

# 引入调整项的模糊孪生支持向量机

李凯 顾丽凤 胡少方

(河北大学计算机科学与技术学院 保定 071000)

**摘要** 模糊孪生支持向量机是一种重要的机器学习方法,克服了噪声或异常数据对分类的影响;然而,该方法考虑的仍是经验风险,从而使得训练过程易出现过拟合现象。为了解决该问题,通过引入调整项,提出了一种改进的模糊孪生支持向量机模型,利用二次规划求解方法和超松弛迭代法对模型进行求解,获得了用于分类的决策面。实验中选取UCI标准数据集验证了所提方法的有效性。

**关键词** 孪生支持向量机,结构风险,经验风险,模糊隶属度

**中图分类号** TP181 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.08.044

## Regularized Fuzzy Twin Support Vector Machine

LI Kai GU Li-feng HU Shao-fang

(College of Computer Science and Technology, Hebei University, Baoding 071000, China)

**Abstract** Fuzzy twin support vector machine is an important machine learning method and it overcomes the impact of noise and outlier data on classification. However, this method still accomplishes minimization of empirical risk so that overfitting is easily produced in the process of training. In order to solve this problem, a modified fuzzy twin support vector machine model was presented by introducing regularized item. Classifier was obtained by using quadratic programming and over-relaxation method to solve the model. Some UCI datasets were selected to conduct the experiments. The results validates the effectiveness of the proposed method.

**Keywords** Twin support vector machine, Structural risk, Empirical risk, Fuzzy membership

## 1 引言

支持向量机是由 Vapnik 等<sup>[1]</sup>提出的一种机器学习方法,其理论基础是统计学习理论的 VC 维和结构风险最小化原理。此后,研究人员对支持向量机进行了广泛的研究,提出了许多不同的支持向量机<sup>[2-4]</sup>。由于支持向量机只考虑数据集中类间的可分性,忽视了类内数据的可分性信息,因此研究人员将数据的结构信息引入到支持向量机中,提出了结构型支持向量机<sup>[5]</sup>,进一步提高了支持向量机的泛化性能。为了有效地解决噪声或异常数据对支持向量机的影响,研究人员将模糊理论或粗糙理论引入到支持向量机中,提出了模糊支持向量机或粗糙支持向量机<sup>[6-12]</sup>。对于这些支持向量机,为了获得所需要的分类面,通常要求解一个二次规划问题,当数据规模较大时,该二次规划问题的求解是较困难的,且需要大量的计算时间。为了解决该问题,Fung 等<sup>[13]</sup>提出了支持向量机 PSVM,该方法通过求解线性方程组即可获得所需要的平行超平面。之后,Mangasarian 等<sup>[14]</sup>基于 PSVM 且减少了对两个超平面的平行约束,提出了广义特征值的支持向量机算法 GEPSVM。此外,Jayadeva 等<sup>[15]</sup>提出了孪生支持向量机

TWSVM,该方法通过构造两个非平行超平面,使得每类数据尽量分布在其分类面周围。与 GEPSVM 方法不同,TWSVM 通过求解两个形如 SVM 的凸二次规划问题来获得支持向量机分类模型,从而使运行时间缩短为原 SVM 的 1/4。由于孪生支持向量机的平衡因子只能控制经验风险的比例,因此研究人员针对  $\nu$ -SVM 提出了孪生支持向量机  $\nu$ -TSVM 及其改进方法<sup>[16-17]</sup>。Shao 等<sup>[18]</sup>针对 TWSVM,提出了孪生有界支持向量机 TBSVM,该方法在计算时间和分类正确率两方面都优于 TWSVM。TBSVM 与  $\nu$ -TSVM 的不同之处在于它们分别使用了不同的支持向量机,也就是说,TBSVM 使用了传统的 SVM,而  $\nu$ -TSVM 却使用了  $\nu$ -SVM。另外,Li 等<sup>[19]</sup>在  $\nu$ -TSVM 的基础上,通过考虑不同数据样本点对最优超平面的影响,提出了模糊孪生支持向量机,从而降低了噪声或干扰信息对最优分类面产生的影响。然而,该模糊孪生支持向量机仅考虑了经验风险,从而使得噪声或异常数据对分类的影响较大。为此,本文结合孪生支持向量机 TWSVM 和  $\nu$ -TSVM,将模糊方法与调整项引入到对应的优化问题中,对孪生支持向量机进行了改进,从而实现了结构风险最小化,进一步减少了噪声或异常数据对分类的影响。

到稿日期:2016-07-29 返修日期:2016-12-07

李凯(1963-),男,博士,教授,主要研究方向为机器学习、模式识别,E-mail:likai@hbu.edu.cn;顾丽凤(1992-),女,硕士生,主要研究方向为机器学习;胡少方(1990-),女,硕士生,主要研究方向为机器学习。

