

多分类最大间隔孪生支持向量机^①

高斌斌, 王建军

西南大学 数学与统计学院, 重庆 400715

摘要: 提出一种新的多分类最大间隔孪生支持向量机算法. 该算法通过引入间隔以结构风险最小为优化目标建立分类模型, 并采用一对一对余的结构训练子分类器. 仿真实验和真实数据实验表明: 所提算法能有效提高模型的泛化性能.

关键词: 多分类; 孪生支持向量机; 最大间隔; 一对一对余; 结构风险最小化原则

中图分类号: TP181

文献标志码: A

支持向量机(Support vector machine, SVM)^[1] 是 Vapnik 等人提出的一种基于统计学习理论的机器学习方法, 由于具有良好的学习能力和泛化性能得到广泛应用. 为了改善 SVM 执行效率, Jayadeva 等学者提出了孪生支持向量机(TSVM)^[2] 为解决二元分类问题. 近年来基于 TSVM 思想发展了许多非平行超平面学习方法: LSTSVM^[3]、TBSVM^[4]、TPMSVM^[5]. 这些算法都适宜于解决二分类问题, 但是在现实世界中普遍存在着大量多分类问题, 然而目前基于 TSVM 多分类的研究还比较少. 尽管已有学者提出通过二分类来实现多类问题的分类^[6], 但是研究^[7] 表明, 直接应用一些二分类算法到多类问题时往往并不那么有效. 最近徐等提出 Twin-KSVC 算法为解决多类分类, 但是该算法仅仅以经验风险最小为优化目标, 且对偶问题中矩阵奇异性导致其不能得到较好的分类性能. 针对 Twin-KSVC 不足, 本文提出了一种新的多分类算法, 称之为多分类最大间隔孪生支持向量机(K-MTSVM).

1 最大间隔孪生支持向量多分类机(K-MTSVM)

对于 K 分类问题, 训练集 $T = \{(\mathbf{x}_i, y_i) \mid \mathbf{x}_i \in R^n, y_i \in \{1, 2, \dots, k\}\} (i=1, 2, \dots, l)$, 其中 y_i 为训练样本 \mathbf{x}_i 的类别标号, K-MTSVM 通过 3 分类间接实现 K 分类问题, 因此训练集被划分为三部分. 记矩阵 $\mathbf{X}_+ \in R^{l_1 \times n}$ 、 $\mathbf{X}_- \in R^{l_2 \times n}$ 分别表示两类样本, 矩阵 $\mathbf{X}_0 \in R^{l_3 \times n}$ 为其余类样本, 其样本标号分别为“+1”, “-1”, “0”, 其中 l 为样本总数且 $l = l_1 + l_2 + l_3$.

1.1 线性 K-MTSVM

在线性可分情形下, 线性 K-MTSVM 目标是在 n 维空间寻求两不平行的超平面:

$$f_+(\mathbf{x}) = \mathbf{x}^T \mathbf{w}_+ + b_+ = 0, \quad f_-(\mathbf{x}) = \mathbf{x}^T \mathbf{w}_- + b_- = 0 \quad (1)$$

构造第一类超平面 $f_+(\mathbf{x}) = 0$ 的原始优化问题

$$\begin{aligned} \min_{\mathbf{w}_+, b_+, \xi_-, \eta_-} & \frac{1}{2} \gamma_+ \|\mathbf{w}_+\|^2 + \frac{1}{2} \|\mathbf{X}_+ \mathbf{w}_+ + \mathbf{e}_+ b_+\|^2 + c_+ \mathbf{e}_+^T \xi_- + d_+ \mathbf{e}_0^T \eta_- \\ \text{s. t.} & -(\mathbf{X}_- \mathbf{w}_+ + \mathbf{e}_- b_+) + \xi_- \geq \mathbf{e}_-, \quad \xi_- \geq 0 \\ & -(\mathbf{X}_0 \mathbf{w}_+ + \mathbf{e}_0 b_+) + \eta_- \geq \mathbf{e}_0(1 - \epsilon), \quad \eta_- \geq 0 \end{aligned} \quad (2)$$

① 收稿日期: 2013-01-11

基金项目: 国家自然科学基金资助项目(61273020, 11001227); 中央高校基本科研业务费专项资金资助(XDJK2010B005).

