文章编号: 1001-0920 (2016) 05-0949-04 DOI: 10.13195/j.kzyjc.2015.0245

一种新的孪生大间隔分布机算法

程昊翔, 王 坚

(同济大学 a. 电子与信息工程学院，b. CIMS 研究中心，上海 201804)

摘 要: 为了提高孪生支持向量机的泛化能力, 提出一种新的孪生大间隔分布机算法, 以增加间隔分布对于训练模型的影响. 理论研究表明, 间隔分布对于模型的泛化性能有着非常重要的影响. 该算法在标准孪生支持向量机优化目标函数上增加了间隔分布的影响, 间隔分布通过一阶和二阶数据统计特征来体现. 在标准数据集上的实验结果表明, 所提出的算法比 SVM、TWSVM、TBSVM 算法的分类精确度更高.

关键词: 泛化；孪生支持向量机；间隔分布

中图分类号: TP273 **文献标志码:** A

A novel twin large margin distribution machine

CHENG Hao-xiang, *WANG Jian*

(a. College of Electronics and Information Engineering，b. CIMS Research Center，Tongji University，Shanghai 201804, China．Correspondent：CHENG Hao-xiang，E-mail：373668304@qq.com)

Abstract: In order to improve the generalization ability of the twin support vector machine(TWSVM), a novel twin large margin distribution machine(TLDM) which increases the impact of the margin distribution on the training model is proposed. Theoretical studies show that margin distribution has important influence on the generalization performance of the model. The proposed approach based on the standard twin support vector machine adds the affection of margin distribution to the optimization objective function. The margin distribution is characterized by first order and second order statistics. The experimental results based on benchmark data sets show that the proposed approach has better classification accuracy than other three algorithms including SVM, TWSVM and TBSVM.

Keywords: generalization；twin support vector machine；margin distribution

0 引 言

支持向量机 (SVM)[1]由 Vapnik 等在 1995 年提出, 已经成为最受欢迎的机器学习算法之一, 并被广泛应用在分类和回归领域. 支持向量机算法基于结构风险最小化原则和 VC 维理论[2-3], 是拥有很好泛化能力的机器学习算法. SVM 通过最大化两个分类之间的最小间隔来获取分类超平面. SVM 的算法复杂度为 $O(l^3)$, l 是训练样本总数. SVM 的优点总结如下: 1) SVM 基于结构风险最小化理论, 拥有很好的泛化能力; 2) SVM 求解的是凸二次优化问题, 能够获得全局最优解. 通过引入核函数, SVM 能够有效地解决非线性分类问题.

Jayadeva 提出了孪生支持向量机 (TWSVM)[5], 其灵感来自于广义特征值近似支持向量机 (GEPSVM)[4]. 相比于 SVM 求解一个二次规划问题, TWSVM 求解两个规模较小的二次规划问题来获得两个非平行的超平面. TSVM 的模型学习速度相比于标准 SVM 约提高了 4 倍. 通过在标准数据集上的实验结果表明[5], 相比于标准 SVM 和 GEPSVM 算法, TWSVM 在性能上有很好的提升. Shao 提出了限定双子支持向量机 (TBSVM)[6], 是对标准 TWSVM 进行了扩展, 将目标函数正则化, 从经验风险最小化到结构风险最小化, 实验表明该算法较 TWSVM 有更高的分类精度. 近几年, 对于 TWSVM 还有很多不同的扩展研究[6-8], 但这些算法都没有考虑到间隔分布对于 TWSVM 训练模型的影响. Gao 等[9]证明了边缘分布对于算法的泛化性能有着非常重要影响, 且间隔分布通过其一阶和二阶统计属性来描述. 本文中, 间隔分布采用间隔均值和间隔方差来表现.

本文提出一种孪生大间隔分布机算法 (TLDM),

收稿日期: 2015-03-01；**修回日期:** 2015-07-29.

基金项目: 国家自然科学基金面上项目(71273188)；国家自然科学基金重大项目(91024031).

作者简介: 程昊翔(1986–), 男, 博士, 从事系统工程、机器学习的研究；王坚(1961–), 男, 教授, 博士生导师, 从事系统工程、智能优化算法等研究.

该算法通过优化间隔分布来获得具有更强泛化能力的模型. 所提出的TLDM算法在TWSVM目标函数的基础上增加了间隔分布对训练模型的影响, 通过同时最大化间隔均值以及最小化间隔方差来获得新的优化目标. 选取了7个UCI数据集进行了实验, 分析表明, TLDM算法相比于SVM、TWSVM和TBSVM, 在分类准确度上有很大程度的提高, 验证了本文所提出算法的有效性.