## 2 模糊孪生支持向量机

设矩阵  $A \in R^{m_1 \times n}$  和  $B \in R^{m_2 \times n}$  分别表示 +1 类和 -1 类的数据样本,其中矩阵  $A$  和  $B$  的每行分别表示一个数据样本,  $m_1$  和  $m_2$  分别表示 +1 类和 -1 类的数据样本的个数,  $n$  是数据样本的维数。模糊孪生支持向量机就是寻找两个非平行超平面:  $f_1(x) = (w^{(1)})^T x + b^{(1)} = 0$ ,  $f_2(x) = (w^{(2)})^T x + b^{(2)} = 0$ 。其中,  $w^{(1)} \in R^n$ ,  $w^{(2)} \in R^n$ ,  $b^{(1)} \in R$  和  $b^{(2)} \in R$ 。

### 2.1 模糊孪生支持向量机 FTSVM

模糊孪生支持向量机 FTSVM 的优化问题为:

$$\min_{w^{(1)}, b^{(1)}, \xi} \frac{1}{2} (Aw^{(1)} + e_1 b^{(1)})^T (Aw^{(1)} + e_1 b^{(1)}) + c_1 S_A e_2^T \xi$$

$$\text{s. t. } -(Bw^{(1)} + e_2 b^{(1)})^T + \xi \geq e_2, \xi \geq 0 \quad (1)$$

$$\min_{w^{(2)}, b^{(2)}, \eta} \frac{1}{2} (Bw^{(2)} + e_2 b^{(2)})^T (Bw^{(2)} + e_2 b^{(2)}) + c_2 S_B e_1^T \eta$$

$$\text{s. t. } (Aw^{(2)} + e_1 b^{(2)})^T + \eta \geq e_1, \eta \geq 0 \quad (2)$$

其中,  $S_A$  和  $S_B$  分别表示每类样本的模糊隶属度,  $e_1$  和  $e_2$  为元素为 1 的向量,  $\xi$  和  $\eta$  为松弛变量构成的向量。

式(1)和式(2)的对偶问题分别为:

$$\max_{\alpha} e_2^T \alpha - \frac{1}{2} \alpha^T G (H^T H)^{-1} G^T \alpha$$

$$\text{s. t. } 0 \leq \alpha \leq c_1 S_A \quad (3)$$

$$\max_{\gamma} e_1^T \gamma - \frac{1}{2} \gamma^T H (G^T G)^{-1} H^T \gamma$$

$$\text{s. t. } 0 \leq \gamma \leq c_2 S_B \quad (4)$$

其中,  $H = [A \ e_1]$ ,  $G = [B \ e_2]$ ;  $\alpha$  和  $\gamma$  分别为由拉格朗日乘子构成的向量。

当数据为非线性可分时,引入核矩阵  $K(x^T, C^T) = \varphi(x^T) \varphi(C^T)$ , 则所求的分类面为  $K(x^T, C^T)u^{(1)} + b^{(1)} = 0$  和  $K(x^T, C^T)u^{(2)} + b^{(2)} = 0$ , 其中  $C^T = [A \ B]^T$ ,  $K(x, y)$  是核函数。此时,求解分类面的优化问题为:

$$\min_{u^{(1)}, b^{(1)}, \xi} \frac{1}{2} \|K(A, C^T)u^{(1)} + e_1 b^{(1)}\|^2 + c_1 S_A e_2^T \xi$$

$$\text{s. t. } -(K(B, C^T)u^{(1)} + e_2 b^{(1)})^T + \xi \geq e_2, \xi \geq 0 \quad (5)$$

$$\min_{u^{(2)}, b^{(2)}, \eta} \frac{1}{2} \|K(B, C^T)u^{(2)} + e_2 b^{(2)}\|^2 + c_2 S_B e_1^T \eta$$

$$\text{s. t. } (K(A, C^T)u^{(2)} + e_1 b^{(2)})^T + \eta \geq e_1, \eta \geq 0 \quad (6)$$

令  $S = [K(A, C^T) \ e_1]$ ,  $R = [K(B, C^T) \ e_2]$ , 其对偶问题为:

$$\max_{\alpha} e_2^T \alpha - \frac{1}{2} \alpha^T R (S^T S)^{-1} R^T \alpha$$

$$\text{s. t. } 0 \leq \alpha \leq c_1 S_A \quad (7)$$

$$\max_{\gamma} e_1^T \gamma - \frac{1}{2} \gamma^T S (R^T R)^{-1} S^T \gamma$$

$$\text{s. t. } 0 \leq \gamma \leq c_2 S_B \quad (8)$$

### 2.2 v 型模糊孪生支持向量机 v-FTSVM

v 型模糊孪生支持向量机 v-FTSVM 的优化问题为:

$$\min_{w^{(1)}, b^{(1)}, \xi, \rho_1} \frac{1}{2} \|Aw^{(1)} + e_1 b^{(1)}\|^2 - v_1 \rho_1 + \frac{1}{l_2} S_B^T \xi_2$$

