对比分析算法有效性，见表 2.

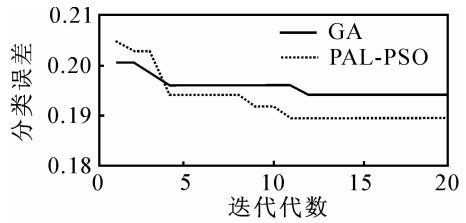
表 2　PAL-PSO、GA 与网格法的参数搜索性能对比

搜索算法	数据集/(n_b/n_w)			
	heart disease	hepatitis	balance	yeast
PAL-PSO	3/0	3/0	3/0	3/0
GA	1/2	3/0	0/1	1/2
Grid	0/0	1/0	0/0	0/1

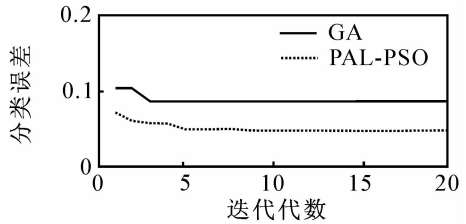
对每个数据集进行的 3 次实验中，n_b 表示一种算法优于其他算法的次数(以测试误差为指标)，n_w 表示一种算法差于其他算法的次数. 网格法仅在其中一个数据集上搜索到最佳参数组合，GA 在 3 个数据集上搜索到最佳参数组合，而 PAL-PSO 方法在所有数据集上搜索到最佳参数组合. 这组结果说明，采用 GA 与 PAL-PSO 进行参数搜索优于网格法. 同时从表 2 中也可以看出，GA 方法不稳定，12 次实验中共 5 次得到最差结果. 这是由于 GA 方法收敛较慢，而实验中为了考查搜索算法的时效，采用的迭代代数较小.(这里最佳组合与最差结果均限于本文得出的实验结果.)

图 1 显示了 PAL-PSO 与 GA 方法的性能对比. 相比 GA，PAL-PSO 在第 10 代和第 20 代得到的分类误差分别降低 9.62% 和 9.03%，在第 5 代、第 10 代时得到的最低分类误差均值分别比 GA 运行 20 代得到的分类误差均值降低 5.67% 和

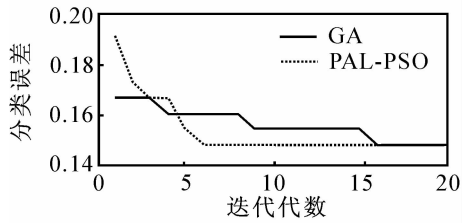
8.00%,这充分说明了采用 PAL-PSO 方法搜索最优参数组合的早期收敛性能优越.即在相同的分类精度要求下,运行 PAL-PSO 需要更少的迭代代数.根据本文实验,PAL-PSO 方法运行 10 代能得出稳定优于 GA 运行 20 代的结果.即在同样的精度要求下,相比 GA 算法,采用 PAL-PSO 只需要一半的迭代次数.因此可以说 PAL-PSO 方法的效率比 GA 算法高一倍以上.



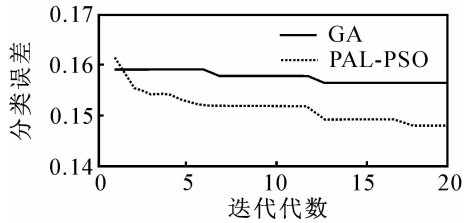
(a) heart disease数据集



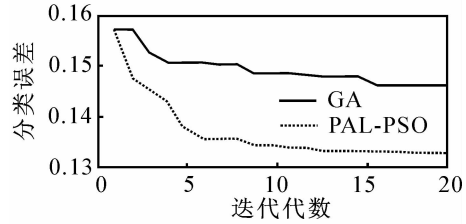
(b) hepatitis数据集



(c) balance数据集



(d) yeast数据集



(e) 平均分类误差对比

图 1 GA 和 PAL-PSO 的测试误差收敛曲线

相比 GA 方法,PSO 方法本身收敛速度较快,同时采用参数范围预估计和对数刻度方法大大提高了粒子落入最优值邻域的几率,每次实验均收敛到最优值说明 PAL-PSO 方法采取的参数范围是可行的.事实上,GA 方法具有很强的全局搜索能力,如果采用更大的迭代代数,则 GA 方法可能搜索到更好的参数组合.但是对于 SVM 参数寻优问题,每进化一代都需要训练若干 SVM,而训练 SVM 的时间复杂度很高,是参数优化算法中时间消耗的主要因素,所以要求算法的迭代次数要少,从而算法的收敛速度成为一项重要指标.

5 结 论

本文根据 RBF 核函数的特点,通过计算训练样本的平均最近距离和平均最远距离预设核参数的范围,并采用对数刻度进行参数寻优,不但有效减小了模型优化算法的搜索空间,而且提高了算法在小值参数附近的寻优能力.同时,采用 PSO 方法提高了参数寻优的速度.实验表明,本文方法不仅搜索速度更高,而且能够收敛到更优的参数组合,是一种有效的 SVM 超参数搜索方法.

参考文献(References)

[1] Dong Y L, Xia Z H, Xia Z Q. A two-level approach to choose the cost parameter in support vector machines [J]. Expert Systems with Applications, 2008, 34(2): 1366-1370.

[2] Ayat N E, Cheriet M, Suen C Y. Automatic model selection for the optimization of SVM kernels [J]. Pattern Recognition, 2005, 38(10): 1733-1745.

[3] 刘向东,骆斌,陈兆乾. 支持向量机最优模型选择的研究[J]. 计算机研究与发展, 2005, 42(4): 576-581.
(Liu X D, Luo B, Chen Z Q. Optimal model selection for support vector machines[J]. J of Computer Research and Development, 2005, 42(4):576-581.)

[4] 朱家元,杨云,张恒喜,等. 支持向量机的多层动态自适应参数优化[J]. 控制与决策, 2004, 19(2): 223-225.
(Zhu J Y, Yang Y, Zhang H X, et al. Multi-layer adaptive parameters optimization approach for support vector machines [J]. Control and Decision, 2004, 19(2): 223-225.)

[5] 郑春红,焦李成,丁爱玲. 基于启发式遗传算法的 SVM 模型自动选择[J]. 控制理论与应用, 2006, 23(2): 187-192.
(Zheng C H, Jiao L C, Ding A L. Automatic model selection for support vector machines using heuristic genetic algorithm[J]. Control Theory and Applications, 2006, 23(2): 187-192.)

(下转第 377 页)