

粒子群算法优化双核支持向量机及应用*

聂立新^{1,2}, 张天侠¹, 赵波²

(1. 东北大学机械工程与自动化学院 沈阳, 110819) (2. 河南理工大学机械与动力工程学院 焦作, 454000)

摘要 针对支持向量机核函数和控制参数选取难度较大的问题, 提出了一种主动划分参数区间的双尺度径向基核支持向量机, 并用并行定向变异混合粒子群优化算法选取其控制参数。试验分析了利用标准数据集经多次独立重复试验得到的均值等统计量, 验证、测试了上述支持向量机模型, 同时考虑了类间数据不平衡的影响。结果表明, 双尺度径向基核函数的性能在多数情况下优于单径向基核函数, 并行定向变异的混合粒子群优化算法优于标准粒子群优化算法, 能够有效抑制早熟收敛, 有利于搜索到更优的支持向量机控制参数。

关键词 支持向量机; 双尺度核函数; 粒子群优化算法; 参数优化; 故障诊断

中图分类号 TH165.3

引言

支持向量机^[1] (support vector machine, 简称 SVM) 是一种基于统计学的 VC 维理论和结构风险最小化原理的小样本学习方法, 在模式分类上具有良好的泛化性能。支持向量机的理论基础已经较为完善, 但在具体应用中, 必须慎重处理如何选择核函数和控制参数的问题。常用的核函数有多项式核函数、径向基核函数以及多层感知器核函数等三种^[2], 针对不同的数据集, 核函数的表现也不尽相同。以支持向量机的交叉验证正分率为目标函数, 使用常用的优化方法均可搜索到支持向量机控制参数的较优值, 但未必是全局最优解。

笔者提出了一种用并行定向变异粒子群优化算法去优选双尺度径向基核支持向量机的控制参数的方法, 用标准数据集验证了该方法的效果, 并应用于发动机的故障诊断。

1 核方法与支持向量机

1.1 核方法

核方法^[3]是解决非线性模式识别问题的有效途径, 它利用较为简单的核函数运算, 既避免了在特征空间的复杂内积运算, 又避免了特征空间(学习机器)本身的设计。设原空间数据集 $S =$

$\{(x_1, y_1), \dots, (x_i, y_i), \dots, (x_l, y_l)\} \in X \times Y$, 其中: x_i 属于输入空间 $X \subseteq R^n$; y_i 属于输出空间 $Y \subseteq R^m$ 。通过一个非线性映射

$$\begin{aligned} \Phi: X &\rightarrow F \\ x &\mapsto \Phi(x) \end{aligned} \quad (1)$$

将原空间输入数据映射到新的特征空间 $F = \{\Phi(x) \mid x \in X\}$, 其中: $F \subseteq R^n$ 。该映射将原空间数据集转化为特征空间的数据集

$$S^\Phi = \{(\Phi(x_1), y_1), \dots, (\Phi(x_i), y_i), \dots, (\Phi(x_l), y_l)\} \in F \times Y \quad (2)$$

核方法利用上述映射将原空间非线性可分的问题转化为特征空间线性可分或近似线性可分的问题, 并且可分性的优劣取决于核函数的选取是否合适。目前, 核函数选取的研究主要有 3 个方面。

1) 构造特定的单核函数。文献[4]则将小波核函数用于支持向量机的决策和分类, 对于某具体问题, 总有一种核函数对其有良好的表达能力。

2) 合成核方法。可以证明, 核函数的凸组合仍然满足作为核函数的条件, 即 Mercer 条件。将不同特性的核函数进行组合, 集各类核函数之优点, 能够有效提高核方法的性能。文献[5]提出了一种多项式核和径向基核的组合核函数, 兼顾了内推和预测的能力, 现已得到较为广泛的应用。

3) 多尺度核方法。该方法是多核合成方法的特例, 所采用的原始核 $k_m(x, z)$ 均为同一核函数,

* 国家自然科学基金资助项目(51175153); 河南理工大学博士基金资助项目(B2012-105)

收稿日期: 2014-01-25; 修回日期: 2014-02-20

但核参数各不相同。这种方法灵活性很强,能够提供更完备的尺度选择。多尺度核方法的关键是找到一组具有多尺度表达能力的核函数,并且核函数组合能够最大限度地区分不同类别的样本。常用的多尺度核函数有径向基核函数和小波核函数等。