$$\text{s. t. } -(Bw^{(1)} + e_2 b^{(1)}) \geq \rho_1 - \xi_2$$

$$\xi_2 \geq 0, \rho_1 \geq 0 \quad (9)$$

$$\min_{w^{(2)}, b^{(2)}, \xi_1, \rho_2} \frac{1}{2} \|Bw^{(2)} + e_2 b^{(2)}\|^2 - v_2 \rho_2 + \frac{1}{l_1} S_A^T \xi_1$$

$$\text{s. t. } (Aw^{(2)} + e_1 b^{(2)}) \geq \rho_2 - \xi_1$$

$$\xi_1 \geq 0, \rho_2 \geq 0 \quad (10)$$

其中,  $v_1$  和  $v_2$  为参数,  $l_1$  和  $l_2$  分别为 +1 类和 -1 类样本的个数。

式(9)和式(10)的对偶问题分别为:

$$\max_{\alpha} -\frac{1}{2} \alpha^T G (H^T H)^{-1} G^T \alpha$$

$$\text{s. t. } 0 \leq \alpha \leq \frac{S_B}{l_2}, e_2^T \alpha \geq v_1 \quad (11)$$

$$\max_{\gamma} -\frac{1}{2} \gamma^T H (G^T G)^{-1} H^T \gamma$$

$$\text{s. t. } 0 \leq \gamma \leq \frac{S_A}{l_1}, e_1^T \gamma \geq v_2 \quad (12)$$

当数据为非线性可分时,所求的分类面为  $K(x^T, C^T)u^{(1)} + b^{(1)} = 0$  和  $K(x^T, C^T)u^{(2)} + b^{(2)} = 0$ , 其中  $C^T = [A \ B]^T$ ,  $K(x, y)$  是核函数。其优化问题为:

$$\min_{u^{(1)}, b^{(1)}, \xi_2, \rho_1} \frac{1}{2} \|K(A, C^T)u^{(1)} + e_1 b^{(1)}\|^2 - v_1 \rho_1 + \frac{1}{l_2} S_B^T \xi_2$$

$$\text{s. t. } -(K(B, C^T)u^{(1)} + e_2 b^{(1)}) \geq \rho_1 - \xi_2$$

$$\xi_2 \geq 0, \rho_1 \geq 0 \quad (13)$$

$$\min_{u^{(2)}, b^{(2)}, \xi_1, \rho_2} \frac{1}{2} \|K(B, C^T)u^{(2)} + e_2 b^{(2)}\|^2 - v_2 \rho_2 + \frac{1}{l_1} S_A^T \xi_1$$

$$\text{s. t. } (K(A, C^T)u^{(2)} + e_1 b^{(2)}) \geq \rho_2 - \xi_1$$

$$\xi_1 \geq 0, \rho_2 \geq 0 \quad (14)$$

式(13)和式(14)的对偶问题分别为:

$$\max_{\alpha} -\frac{1}{2} \alpha^T R (S^T S)^{-1} R^T \alpha$$

$$\text{s. t. } 0 \leq \alpha \leq \frac{S_B}{l_2}, e_2^T \alpha \geq v_1 \quad (15)$$

$$\max_{\gamma} -\frac{1}{2} \gamma^T S (R^T R)^{-1} S^T \gamma$$

$$\text{s. t. } 0 \leq \gamma \leq \frac{S_A}{l_1}, e_1^T \gamma \geq v_2 \quad (16)$$

其中,  $S = [K(A, C^T) \ e_1]$ ,  $R = [K(B, C^T) \ e_2]$ 。

## 3 模糊孪生支持向量机的改进方法

### 3.1 线性可分

针对模糊孪生支持向量机 FTSVM, 为了实现结构风险最小化, 在目标函数中引入调整项, 从而获得一种改进的模糊孪生支持向量机 FTBSVM, 其对应的优化问题为:

$$\min_{w^{(1)}, b^{(1)}, \xi, \xi^*} \frac{1}{2} c_3 (\|w^{(1)}\|^2 + b^{(1)2}) + \frac{1}{2} \xi^{*T} \xi^* + c_1 S_A e_2^T \xi$$

$$\text{s. t. } Aw^{(1)} + e_1 b^{(1)} = \xi^*$$

$$-(Bw^{(1)} + e_2 b^{(1)}) + \xi \geq e_2, \xi \geq 0 \quad (17)$$

$$\min_{w^{(2)}, b^{(2)}, \eta, \eta^*} \frac{1}{2} c_4 (\|w^{(2)}\|^2 + b^{(2)2}) + \frac{1}{2} \eta^{*T} \eta^* + c_2 S_B e_1^T \eta$$

$$\text{s. t. } Bw^{(2)} + e_2 b^{(2)} = \eta^*$$

$$Aw^{(2)} + e_1 b^{(2)} + \eta \geq e_1, \eta \geq 0 \quad (18)$$

式(17)与式(18)的对偶问题为:

$$\max_{\alpha} e_2^T \alpha - \frac{1}{2} \alpha^T G(H^T H + c_3 I)^{-1} G^T \alpha \quad (19)$$