作者简介: 高斌斌(1986-), 男, 甘肃天水人, 硕士研究生, 主要从事神经网络与数据挖掘方面的研究.

通信作者: 王建军, 教授.

其中 $\|\cdot\|$ 是 L_2 范数, γ_+ 是间隔参数, $c_+ > 0$, $d_+ > 0$ 是惩罚参数, ξ_+ , η_+ 为松弛变量, e_+ , e_- , e_0 是单位列向量, 可调参数 $0 < \epsilon < 1$.

最优超平面 $f_+(x) = 0$ 与边界超平面 $f_+(x) = -1$ 的间隔为 $1/\|w_+\|$, 极大化此间隔与极小化 $\|w_+\|^2$ 等价. 优化目标函数(2)中引入的 $\frac{1}{2}\gamma_+ \|w_+\|^2$ 控制模型获得第一类最优超平面; 目标函数中第二项极小化“+”类训练样本到最优超平面的均方距离和, 导致“+”类训练样本尽可能地靠近 $f_+(x) = 0$; 第一个约束条件限制“-”类样本至少以距离 1 远离 $f_+(x) = 0$; 第二个约束条件限制“0”类样本至少以 $1 - \epsilon$ 距离远离 $f_+(x) = 0$, 其中 ϵ 为一个可调参数.

为方便获得最优超平面, 需要求得原始二次规划问题(2)的对偶问题, 其 Lagrangian 函数

$$L = \frac{1}{2}\gamma_+ \|w_+\|^2 + \frac{1}{2}\|X_+ w_+ + e_+ b_+\|^2 + c_+ e_-^T \xi_- + d_+ e_0^T \eta_- - \lambda_1^T \xi_- - \lambda_2^T \eta_- + \quad (3)$$

$$\alpha_1^T (X_- w_+ + e_- b_+ - \xi_- + e_-) + \alpha_2^T (X_0 w_+ + e_0 b_+ - \eta_- + e_0(1 - \epsilon))$$

其中 $\lambda_i \geq 0$, $\alpha_i \geq 0$ ($i = 1, 2$) 是 Lagrangian 乘子, 根据 KKT 条件

$$\nabla_{w_+} L = \gamma_+ w_+ + X_+^T (X_+ w_+ + e_+ b_+) + X_+^T \alpha_1 + X_0^T \alpha_2 = 0 \quad (4)$$

$$\nabla_{b_+} L = e_+^T (X_+ w_+ + e_+ b_+) + e_+^T \alpha_1 + e_0^T \alpha_2 = 0 \quad (5)$$

$$\nabla_{\xi_-} L = c_+ e_- - \lambda_1 - \alpha_1 = 0 \quad (6)$$

$$\nabla_{\eta_-} L = d_+ e_0 - \lambda_2 - \alpha_2 = 0 \quad (7)$$

$$\alpha_1^T (X_- w_+ + e_- b_+ - \xi_- + e_-) = 0 \quad (8)$$

$$\alpha_2^T (X_0 w_+ + e_0 b_+ - \eta_- + e_0(1 - \epsilon)) = 0 \quad (9)$$

$$\lambda_1^T \xi_- = 0, \lambda_2^T \eta_- = 0 \quad (10)$$

因为 $\lambda_1 \geq 0$, $\lambda_2 \geq 0$, 从(6)、(7)式可得

$$0 \leq \alpha_1 \leq c_+ e_-, 0 \leq \alpha_2 \leq d_+ e_0 \quad (11)$$

令 $\alpha = [\alpha_1; \alpha_2]$, $F_+ = [c_+ e_-; d_+ e_0]$, 则 $0 \leq \alpha \leq F_+$. 令 $E = \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}$, 合并(4)、(5)可得

$$([X_+, e_+]^T [X_+, e_+] + \gamma_+ E) [w_+^T, b_+]^T + [X_-, e_-]^T \alpha_1 + [X_0, e_0]^T \alpha_2 = 0 \quad (12)$$