1.2 支持向量机

支持向量机是一种典型的核学习方法,有力地推动了核方法的应用,其求解模型为

$$\min \left\{ \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^l \xi_i \right\} \quad (3)$$

$$(\text{s. t. } y_i(\mathbf{w}^T \boldsymbol{\varphi}(x_i) + b) \geq 1 - \xi_i;$$

$$i = 1, 2, \dots, l; \xi_i \geq 0)$$

其中: $C \geq 0$ 为惩罚参数,起平衡区间距离最大化和分类误差最小化的作用。

式(4)可转化为以下对偶问题

$$\min \left(\frac{1}{2} \boldsymbol{\alpha}^T \mathbf{Q} \boldsymbol{\alpha} - \mathbf{e}^T \boldsymbol{\alpha} \right) \quad (4)$$

$$(\text{s. t. } \mathbf{y}^T \boldsymbol{\alpha} = 0; 0 \leq \alpha_i \leq C; i = 1, 2, \dots, l)$$

其中: $\mathbf{e} = [1, 1, \dots, 1]^T$; \mathbf{Q} 为 $l \times l$ 半正定矩阵,其中 $Q_{i,j} = y_i y_j \mathbf{K}(x_i, x_j)$, $\mathbf{K}(x_i, x_j) = \boldsymbol{\varphi}(x_i)^T \boldsymbol{\varphi}(x_j)$ 为核函数。

理论上,支持向量机只能解决“非此即彼”的二分类问题。工程应用中的多分类问题,可以通过多个二分类支持向量机的组合判决来实现^[4]。

在确定多分类支持向量机的组合方式以及核函数后,惩罚参数 C 以及核函数的核参数的选取是否合理成为支持向量机分类性能优劣的决定因素。最常用的算法性能的评价标准是 K -折交叉验证 (K -fold cross-validation, 简称 K -CV)^[6], 以惩罚参数 C 以及核函数的核参数为决策变量,交叉验证正分率为目标函数的优化模型经求解后得到最优参数,从而形成有较高鲁棒性的支持向量机分类器。上述优化模型通常具有多个局部最优解,一般的优化算法极易陷入局部最优,丧失了搜索到全局最优解的机会。对于上述多峰目标函数优化问题,群智能优化较常规优化算法有更大优势。

2 粒子群优化算法

粒子群优化 (particle swarm optimization, 简称 PSO) 算法是 Kennedy 等^[7]提出的一种基于群智能的优化算法,其速度更新和位置更新公式为

$$v_{ij}(t+1) = \omega v_{ij}(t) + c_1 r_1 (\text{pbest}_{ij}(t) - x_{ij}(t)) + c_2 r_2 (\text{gbest}_j(t) - x_{ij}(t)) \quad (5)$$

$$x_{ij}(t+1) = x_{ij}(t) + v_{ij}(t+1) \quad (6)$$

其中: x_{ij} 和 v_{ij} 分别为粒子 i 第 j 维的位置和速度; pbest_{ij} 为第 i 个粒子第 j 维的历史最优位置; gbest_j 为群体第 j 维历史最优位置; c_1, c_2 分别为粒子自身加速系数和群体加速系数,均为非负实数; $r_1, r_2 \sim U[0, 1]$; ω 称为惯性权重; t 为粒子的经历时间或迭代次数。

尽管与传统的基于梯度法的优化方法相比,PSO 算法能够有良好的表现,但其也有一个致命缺陷,即在迭代过程中极易出现早熟收敛,导致其陷入局部最优^[8]。为抑制早熟收敛,文献^[9]提出了遗传算法与粒子群算法相融合的 GAPSO 混合算法。文献^[10]提出了一种在最优解周围的区域内进行混沌搜索的混沌 PSO 算法,取得了较好的效果。但现有的改进算法通常同时对所有优化决策变量进行扰动变异,存在文献^[8]所述的“进 2 退 1”(two steps forward, one step back)现象。