$$s. t. 0 \leq \alpha \leq c_1 S_A$$

$$\max_{\gamma} e_1^T \gamma - \frac{1}{2} \gamma^T H(G^T G + c_4 I)^{-1} H^T \gamma \quad (20)$$

$$s. t. 0 \leq \gamma \leq c_2 S_B$$

按照类似的方法,对于  $\nu$  型模糊孪生支持向量机  $\nu$ -FTSVM,通过引入调整项来获得改进的  $\nu$  型模糊孪生支持向量机,将其记为  $\nu$ -FTBSVM,其对应的优化问题为:

$$\min_{w^{(1)}, b^{(1)}, \xi, \xi^*} \frac{1}{2} c_3 (\|w^{(1)}\|^2 + b^{(1)^2}) + \frac{1}{2} \xi^{*T} \xi^* - \nu_1 \rho_1 +$$

$$\frac{1}{l_2} S_B e_2^T \xi$$

$$s. t. A w^{(1)} + e_1 b^{(1)} = \xi^* \\ -(B w^{(1)} + e_2 b^{(1)}) \geq \rho_1 - \xi \quad (21)$$

$$\xi \geq 0, \rho_1 \geq 0$$

$$\min_{w^{(2)}, b^{(2)}, \eta, \eta^*} \frac{1}{2} c_4 (\|w^{(2)}\|^2 + b^{(2)^2}) + \frac{1}{2} \eta^{*T} \eta^* - \nu_2 \rho_2 +$$

$$\frac{1}{l_1} S_A e_1^T \eta$$

$$s. t. B w^{(2)} + e_2 b^{(2)} = \eta^* \\ A w^{(2)} + e_1 b^{(2)} \geq \rho_2 - \eta \quad (22)$$

$$\eta \geq 0, \rho_2 \geq 0$$

式(21)和式(22)的对偶问题分别为:

$$\max_{\alpha} -\frac{1}{2} \alpha^T G(H^T H + c_3 I)^{-1} G^T \alpha \quad (23)$$

$$s. t. 0 \leq \alpha \leq \frac{S_B}{l_2}, e_2^T \alpha \geq \nu_1$$

$$\max_{\gamma} -\frac{1}{2} \gamma^T H(G^T G + c_4 I)^{-1} H^T \gamma \quad (24)$$

$$s. t. 0 \leq \gamma \leq \frac{S_A}{l_1}, e_1^T \gamma \geq \nu_2$$

### 3.2 非线性可分

当数据为非线性可分时,通过引入核函数,将非线性可分问题转化为线性可分问题,所求的分类面为  $K(x^T, C^T)u^{(1)} + b^{(1)} = 0$  和  $K(x^T, C^T)u^{(2)} + b^{(2)} = 0$ ,其中  $C^T = [A \ B]^T$ ,  $K(x, y)$  是核函数,  $c_1 > 0, c_2 > 0, c_3 > 0$  和  $c_4 > 0$ ,因此,相对于 FTBSVM,所要求解的优化问题为:

$$\min_{u^{(1)}, b^{(1)}, \xi, \xi^*} \frac{1}{2} c_3 (\|u^{(1)}\|^2 + b^{(1)^2}) + \frac{1}{2} \xi^{*T} \xi^* + c_1 S_A e_2^T \xi$$

$$s. t. K(A, C^T)u^{(1)} + e_1 b^{(1)} = \xi^* \\ -(K(B, C^T)u^{(1)} + e_2 b^{(1)}) + \xi \geq e_2, \xi \geq 0 \quad (25)$$

$$\min_{u^{(2)}, b^{(2)}, \eta, \eta^*} \frac{1}{2} c_4 (\|u^{(2)}\|^2 + b^{(2)^2}) + \frac{1}{2} \eta^{*T} \eta^* + c_2 S_B e_1^T \eta$$

$$s. t. K(B, C^T)u^{(2)} + e_2 b^{(2)} = \eta^* \\ (K(A, C^T)u^{(2)} + e_1 b^{(2)}) + \eta \geq e_1, \eta \geq 0 \quad (26)$$

令  $S = [K(A, C^T) \ e_1], R = [K(B, C^T) \ e_2]$ ,则式(25)和式(26)的对偶问题为:

$$\max_{\alpha} e_2^T \alpha - \frac{1}{2} \alpha^T R(S^T S + c_3 I)^{-1} R^T \alpha \quad (27)$$

$$s. t. 0 \leq \alpha \leq c_1 S_A$$

$$\max_{\gamma} e_1^T \gamma - \frac{1}{2} \gamma^T S(R^T R + c_4 I)^{-1} S^T \gamma \quad (28)$$

$$s. t. 0 \leq \gamma \leq c_2 S_B$$

对于  $\nu$ -FTBSVM,令  $C^T = [A \ B]^T$ ,并引入核函数  $K(x, y)$ ,则最优分类面为  $K(x^T, C^T)u^{(1)} + b^{(1)} = 0$  和  $K(x^T, C^T)u^{(2)} + b^{(2)} = 0$ ,其对应的优化问题为:

$$\min_{w^{(1)}, b^{(1)}, \xi, \xi^*} \frac{1}{2} c_3 (\|u^{(1)}\|^2 + b^{(1)^2}) + \frac{1}{2} \xi^{*T} \xi^* - \nu_1 \rho_1 +$$

$$\frac{1}{l_2} S_B e_2^T \xi$$

$$s. t. K(A, C^T)u^{(1)} + e_1 b^{(1)} = \xi^* \\ -(K(B, C^T)u^{(1)} + e_2 b^{(1)}) \geq \rho_1 - \xi \quad (29)$$

$$\xi \geq 0, \rho_1 \geq 0$$

$$\min_{w^{(2)}, b^{(2)}, \eta, \eta^*} \frac{1}{2} c_4 (\|u^{(2)}\|^2 + b^{(2)^2}) + \frac{1}{2} \eta^{*T} \eta^* - \nu_2 \rho_2 +$$

$$\frac{1}{l_1} S_A e_1^T \eta$$

$$s. t. K(B, C^T)u^{(2)} + e_2 b^{(2)} = \eta^* \\ (K(A, C^T)u^{(2)} + e_1 b^{(2)}) \geq \rho_2 - \eta \quad (30)$$

$$\eta \geq 0, \rho_2 \geq 0$$

令  $S = [K(A, C^T) \ e_1], R = [K(B, C^T) \ e_2]$ ,则式(29)和式(30)的对偶问题为:

$$\max_{\alpha} -\frac{1}{2} \alpha^T R(S^T S + c_3 I)^{-1} R^T \alpha \quad (31)$$

$$s. t. 0 \leq \alpha \leq \frac{S_B}{l_2}, e_2^T \alpha \geq \nu_1$$

$$\max_{\gamma} -\frac{1}{2} \gamma^T S(R^T R + c_4 I)^{-1} S^T \gamma \quad (32)$$

$$s. t. 0 \leq \gamma \leq \frac{S_A}{l_1}, e_1^T \gamma \geq \nu_2$$

对于上述模糊支持向量机,我们只给出了不同支持向量机所对应的优化问题。由于这些优化问题都是关于拉格朗日乘子的凸二次规划,因此可以直接使用求解凸二次规划问题的方法进行求解,也可以使用超松弛迭代方法求解<sup>[20]</sup>。

### 4 实验研究

为了验证改进的模糊孪生支持向量机的性能,选择 UCI<sup>[21]</sup> 中的两类数据集,如表 1 所列。本文实验采用了十重交叉验证方法,分别在线性可分与非线性可分的情况下进行研究,选择的核函数分别为线性核函数和高斯核函数;为了便于比较,实验选择了模糊孪生支持向量机 FTBSVM 和  $\nu$  型模糊孪生支持向量机  $\nu$ -FTBSVM。

表 1 数据集的特性

数据集	样本数	属性
heart-statlog	270	14
sonar	208	60
ionosphere	351	35
banknote_authentication	1372	6
diabetes	768	9
breast_cancer	683	11
housing	506	15
wholesale_customers	440	9
vertebral_column	310	7
parkinsons	197	24

改进的模糊孪生支持向量机的实验结果如表 2 所列。由表 2 可知,无论是采用线性核函数还是高斯核函数,提出的方法 FTBSVM 和  $\nu$ -FTBSVM 的正确率都优于 FTBSVM 和  $\nu$ -

FTSVM。例如,对于 ionosphere 数据集,在使用线性核函数和高斯核函数的情况下,FTSVM 算法的正确率分别为 74.93%和 92.58%,而使用 FTBSVM 算法的分类正确率分别为 75.14%和 92.64%;使用  $\nu$ -FTSVM 算法的正确率分别为 74.96%和 92.32%,而使用  $\nu$ -FTBSVM 算法的正确率分别为 75.14%和 92.37%。