令 $H_i = [X_i, e_i]$ ($i = +, -, 0$), $P_- = [H_-; H_0]$, $v_+ = [w_+^T, b_+]^T$, (12) 式化为

$$(H_+^T H_+ + \gamma_+ E) v_+ + H_+^T \alpha_1 + H_0^T \alpha_2 = 0 \quad (13)$$

$$v_+ = -(H_+^T H_+ + \gamma_+ E)^{-1} (P_-^T \alpha) \quad (14)$$

令 $e_1 = [e_-; (1 - \epsilon)e_0]$, 将(14)式带入 Lagrangian 函数(3), 使用条件(4)–(10)可以得到(2)的 Wolfe 对偶问题

$$\max_{\alpha} e_1^T \alpha - \frac{1}{2} \alpha^T P_- (H_+^T H_+ + \gamma_+ E)^{-1} P_-^T \alpha \quad (15)$$

$$\text{s. t. } 0 \leq \alpha \leq F_+$$

第二类超平面的 $f_-(x) = 0$ 原始优化问题可以类似的构造为

$$\min_{w_-, b_-, \xi_+} \frac{1}{2}\gamma_- \|w_-\|^2 + \frac{1}{2}\|X_- w_- + e_- b_-\|^2 + c_- e_+^T \xi_+ + d_- e_0^T \eta_+ \quad (16)$$

$$\text{s. t. } (X_+ w_- + e_+ b_-) + \xi_+ \geq e_+, \xi_+ \geq 0$$

$$(X_0 w_- + e_0 b_-) + \eta_+ \geq e_0(1 - \epsilon), \eta_+ \geq 0$$

同样的方法可以求得(16)的对偶问题.

线性 K-MTSVM 采用一对一对余的结构训练模型, 因此对于 K 分类问题, 需要构造 $\frac{K(K-1)}{2}$ 个子分类器. 子分类器构造一个 3 分类问题训练模型, 其决策函数生成三元的输出 $\{+1, -1, 0\}$. 对一个新的输入 $x \in R^n$ 子分类器确定其类别标号依据决策函数

$$f(x) = \begin{cases} 1 & x^T w_+ + b_+ > -1 + \epsilon \\ -1 & x^T w_- + b_- < 1 - \epsilon \\ 0 & \text{其它} \end{cases} \quad (17)$$

K-MTSVM 对输入 \mathbf{x} 最终的类别判别依据子分类器综合投票的原则, 具体算法描述如下:

- Step 1. 输入训练样本 T , 预测样本 \mathbf{x} , 初始化 $i = 1, \mathbf{V} \leftarrow \mathbf{0} \in R^{K \times 1}$.
- Step 2. $j = i + 1$, 初始化样本标签 $L \leftarrow \{L_1, L_2, \dots, L_K\}$.
- Step 3. 初始化训练样本, X_+, X_- 分别为第 L_i, L_j 类样本, X_0 为其余类($L/\{L_i, L_j\}$) 样本.
- Step 4. 求解 K-MTSVM 对偶问题得到最优解, 由(17) 式计算 $f(\mathbf{x})$.
- Step 5. 如果 $f(\mathbf{x}) = 1$, 则 $V(i) \leftarrow V(i) + 1$; 如果 $f(\mathbf{x}) = -1$, 则 $V(j) \leftarrow V(j) + 1$;
如果 $f(\mathbf{x}) = 0$, 则 $V(i) \leftarrow V(i) - 1, V(j) \leftarrow V(j) - 1$.
- Step 6. 如果 $j < K$, 则 $j = j + 1$, 返回 Step 3, 否则执行 Step 7.
- Step 7. 如果 $i < K - 1$, 则 $i = i + 1$, 返回 Step 2, 否则执行 Step 8.
- Step 8. 计算向量 V 中元素最大值的下标 t , 返回 L_t , 即为 \mathbf{x} 的类别.

1.2 非线性 K-MTSVM

为解决线性不可分问题, 引进从输入空间 R^n 到一个高维的 Hilbert 空间 \mathcal{H} 的变换 $\varphi(\cdot)$, 这样在原空间寻找超平面的问题转化为在特征空间 \mathcal{H} 中寻找两个由核生成的超平面的问题

$$f_+(\mathbf{x}) = \kappa(\mathbf{x}, X^T)\mathbf{u}_+ + b_+ = 0, f_-(\mathbf{x}) = \kappa(\mathbf{x}, X^T)\mathbf{u}_- + b_- = 0 \quad (18)$$