1 孪生支持向量机

对于分类问题, 假设数据集T为

$$T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_l,y_l)\}\in(R^n\times Y)^l.$$

其中: $x_i\in R^n$, $y_i\in\{-1,+1\}$, $i=1,2,\cdots,l$, l是数据集的大小, n是输入空间的维度. 对于数据集T, l_1是$y_i=+1$的总个数, l_2是$y_i=-1$的总个数, $l_1+l_2=l$. $A\in R^{l_1\times n}$表示所有$y_i=+1$的数据点组成的矩阵, $B\in R^{l_2\times n}$表示所有$y_i=-1$的数据点组成的矩阵. 矩阵A和B的每一行代表一个数据点. 矩阵$X\in R^{l\times n}$表示数据集T.

孪生支持向量机的目标是通过学习获得如下两个非平行的超平面:

$$f_1(x)=w_1^{\mathrm{T}}x+b_1=0,$$
$$f_2(x)=w_2^{\mathrm{T}}x+b_2=0.$$

其中: $w_1,w_2\in R^n$, $b_1,b_2\in R$.

孪生支持向量机求解如下两个二次规划问题:

$$\min_{w_1,b_1,\xi}\ \frac{1}{2}(Aw_1+b_1)^{\mathrm{T}}(Aw_1+b_1)+c_1e_2^{\mathrm{T}}\xi,$$
$$\text{s.t.}-(Bw_1+e_2b_1)+\xi\geqslant e_2,\ \xi\geqslant 0;\tag{1}$$
$$\min_{w_2,b_2,\eta}\ \frac{1}{2}(Bw_2+b_2)^{\mathrm{T}}(Bw_2+b_2)+c_2e_1^{\mathrm{T}}\eta,$$
$$\text{s.t. }Aw_2+e_1b_2+\eta\geqslant e_1,\ \eta\geqslant 0.\tag{2}$$

引入拉格朗日乘子, 将问题(1)和(2)转化为如下对偶问题:

$$\max_{\alpha}\ e_2^{\mathrm{T}}\alpha-\frac{1}{2}\alpha^{\mathrm{T}}G(H^{\mathrm{T}}H)^{-1}G^{\mathrm{T}}\alpha,$$
$$\text{s.t. }0\leqslant\alpha\leqslant c_1;\tag{3}$$
$$\max_{\gamma}\ e_1^{\mathrm{T}}\gamma-\frac{1}{2}\gamma^{\mathrm{T}}H(G^{\mathrm{T}}G)^{-1}H^{\mathrm{T}}\gamma,$$
$$\text{s.t. }0\leqslant\gamma\leqslant c_2.\tag{4}$$

其中: $G=[B\quad e_2]$, $H=[A\quad e_1]$. 为了避免奇异矩阵, 将上述问题(1)和(2)中的$(H^{\mathrm{T}}H)^{-1}$和$(G^{\mathrm{T}}G)^{-1}$替换为$(H^{\mathrm{T}}H+\varepsilon)^{-1}$和$(G^{\mathrm{T}}G+\varepsilon)^{-1}$. 这里, 取$\varepsilon=10^{-6}$.

通过Karush-Kuhn-Tucker(KKT)条件, 可以得到下面两个等式:

$$v_1=-(H^{\mathrm{T}}H+\varepsilon I)^{-1}G^{\mathrm{T}}\alpha,$$
$$v_2=(G^{\mathrm{T}}G+\varepsilon I)^{-1}H^{\mathrm{T}}\gamma.$$

其中: $v_1=[w_1\quad b_1]$, $v_2=[w_2\quad b_2]$. 通过求解问题(1)和(2)获得α和γ, 能够得到预测模型. 对于新输入$x\in R^n$, 通过$i=\arg\min\limits_{k=1,2}\dfrac{|w_k^{\mathrm{T}}x+b_k|}{|w_k|}$来判断属于哪一类.

引入核函数K, 可以将TWSVM扩展到解决非线性分类问题. 两个超平面变为如下的形式:

$$f_1(x)=w_1^{\mathrm{T}}K(x^{\mathrm{T}},X^{\mathrm{T}})+b_1=0,$$
$$f_2(x)=w_2^{\mathrm{T}}K(x^{\mathrm{T}},X^{\mathrm{T}})+b_2=0.$$