表 2 不同算法的正确率

数据集	核函数	FTSVM	FTBSVM	$\nu$ -FTSVM	$\nu$ -FTBSVM
heart-statlog	线性核	85.56±2.84	85.60±3.12	85.56±4.45	85.59±4.21
	高斯核	85.65±3.44	85.65±3.44	84.81±4.32	84.81±4.32
sonar	线性核	75.51±2.30	75.55±2.24	77.46±3.14	77.27±3.53
	高斯核	80.61±2.40	81.47±2.54	76.46±2.13	76.59±2.08
ionosphere	线性核	74.93±3.60	75.14±2.36	74.96±4.50	75.14±3.46
	高斯核	92.58±3.10	92.64±2.61	92.32±1.31	92.37±1.26
banknote_authentication	线性核	97.53±0.68	97.53±0.37	97.38±0.67	97.24±0.51
	高斯核	97.89±0.14	97.92±0.11	98.25±0.09	98.30±0.04
diabetes	线性核	64.43±4.07	64.53±3.14	65.09±3.16	65.17±2.59
	高斯核	77.22±4.51	77.24±4.49	65.1±4.05	67.28±3.47
breast_cancer	线性核	65.04±5.84	65.30±4.28	64.99±6.07	65.18±5.27
	高斯核	65.02±5.81	65.24±5.69	65.6±4.17	65.65±3.47
housing	线性核	83.61±1.03	84.08±2.12	81.25±6.50	81.25±6.50
	高斯核	80.75±5.23	82.42±5.51	77.28±1.74	78.19±2.38
wholesale_customers	线性核	85.91±1.60	85.88±1.32	84.31±1.14	83.61±2.17
	高斯核	83.64±0.68	83.67±0.74	84.09±0.91	85.11±0.61
vertebral_column	线性核	80.64±2.00	80.64±2.21	79.35±4.21	80.04±3.76
	高斯核	80.00±1.29	80.23±1.07	78.06±0.97	79.32±2.61
parkinsons	线性核	79.16±1.50	79.25±1.43	72.50±1.00	72.63±0.98
	高斯核	85.50±1.16	85.41±1.39	76.83±1.01	77.17±0.81

另外,针对改进的模糊孪生支持向量机,使用迭代方法对优化问题进行求解。为了验证迭代方法的有效性,分别对 FTSVM 和  $\nu$ -FTSVM 使用传统求解方法和迭代方法进行实验,实验结果如表 3 和表 4 所列。需要注意的是,下文算法中的后缀“-SOR”表示使用迭代法求解的优化问题。

表 3 不同求解方法时 FTBSVM 和 FTSVM 的正确率

数据集	核函数	FTSVM	FTSVM-SOR	FTBSVM	FTBSVM-SOR
heart-statlog	线性核	85.56±2.84	85.92±3.60	85.60±3.12	85.60±3.12
	高斯核	85.65±3.44	85.72±4.32	85.65±3.44	85.75±4.29
sonar	线性核	75.51±2.30	75.51±3.10	75.55±2.24	76.11±2.87
	高斯核	80.61±2.40	82.36±3.67	80.47±2.54	82.71±3.15
ionosphere	线性核	74.93±3.60	78.04±4.25	75.14±2.36	78.21±3.28
	高斯核	92.58±3.10	93.16±2.12	92.64±2.61	93.97±2.06
banknote_authentication	线性核	97.53±0.68	97.74±0.78	97.53±0.37	97.74±0.63
	高斯核	97.89±0.14	98.10±0.37	97.92±0.11	98.45±0.18
diabetes	线性核	64.43±4.07	64.65±7.54	64.53±3.14	66.27±6.23
	高斯核	77.22±4.51	77.60±4.82	77.24±4.49	77.65±0.32
breast_cancer	线性核	65.04±5.84	65.32±7.05	65.30±4.28	65.37±7.10
	高斯核	65.02±5.81	65.72±5.81	65.24±5.69	65.72±5.81
housing	线性核	83.61±1.03	84.18±3.75	84.08±2.12	84.21±3.78
	高斯核	80.75±5.23	85.39±1.73	82.42±5.51	85.50±1.01
wholesale_customers	线性核	85.91±1.60	86.09±0.46	85.88±1.32	86.13±0.50
	高斯核	83.64±0.68	87.72±0.68	83.67±0.74	87.81±0.77
vertebral_column	线性核	80.64±2.00	80.96±1.93	80.64±2.21	80.94±1.87
	高斯核	80.00±1.29	80.35±1.65	80.23±1.07	80.41±1.71
parkinsons	线性核	79.16±1.50	74.16±1.96	79.25±1.43	79.93±1.05
	高斯核	85.50±1.16	85.50±5.83	85.41±1.39	85.50±5.83

由表 3 可以看出,当使用迭代法求解模糊孪生支持向量机与改进的模糊孪生支持向量机中的拉格朗日乘子时,其分类正确率在整体上得到了提升。例如,对于 FTBSVM,当使用线性核函数和高斯核函数时,在 sonar 数据集上,采用传统方法求解得到的正确率分别为 75.55%和 80.47%,而采用迭

代法求解的正确率分别为 76.11%和 82.71%。对于 FTSVM,当使用高斯核和线性核时,在 housing 数据集上,采用传统方法求解得到的正确率分别为 83.61%和 80.75%,而采用迭代法求解的正确率分别为 84.18%和 85.39%。