其中 $X = [X_+; X_-; X_0]$, $\kappa(x, y) = \langle \varphi(x) \cdot \varphi(y) \rangle$ 为核函数. 类似于线性情形, 构造第一类核超平面的原始优化问题

$$\begin{aligned} \min_{\mathbf{u}_+, b_+, \xi_-} \quad & \frac{1}{2} \gamma_+ \|\mathbf{u}_+\|^2 + \frac{1}{2} \|\kappa(X_+, X^T)\mathbf{u}_+ + \mathbf{e}_+ b_+\|^2 + c_+ \mathbf{e}_-^T \xi_- + d_+ \mathbf{e}_0^T \boldsymbol{\eta}_- \\ \text{s. t.} \quad & -(\kappa(X_-, X^T)\mathbf{u}_+ + \mathbf{e}_- b_+) + \xi_- \geq \mathbf{e}_-, \xi_- \geq \mathbf{0} \\ & -(\kappa(X_0, X^T)\mathbf{u}_+ + \mathbf{e}_0 b_+) + \boldsymbol{\eta}_- \geq \mathbf{e}_0(1 - \epsilon), \boldsymbol{\eta}_- \geq \mathbf{0} \end{aligned} \quad (19)$$

非线性 K-MTSVM 在形式上完全类似于线性 K-MTSVM, 不同的是非线性 K-MTSVM 将线性 K-MTSVM 模型中的 X_+, X_-, X_0 分别置换为 $\kappa(X_+, X^T), \kappa(X_-, X^T), \kappa(X_0, X^T)$, 因此其对偶问题的求解以及算法的执行过程完全同于线性 K-MTSVM. 值得注意的是, 核空间的 K-MTSVM 算法是在高维空间执行分类任务, 无疑相对线性 K-MTSVM 计算的开销会大幅增加.

1.3 算法的比较与分析

K-MTSVM 通过最大化间隔以结构风险最小化为目标建立分类模型, 因此提高了分类性能, 而 Twin-KSVC 以经验风险最小化为优化目标, 不存在类似 K-MTSVM 的间隔信息, 泛化能力较差. 尽管 Twin-KSVC 与 K-MTSVM 的对偶问题在形式上非常类似, 然而存在本质区别, Twin-KSVC 中 δ 是一个固定的正数, 而 K-MTSVM 中的参数 γ_+, γ_- 起着权衡正则项与经验风险的作用, 因此它们反映着结构风险最小化的原则. 实验表明, 选择适合的参数 γ_+, γ_- 可以提高分类准确率.

K-MTSVM、Twin-KSVC 的求解归结于具有线性约束的二次凸规划问题, 此类问题的计算复杂度主要取决于规划中变量的个数和线性约束方程的个数. 根据文献[8] 的方法, 可以证明 Twin-KSVC 与 K-MTSVM 的复杂度同级.

2 实验与分析

在人工数据集与 UCI 标准数据集上分别对线性和非线性的 Twin-KSVC、K-MTSVM 算法进行实验. 实验环境: Windows 7 操作系统, CPU 为 Intel core i3(2.4 GHz), 内存为 2 GB, 运行软件为 MATLAB7. 10. 非线性情形下使用高斯核函数: $\kappa(x, y) = \exp(-\|x - y\|^2 / (2\sigma^2))$. 所有算法对偶问题使用 quadprog.m 函数求解, 分类准确率(Accuracy) 采用标准的 10 折交叉验证方法.

2.1 人工数据集

为直观说明 K-MTSVM 的分类性能, 模拟生成正态分布样本数为 300 的二维 3 分类数据. 线性情形下子分类器的分类结果如图 1(a)(b)(c) 所示, 最终的分类超平面如图 1(d) 所示. 图 2 展示了非线性 K-MTSVM 算法在该数据集上的分类结果. 图 1(d) 与图 2(d) 表明非线性的 K-MTSVM 相对线性获得了较好的分类性能.