3 HPSO-PDT 算法优化双核支持向量机

3.1 双尺度径向基核函数

径向基核函数可以表示为

$$k(x, z) = \exp(-\gamma \|x - z\|^2) \quad (7)$$

其中: $\gamma \in [2^{-15}, 2^{15}]$ 为核参数。

当 γ 较大时,可以分辨剧烈变化的样本;当 γ 较小时,可对平缓变化的样本进行分类^[3]。根据上文所述的合成核方法和多尺度核方法,文中拟采用如下核函数

$$k(x, z) = \alpha \exp(-\gamma_1 \|x - z\|^2) + (1 - \alpha) \exp(-\gamma_2 \|x - z\|^2) \quad (8)$$

$$(\text{s. t. } \alpha \in [0, 1]; \gamma_1 \in [2^{-15}, 1]; \gamma_2 \in [1, 2^{15}])$$

可以证明,式(8)所示核函数满足 Mercer 条件要求,可以用于核方法运算。该核函数主动将核参数区间分为两部分,通过调整权参数 α 大小去平衡两个尺度的子核函数的效能。

3.2 基于不平衡数据的支持向量机

在支持向量机的具体应用中,参与训练和测试的数据集通常是不均衡的,如在设备的故障诊断中,正常数据一般要比故障数据多一些。在数据不均衡场合,为每类数据设置不同惩罚参数 C ,有可能使支持向量机模型有更强的分类能力^[11]。设数据有 c 类,则不平衡支持向量机模型表达如下

$$\min \left\{ \frac{1}{2} \|w\|^2 + C_j \sum_{i=1}^{l_j} \xi_i + C_k \sum_{i=1}^{l_k} \xi_i \right\} \quad (9)$$

(s. t. $C_j, C_k \in [2^{-15}, 2^{15}]$; $j, k \in \{1, 2, \dots, c\}$)

3.3 HPSO-PDT 算法

笔者采用一种并行定向扰动的混合粒子群优化 (hybrid particle swarm optimization based on parallel directional turbulence, 简称 HPSO-PDT) 算法^[12]。其算法核心及特点是当算法出现早熟收敛时, 群体中产生变异的每个粒子仅发生单坐标方向扰动, 从而避免了“进 2 退 1”现象的出现, 避免了优化决策变量之间的干扰。若优化目标为最小化问题, 则 HPSO-PDT 的算法流程见图 1。

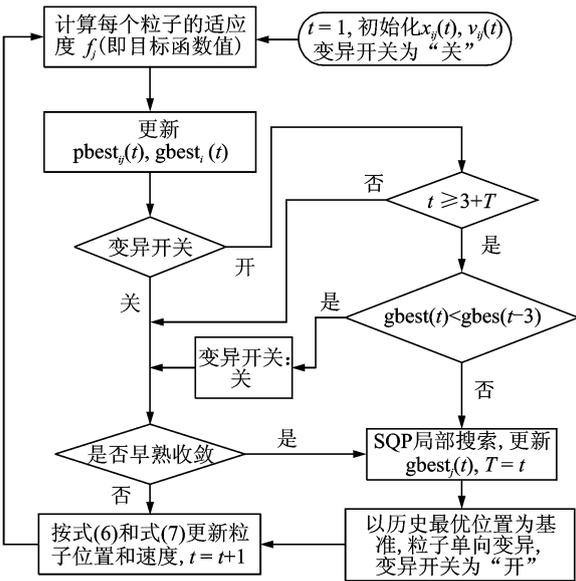


图 1 HPSO-PDT 算法的流程

Fig. 1 Flow chart of HPSO-PDT algorithm

HPSO-PDT 算法由标准 PSO 算法迭代、序列二次规划 (sequential quadratic programming, 简称 SQP) 算法局部搜索和粒子单决策变量扰动变异三部分构成。标准 PSO 算法使粒子在按规则运动过程中不断更新个体和群体最优值; SQP 算法具有优秀的局部搜索能力, 能够快速到达当前所在区域的“谷底”; 粒子的并行定向变异可使每个参与变异的粒子分别沿选定的单个坐标方向 (即单优化决策变量) 进行搜索, 继承了现有的群体最优位置信息。

3.4 HPSO-PDT 算法优化双核 SVM 的数据验证

取支持向量机模型性能的评价函数为 10-折交叉验证 (即 $K=10$) 正分率, 并将其作为 PSO 算法的

适应度函数, 即优化目标函数, 便可通过 PSO 算法得到使交叉验证正分率最高 (或错分率最低) 的决策变量, 即径向基核函数参数 γ_1, γ_2 、双核权参数 α 以及各数据类别的惩罚参数 C_k 。