表 4 不同求解方法的  $\nu$ -FTBSVM 和  $\nu$ -FTSVM 的正确率

数据集	核函数	$\nu$ -FTSVM	$\nu$ -FTSVM-SOR	$\nu$ -FTBSVM	$\nu$ -FTBSVM-SOR
heart-statlog	线性核	85.56±4.45	86.29±4.32	85.59±4.21	85.37±4.24
	高斯核	84.81±4.32	85.18±4.51	84.81±4.32	85.39±4.40
sonar	线性核	77.46±3.14	77.46±4.50	77.27±3.53	77.46±4.50
	高斯核	76.46±2.13	78.27±3.46	76.59±2.08	78.27±3.46
ionosphere	线性核	74.96±4.50	78.38±3.42	75.14±3.46	78.41±3.38
	高斯核	92.32±1.31	92.57±3.20	92.37±1.26	92.60±2.73
banknote_authentication	线性核	97.38±0.67	97.74±0.56	97.24±0.51	97.74±0.63
	高斯核	98.25±0.09	99.05±0.10	98.30±0.04	99.05±0.10
diabetes	线性核	65.09±3.16	76.18±5.48	65.17±2.59	76.37±5.19
	高斯核	65.1±4.05	74.97±5.06	67.28±3.47	75.18±3.84
breast_cancer	线性核	64.99±6.07	65.06±7.64	65.18±5.27	65.32±7.05
	高斯核	65.6±4.17	65.71±6.25	65.65±3.47	65.77±6.19
housing	线性核	81.25±6.50	81.24±6.73	81.25±6.50	82.81±2.33
	高斯核	77.28±1.74	78.46±2.37	78.19±2.38	78.96±1.67
wholesale_customers	线性核	84.31±1.14	85.22±1.36	83.61±2.17	85.83±1.44
	高斯核	84.09±0.91	87.27±0.72	85.11±0.61	87.72±0.68
vertebral_column	线性核	79.35±4.21	80.00±2.26	80.04±3.76	80.96±1.93
	高斯核	78.06±0.97	81.29±2.62	79.32±2.61	81.34±2.11
parkinsons	线性核	72.50±1.00	72.67±2.30	72.63±0.98	74.39±1.51
	高斯核	76.83±1.01	77.00±6.33	77.17±0.81	77.00±6.33

同样地,针对选择的数据集,对  $\nu$ 型模糊孪生支持向量机  $\nu$ -FTSVM 及其改进方法  $\nu$ -FTBSVM 进行了实验研究,结果如表 4 所列。可以看到,对于改进的模糊孪生支持向量机和模糊孪生支持向量机,利用迭代法求解拉格朗日乘子时分类正确率均有一定程度的提升。例如,对于  $\nu$ -FTBSVM,当使用线性核函数和高斯核函数时,在 ionosphere 数据集上,采用传统方法求解的正确率分别为 75.14%和 92.37%,而采用迭代法求解的正确率分别为 78.41%和 92.60%。对于  $\nu$ -FTSVM,当使用线性核函数和高斯核函数时,在 banknote\_authentication 数据集上,采用传统方法求解的正确率分别为 97.38%和 98.25%,而采用迭代法求解的正确率分别为 97.74%和 99.05%。

为了更好地表明不同算法分类时的速度,对 FTBSVM, FTBSVM-SOR,  $\nu$ -FTBSVM 和  $\nu$ -FTBSVM-SOR 等不同算法进行了实验,给出了使用不同算法分类时所用的时间,实验结果如表 5 所列。可以看到,应用迭代法求解比使用二次规划方法求解在时间性能上有了较大的提高。例如,在 heart-statlog 数据集上,应用  $\nu$ -FTBSVM 和 FTBSVM 算法的运行时间分别是 0.23s 和 0.42s,而采用  $\nu$ -FTBSVM-SOR 和 FTBSVM-SOR 算法的运行时间分别是 0.12s 和 0.19s。

为了从总体上评价不同算法的时间性能,针对所有选择的数据集,给出了不同算法的平均运行时间,如图 1 所示。可以看出,FTSVM,  $\nu$ -FTSVM, FTBSVM 和  $\nu$ -FTBSVM 算法的运行时间分别是 FTSVM-SOR,  $\nu$ -FTSVM-SOR, FTBSVM-SOR 和  $\nu$ -FTBSVM-SOR 算法运行时间的 2.5 倍左右,表明了基于迭代法的模糊孪生支持向量机降低了计算的时间复杂度,进一步提高了模糊孪生支持向量机的时间性能。

表5 不同算法的运行时间比较

数据集	时间/s							
	$\nu$ -FTBSVM	FTBSVM	$\nu$ -FTBSVM-SOR	FTBSVM-SOR	$\nu$ -FTSVM	FTSVM	$\nu$ -FTSVM-SOR	FTSVM-SOR
heart-statlog	0.23	0.42	0.12	0.19	0.23	0.42	0.13	0.19
sonar	0.22	0.37	0.07	0.18	0.24	0.39	0.06	0.18
ionosphere	0.7	0.68	0.25	0.25	0.72	0.71	0.25	0.28
banknote_authentication	79.26	78.27	30.72	30.39	78.43	77.96	31.43	30.65
diabetes	11.22	11.37	5.72	6.2	11.47	12.5	5.72	6.2
breast_cancer	6.03	6.68	1.73	1.74	6.03	6.68	1.73	1.74
housing	3.71	3.15	1.1	1.1	3.71	3.43	1.11	1.16
wholesale_customers	2.28	2.21	0.42	0.42	2.25	2.2	0.43	0.44
vertebral_column	0.92	0.92	0.19	0.18	0.9	0.9	0.19	0.18
parkinsons	0.24	0.32	0.05	0.05	0.24	0.32	0.05	0.05