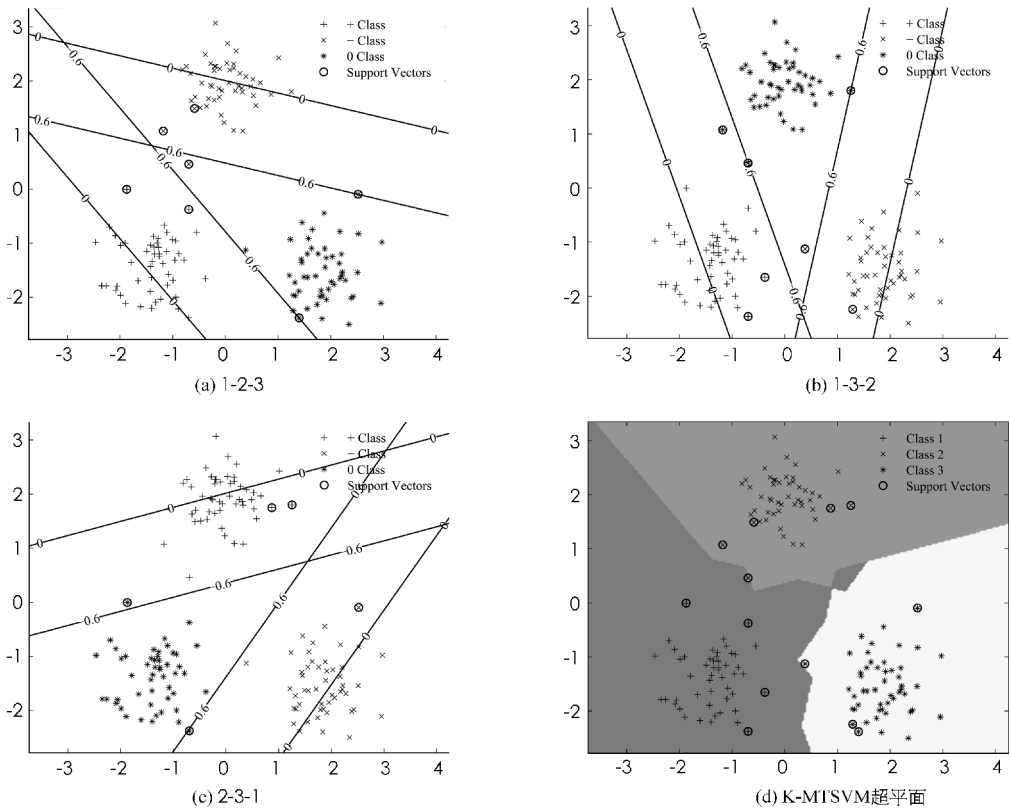


图 1 $\gamma=2^{-1}$, $\epsilon=0.4$ 时线性 K-MTSVM 训练结果

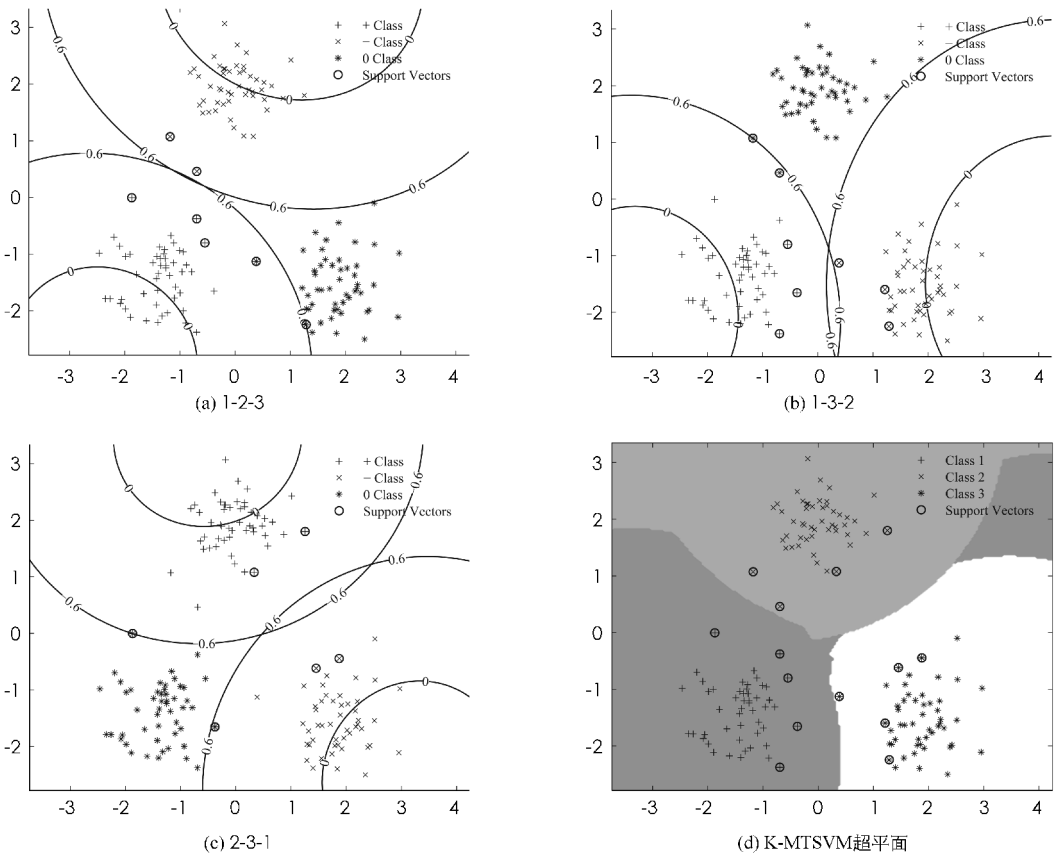


图 2 $\gamma=2^{-1}$, $\epsilon=0.4$, $\sigma=2^3$ 时非线性 K-MTSVM 训练结果

2.2 真实数据集

为了进一步测试 K-MTSVM 算法的有效性和实用性, 我们在 UCI 机器学习数据集上选取 7 个常用的数据集同 Twin-KSVC 对比实验, 表 1 中 $l \times n \times k$ 列举了数据的基本特征, 其中 l 为样本数目, n 为样本维数, k 为类别数. 实验之前, 为了减少样本不同特征的数量差别, 归一化样本数据在区间 $[-1, 1]$ 上.

为了简化参数的搜索难度, 设置 Twin-KSVC、K-MTSVM 惩罚参数 $c_+ = c_- = 1$, $d_+ = d_- = 1$, K-MTSVM 中间隔参数 $\gamma_+ = \gamma_- = \gamma$. 随机选择 20% 的训练集作调试集, 参数是在调试集上通过网格搜索方法以 10 折交叉验证分类准确率作为评价准则选择. 高斯核参数 σ 以及间隔参数 γ 的搜索空间是 $\{2^i \mid i = -6, -4, -2, -1, 0, 1, 2, 4, 6\}$, ϵ 的搜索范围为 $\{0.1, 0.2, 0.3, 0.4\}$, 一旦参数被确定, 调试集将还原到训练集中.

表 1 线性 Twin-KSVC 与 K-MTSVM 的分类结果