为验证上述方法性能, 取台湾大学林智仁所列的 UCI 数据库中的 Iris, Wine 和 Glass 3 个标准数据集进行测试。上述标准数据集的属性见表 1, 每个数据集均随机地划分成训练集和测试集, 并且按类近似平均分配。

表 1 数据集属性及划分

Tab. 1 Properties and partition of data sets

数据集	样本数	特征数	类数	训练集	测试集
				样本数	样本数
Iris	150	4	3	75	75
Wine	178	13	3	90	88
Glass	214	9	6	109	105

训练集数据用于训练支持向量机, 以便计算 10-折交叉验证正分率, 测试集用于验证训练得到的 SVM 的性能。粒子群优化算法的惯性权重 ω 、加速系数 c_1, c_2 以及粒子群规模 ps、最大迭代次数 t_{max} 见表 2。

为比较标准 PSO 和 HPSO-PDT 两种优化算法的优劣, 以及评价径向基核 SVM、双尺度径向基核 SVM 以及双尺度径向基核加多类惩罚参数 SVM 3 种支持向量机模型分类效果, 现采用表 3 列出的 6 种方法进行比较。支持向量机平均正分率的每个数据是按同一方法进行 10 次独立重复试验所得的数据的均值和标准差。

表 2 粒子群优化算法的控制参数

Tab. 2 Controlling parameters of PSO algorithm

参数	t_{max}	c_1	c_2	ps	ω
取值	100	1.5	1.5	25	$0.9 - \frac{t-1}{2(t_{max}-1)}$

从表 3 的数据可以看出:

- 1) HPSO-PDT 算法在多数情况下, 能够比标准 PSO 算法得到更高的交叉验证正分率, HPSO-PDT 算法优于标准 PSO 算法;
- 2) 与径向基核支持向量机相比, 双尺度核支持向量机能够取得更高的交叉验证正分率, 且测试集的正分率也有明显提高;
- 3) 按类惩罚的不平衡数据支持向量机模型由于其控制参数数量的增加, 标准 PSO 算法的优化效果欠稳定, Iris 和 Wine 数据集的交叉验证正分率低

于单惩罚参数的标准支持向量机,但采用 HPSO-PDT 算法后效果有较大提高。

总之,从试验数据的比对可以看出:HPSO-PDT 算法能够取得比标准 PSO 算法更好的优化效

果;双尺度径向基核支持向量机比单径向基核支持向量机有更高的适应性和鲁棒性;按类惩罚不平衡数据支持向量机在某些特定场合能够获得更好的评价性能。

表 3 粒子群优化算法优化支持向量机的平均正分率

Tab. 3 Average correct classification rates of SVMs optimized by PSO algorithms

数据集	正分率类型	PSO 算法			HPSO-PDT 算法			%
		径向基核	双核	双核 C_k	径向基核	双核	双核 C_k	
Wine	交叉验证	98.67±0.47	98.44±0.57	98.33±0.59	98.78±0.35	98.44±0.57	98.44±0.57	
	训练集	99.89±0.35	100±0	99.89±0.35	99.67±0.54	100±0	99.78±0.47	
	测试集	96.7±0.36	97.27±0.96	97.39±0.55	96.93±0.55	97.5±0.9	98.07±0.77	
Glass	交叉验证	68.81±0	70.92±0.87	71.19±0.47	68.53±0.44	70.92±0.87	71.47±0.29	
	训练集	97.61±2.13	99.82±0.39	99.36±1.15	96.06±0.97	100±0	99.91±0.29	
	测试集	71.71±1.01	75.71±1.75	74.95±1.19	73.33±1.19	75.33±1.58	75.71±1.63	
Iris	交叉验证	98.67±0	97.73±0.9	97.73±0.64	98.67±0	98.53±0.42	97.2±0.42	
	训练集	100±0	100±0	98.8±0.98	100±0	100±0	98.67±0.63	
	测试集	93.33±0	93.47±0.98	96.27±1.51	93.33±0	93.47±0.42	96.93±0.9	