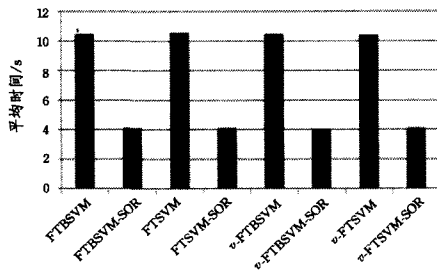


图1 不同算法的平均运行时间

**结束语** 为了实现结构风险最小化,通过引入调整项来获得一种改进的模糊孪生支持向量机模型,利用二次规划和超松弛迭代法求解对偶问题,提出了一种改进的模糊孪生支持向量机。通过选取 UCI 标准数据集进行实验,结果验证了所提方法的性能。另外,选取模糊孪生支持向量机进行实验比较,改进的模糊支持向量机的正确率都有了一定程度的提升,表明改进的模糊孪生支持向量机的分类性能高于模糊孪生支持向量机的分类性能。

### 参考文献

- [1] VAPNIK V N. The nature of statistical learning theory[M]. New York:Springer,1995.
- [2] MANGASARIAN O L, MUSICANT D R. Lagrangian support vector machines[J]. Journal of Machine Learning Research, 2001,1(3):161-177.
- [3] SCHOLKOPF B, SMOLA A J, WILLIAMSON R C, et al. New support vector algorithms[J]. Neural Computation, 2000, 12(5):1207-1245.
- [4] BLOOM V, GRIVA I, KWON B, et al. Exterior-point method for support vector machines[J]. IEEE Transactions on Neural Networks and Learning Systems, 2014, 25(7):1390-1393.
- [5] CHEN D D, TIAN Y J, LIU X H. Structural nonparallel support vector machine for pattern recognition[J]. Pattern Recognition, 2016, 60:296-305.
- [6] LIN C F, WANG S D. Fuzzy support vector machines[J]. IEEE Transaction on Neural Networks, 2002, 13(2):464-471.
- [7] WANG Y Q, WANG S Y, LAI K K. A new fuzzy support vector machine to evaluate credit risk[J]. IEEE Transaction on Fuzzy System, 2005, 13(6):820-831
- [8] YANG X W, ZHANG G Q, LU J, et al. A kernel fuzzy c-means clustering-based fuzzy support vector machine algorithm for classification problems with outliers or noises[J]. IEEE Transactions on Fuzzy Systems, 2011, 19(1):105-115.
- [9] ZHANG J H, WANG Y Y. A rough margin based support vector machine[J]. Information Sciences, 2008, 178(9):2204-2214.
- [10] XU Y T. A rough margin-based linear  $\nu$  support vector regression[J]. Statistics and Probability Letters, 2012, 82(3):528-534.
- [11] CHEN D G, HE Q, WANG X Z. FRSVMs: Fuzzy rough set based support vector machines[J]. Fuzzy Sets and Systems, 2010, 161(4):596-607.
- [12] SHIGEO A. Fuzzy support vector machines for multilabel classification[J]. Pattern Recognition, 2015, 48(6):2110-2117.
- [13] FUNG G, MANGASARIAN O L. Proximal support vector machine classifiers[C]//Proceedings of the 7th International Conference on Knowledge and Data Discovery. 2001:77-86.
- [14] MANGASARIAN O L, WILD E W. Multisurface proximal support vector classification via generalized eigenvalues[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(1):69-74.
- [15] JAYADEVA R K, KHEMCHANDANI R, CHANDRA S. Twin support vector machine for pattern classification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(5):905-910.
- [16] PENG X J. A  $\nu$ -twin support vector machine( $\nu$ -TSVM) classifier and its geometric algorithms[J]. Information Sciences, 2010, 180(20):3863-3875.
- [17] RESHMA K, POOJA S, SURESH C. Improvements on  $\nu$ -twin support vector machine[J]. Neural Networks, 2016, 79:97-107.
- [18] 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.
- [19] LI K, MA H Y. A Fuzzy Twin Support Vector Machine Algorithm[J]. International Journal of Application or Innovation in Engineering & Management, 2013, 2(3):459-465.
- [20] MANGASARIAN O L, MUSICANT D R. Successive overrelaxation for support vector machines[J]. IEEE Transactions on Neural Networks, 1999, 10(5):1032-1037.
- [21] BLAKE C, MERZ C J. UCI Repository for machine learning databases[EB/OL]. (1998-01-12) [2014-09-06]. IrvineCA: University of California, Department of Information and Computer Sciences. <http://www.ics.uci.edu/mllearn/MLRepository.html>.