数据集	$(l \times n \times k)$	Twin-KSVC			K-MTSVM		
		ϵ	Accuracy/%	Time/s	(ϵ, γ)	Accuracy/%	Time/s
Iris	$(150 \times 4 \times 3)$	0.2	76.67 ± 3.51	1.22	$(0.2, 2^{-4})$	80.67 ± 3.78	1.26
Wine	$(178 \times 13 \times 3)$	0.2	96.67 ± 5.37	1.21	$(0.3, 2^{-2})$	98.33 ± 2.68	1.14
Balance	$(625 \times 4 \times 3)$	0.3	90.56 ± 3.82	26.99	$(0.4, 2^2)$	90.57 ± 2.40	10.79
Soybean	$(47 \times 35 \times 4)$	0.3	74.50 ± 15.54	0.43	$(0.3, 2^{-2})$	100.0 ± 0.0	1.06
Ecoli	$(327 \times 7 \times 5)$	0.3	87.43 ± 4.96	7.14	$(0.2, 2^{-4})$	88.31 ± 4.89	6.46
Glass	$(214 \times 9 \times 6)$	0.2	80.26 ± 9.22	13.23	$(0.2, 2^{-4})$	85.18 ± 6.05	22.64
Dermatology	$(358 \times 34 \times 6)$	0.3	94.71 ± 1.88	12.48	$(0.4, 2^2)$	96.93 ± 1.57	23.59