4 基于 HPSO-PDT 算法优化双核 SVM 的发动机故障诊断

为辨别发动机正常工作状态 S_1 以及排气门开裂 S_2 、活塞环断裂 S_3 、失火故障 S_4 和水泵穴蚀 S_5 等 4 种故障状态,分别测量了上述各状态的发动机机油温度 T_{oil} 、催化器温度 T_c 、输出轴扭矩 T_s 、曲柄轴箱压力 P 和主轴转速 N_s 等 5 种信号共 300 组数据。根据文中所述方法,将采集到的数据随机均匀分割成训练集和测试集两部分(各 150 组),训练集数据归一化后用于粒子群优化支持向量机的训练,测试集数据用于验证粒子群优化支持向量机模型的效果。故障诊断的流程如图 2 所示。

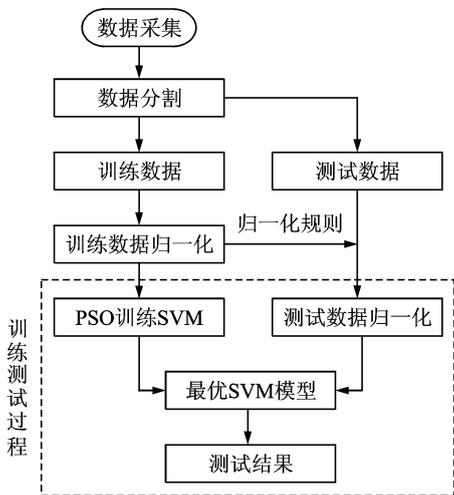


图 2 发动机故障诊断流程

Fig. 2 Flow chart of engine fault diagnosis

表 4 列出了训练集归一化后的部分数据。为保证试验的可靠性,发动机故障诊断中,每种方案均进行 10 次独立重复试验,并根据试验值计算测试结果(测试集正分率)的均值、标准差、最大值和最小值等 4 个统计量,具体数据见表 5。

表 4 发动机状态归一化数据

Tab. 4 Normalization data of engine status

状态	T_{oil}	T_c	T_s	P	N_s
S_1	0.498 5	0.436 6	0.870 6	0.005 1	0.902 1
S_1	0.408 1	0.432 5	0.819 0	0.004 2	0.902 8
S_1	0.624 0	0.444 3	0.862 8	0.004 7	0.906 1
S_1	0.433 2	0.433 2	0.815 9	0.002 0	0.905 0
S_1	0.456 8	0.433 8	0.864 8	0.005 0	0.900 2
S_1	0.355 9	0.431 2	0.820 6	0.002 9	0.898 0
S_2	0.573 8	0.373 9	0.899 7	0.011 4	0.902 1
S_2	0.578 8	0.414 4	0.895 0	0.015 5	0.902 1
S_2	0.583 8	0.345 9	0.899 7	0.023 7	0.902 1
S_3	0.307 7	0.164 1	0.656 4	0.137 2	0.907 8
S_3	0.337 9	0.137 1	0.656 5	0.415 1	0.900 6
S_3	0.388 1	0.166 6	0.650 9	0.120 1	0.903 3
S_4	0.593 9	0.862 6	0.921 9	0.011 2	0.902 1
S_4	0.603 9	0.861 0	0.922 5	0.011 2	0.902 6
S_4	0.614 0	0.861 0	0.923 9	0.011 0	0.903 3
S_5	0.709 3	0.386 6	0.976 0	0.057 2	0.901 8
S_5	0.743 5	0.371 4	0.977 0	0.057 0	0.893 8
S_5	0.672 2	0.363 8	0.959 5	0.056 9	0.904 7
S_5	0.768 1	0.382 9	0.992 4	0.059 0	0.896 5

从表 5 中的测试结果可看出,双尺度径向基核支持向量机的测试结果与单径向基核支持向量机相

比有明显提高,同时,其标准差有较大减小,说明双尺度径向基核 SVM 具有更高的稳定性。按类惩罚的不平衡 SVM 方法与单惩罚参数 SVM 相比,分类效果略有提高,但差别不大。HPSO-PDT 算法优化 SVM 比标准 PSO 算法优化 SVM 的效果有较大提高,能够更加逼近全局最优解。通过核函数构造以及优化算法的改进,故障诊断的正确率提高了 6.49%。最终,可以选取 HPSO-PDT 算法优化双尺度径向基核支持向量机模型用于故障诊断,支持向量机单惩罚参数或按类惩罚均可,但按类惩罚会有更好的性能。