表 2 非线性 Twin-KSVC 与 K-MTSVM 的分类结果

数据集	Twin-KSVC			K-MTSVM		
	(ϵ, σ)	Accuracy/%	Time/s	$(\epsilon, \gamma, \sigma)$	Accuracy/%	Time/s
Iris	$(0.3, 2^4)$	94.67 ± 5.26	2.19	$(0.2, 2^{-6}, 2)$	96.00 ± 4.66	1.63
Wine	$(0.3, 2^0)$	94.31 ± 7.28	2.41	$(0.3, 2^0, 2^2)$	98.36 ± 2.64	2.63
Balance	$(0.3, 2^4)$	94.88 ± 2.58	75.02	$(0.4, 2^1, 2^0)$	95.52 ± 2.10	36.84
Soybean	$(0.3, 2^2)$	100.0 ± 0.00	0.90	$(0.3, 2^{-4}, 2^4)$	100.0 ± 0.0	1.32
Ecoli	$(0.2, 2^2)$	88.75 ± 4.14	15.75	$(0.4, 2^{-4}, 2^1)$	89.05 ± 4.27	15.04
Glass	$(0.2, 2^0)$	88.32 ± 1.09	17.29	$(0.2, 2^{-4}, 2^{-1})$	92.13 ± 5.28	19.25
Dermatology	$(0.2, 2^4)$	96.06 ± 2.75	40.38	$(0.2, 2^{-2}, 2^2)$	96.65 ± 3.15	68.84

表 1、表 2 分别给出了 Twin-KSVC、K-MTSVM 在线性和非线性最优参数下的实验结果, 包括最优参数, 分类准确率以及训练的 CPU 时间. 从统计结果上看, 两种算法执行效率相当; 线性、非线性 K-MTSVM 都获得了相对于 Twin-KSVC 较高的分类性能, 尤其是在线性情形下, K-MTSVM 算法在数据集 Soybeans 和 Glass 上的分类准确率远远高于 Twin-KSVC, 这暗示了 K-MTSVM 算法良好的泛化性能.

3 结 论

本文在 Twin-KSVC 的基础上, 以结构风险最小化思想来重构原始优化问题, 提出了一种新的多分类 K-MTSVM 算法. 该算法计算复杂度与 Twin-KSVC 同级, 对偶问题为凸二次规划问题, 可以得到全局唯一解, 并避免了 Twin-KSVC 中矩阵奇异问题. 在 UCI 数据集上测试证实: 在线性和非线性情况下, K-MTSVM 算法都获得了更好的分类性能.

参考文献:

- [1] CORTES C, VAPNIK V N. Support Vector Networks [J]. Machine Learning, 1995, 20(3): 273—297.
- [2] JAYADEVA R K, KHEMCHANDANI R, CHANDRA S. Twin Support Vector Machines for Pattern Classification [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(5): 905—910.
- [3] KUMAR M A, GOPAL M. Least Squares Twin Support Vector Machines for Pattern Classification [J]. Expert Systems with Applications, 2009, 36(4): 7535—7543.

- [4] SHAO Y H, ZHANG C H, WANG X B, et al. Improvements on Twin Support Vector Machines [J]. IEEE Transactions on neural networks, 2011, 22(6): 962–968.
- [5] PENG X J. TPMSVM: A Novel Twin Parametric-Margin Support Vector Machine for Pattern Recognition [J]. Pattern Recognition, 2011, 44: 2678–2692.
- [6] 谢娟英, 张兵权, 汪万紫. 基于双支持向量机的偏二叉树多类分类算法 [J]. 南京大学学报: 自然科学版, 2011, 47(4): 354–363.
- [7] ZHOU Z H, LIU X Y. Training Cost-Sensitive Neural Networks with Methods Addressing the Class Imbalance Problem [J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 18: 63–77.

On Maximum Margin Twin Support Vector Machine for Multi-Class Classification

GAO Bin-bin, WANG Jian-jun

School of Mathematics and Statistics, Southwest University, Chongqing 400715, China

Abstract: A novel maximum margin twin support vector machine for multi-class classification (K-MTSVM) has been presented in this paper. The K-MTSVM takes structural risk minimization principle as the optimization objective to build classification model by introducing the margin and uses a 1-versus-1-versus-rest structure to train sub-classifiers. The experimental results on both artificial and UCI datasets indicate that our K-MTSVM gets better generalization performance.

Key words: multi-class classification; twin support vector machines; maximum margin; 1-versus-1-versus-rest; structural risk minimization principle

责任编辑 汤振金