表 5 发动机故障诊断正分率

Tab. 5 Correct classification rates of engine fault diagnosis

统计量	PSO 算法			HPSO-PDT 算法		
	径向基	双核	双核	径向基	双核	双核
			C_k			C_k
最小值	88	98.67	98	89.33	98	98
最大值	98.67	100	100	98.67	100	100
均值	93.53	99.53	99.53	95.07	99.53	99.6
标准差	4.93	0.77	0.77	4.28	0.77	0.72

5 结束语

笔者建立了主动划分核参数区间的双尺度径向基核支持向量机,考虑了类间数据不平衡对支持向量机的影响,用并行定向变异的混合粒子群优化算法优选支持向量机的控制参数,用标准数据集验证了其性能,在发动机故障诊断中取得了良好的应用。

参 考 文 献

[1] Vapnik V N. The nature of statistical learning theory [M]. 2nd. Berlin: Springer-Verlag,1999:138-170.

[2] 申秀敏,左曙光,韩乐,等.基于支持向量机的车内噪声声品质预测[J].振动、测试与诊断,2011,31(1):55-58.
Shen Xiumin, Zuo Shuguang, Han Le, et al. Interior vehicle noise quality prediction using support vector machines[J]. Journal of Vibration, Measurement & Diagnosis,2011,31(1):55-58. (in Chinese)

[3] 汪洪桥,孙富春,蔡艳宁,等.多核学习方法[J].自动化学报,2010,36(8):1037-1050.
Wang Hongqiao, Sun Fuchun, Cai Yanning, et al. On multiple kernel learning methods [J]. Acta Automatica Sinica, 2010,36(8):1037-1050. (in Chinese)

[4] 董绍江,汤宝平,宋涛.改进投票策略的 Morlet 小波核支持向量机及应用[J].振动、测试与诊断,2011,31(3):314-317.
Dong Shaojiang, Tang Baoping, Song Tao. Morlet wavelet kernel SVM improved by voting strategy and its application[J]. Journal of Vibration, Measurement & Diagnosis,2011,31(3):314-317. (in Chinese)

[5] Smits G F, Jordaen E M. Improved SVM regression using mixtures of kernels[C] // Proceedings of the 2002 International Joint Conference on Neural Networks . Honolulu: IEEE,2002:2785-2790.

[6] 邓乃扬,田英杰.支持向量机-理论、算法与拓展[M].北京:科学出版社,2009:134-155.

[7] Kennedy J, Eberhart R C. Particle swarm optimization [C]// IEEE International Conference on Neural Networks. Piscataway: IEEE,1995:1942-1948.

[8] Van Den B. An analysis of particle swarm optimizers [D]. South Africa:University of Pretoria, 2002.

[9] Nie Ru, Yue Jianhua. A GA and particle swarm optimization based hybrid algorithm [C]// Proceeding of IEEE Congress on Evolutionary Computation 2008. Hongkong: IEEE, 2008 : 1047-1050.

[10] 刘玲,钟伟民,钱锋.改进的混沌粒子群优化算法[J].华东理工大学学报:自然科学版,2010,36(2):267-272.
Liu Ling, Zhong Weimin, Qian Feng. An improved chaos-particle swarm optimization algorithm[J]. Journal of East China University of Science and Technology: Natural Science Edition,2010,36(2):267-272. (in Chinese)

[11] Chang C C, Lin C J. LIBSVM : a library for support vector machines [EB/OL]. [2014-01-10]. <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.

[12] 聂立新,张天侠,郭立新.并行定向扰动的混合粒子群优化算法[J].计算机应用研究,2013,30(6):1633-1635.
Nie Lixin, Zhang Tianxia, Guo Lixin. Hybrid particle swarm optimization algorithm based on parallel directional turbulence[J]. Application Research of Computers,2013,30(6):1633-1635. (in Chinese)



第一作者简介:聂立新,男,1975年6月生,博士研究生、副教授。主要从事车辆状态监测与故障诊断研究。曾发表《基于DNPSO的支持向量机的发动机故障诊断》(《东北大学学报:自然科学版》2012年第33卷第4期)等论文。
E-mail: nielixin@hpu.edu.cn