对于非线性分类问题, TWSVM解决如下两个二次规划问题:

$$\min_{w_1,b_1,\xi}\ \frac{1}{2}(K(A,X^{\mathrm{T}})w_1+b_1)^{\mathrm{T}}(K(A,X^{\mathrm{T}})w_1+b_1)+c_1e_2^{\mathrm{T}}\xi,$$
$$\text{s.t. }-(K(B,X^{\mathrm{T}})w_1+e_2b_1)+\xi\geqslant e_2,\ \xi\geqslant 0;\tag{5}$$
$$\min_{w_2,b_2,\eta}\ \frac{1}{2}(K(B,X^{\mathrm{T}})w_2+b_2)^{\mathrm{T}}(K(B,X^{\mathrm{T}})w_2+b_2)+c_2e_1^{\mathrm{T}}\eta,$$
$$\text{s.t. }K(A,X^{\mathrm{T}})w_2+e_1b_2+\eta\geqslant e_1,\ \eta\geqslant 0.\tag{6}$$

同样, 可以得到二次规划问题(5)和(6)的对偶表示

$$\max_{\alpha}\ e_2^{\mathrm{T}}\alpha-\frac{1}{2}\alpha^{\mathrm{T}}G(H^{\mathrm{T}}H)^{-1}G^{\mathrm{T}}\alpha,$$
$$\text{s.t. }0\leqslant\alpha\leqslant c_1;\tag{7}$$
$$\max_{\gamma}\ e_1^{\mathrm{T}}\gamma-\frac{1}{2}\gamma^{\mathrm{T}}H(G^{\mathrm{T}}G)^{-1}H^{\mathrm{T}}\gamma,$$
$$\text{s.t. }0\leqslant\gamma\leqslant c_2.\tag{8}$$

其中: $G=[K(B,X^{\mathrm{T}})\quad e_2]$, $H=[K(A,X^{\mathrm{T}})\quad e_1]$, K为高斯核函数.

通过求解问题(7)和(8), 可得到v_1和v_2. 这样, 对于新输入的$x\in R^n$, 可通过

$$i=\arg\min_{k=1,2}\frac{|w_k^{\mathrm{T}}K(x^{\mathrm{T}},X^{\mathrm{T}})+b_k|}{\sqrt{w_k^{\mathrm{T}}K(X,X^{\mathrm{T}})w_k}}$$

来判断属于哪一类.

2 一种新的孪生大间隔分布机算法

下面提出一种孪生大间隔分布机算法(TLDM), 该算法增加了间隔分布对于训练模型的影响. 理论研究表明[9], 间隔分布对于模型的影响通过一阶和二阶统计特性来表现. 本文中, 采用了间隔均值和间隔方差来代表这两个统计特性.

通过文献[3], 每个训练数据(x_i,y_i)所对应的几何间隔为$\mu_i=y_i(w^{\mathrm{T}}x_i+b)$. 可以计算间隔均值和间隔方差. Y为$l\times 1$的列矩阵, 表示样本输出值组成的矩阵; e为所有元素为1的$l\times 1$列矩阵.

考虑f_1以及$v_1=[w_1\quad b_1]$情况下, 可以获得如下的间隔均值$\bar{\mu}_1$和间隔方差$\hat{\mu}_1$:

$$\bar{\mu}_1 = \frac{1}{l}Y^{\mathrm{T}}(Xw_1 + b_1e),$$
$$\hat{\mu}_1 = \frac{1}{l^2}[l(Xw_1 + b_1e)^{\mathrm{T}}(Xw_1 + b_1e) - (Xw_1 + b_1e)^{\mathrm{T}}YY^{\mathrm{T}}(Xw_1 + b_1e)].$$

同样, 对于 f_2 以及 $v_2 = [w_2 \quad b_2]$, 可以计算获得间隔均值 $\bar{\mu}_2$ 和间隔方差 $\hat{\mu}_2$ 为

$$\bar{\mu}_2 = \frac{1}{l}Y^{\mathrm{T}}(Xw_2 + b_2e),$$
$$\hat{\mu}_2 = \frac{1}{l^2}[l(Xw_2 + b_2e)^{\mathrm{T}}(Xw_2 + b_2e) - (Xw_2 + b_2e)^{\mathrm{T}}YY^{\mathrm{T}}(Xw_2 + b_2e)].$$

受最近理论研究的启发, TLDM 算法可在两个优化问题目标函数上同时最小化间隔方差和最大化间隔均值. 结合 TWSVM 优化目标, 扩展问题 (1) 和 (2) 为如下二次规划问题:

$$\min_{w_1,b_1,\xi} \frac{\lambda_1}{2}\hat{\mu}_1 - \lambda_2\bar{\mu}_1 + c_1e_2^{\mathrm{T}}\xi + \frac{1}{2}(Aw_1 + b_1)^{\mathrm{T}}(Aw_1 + b_1),$$
$$\text{s.t. } -(Bw_1 + e_2b_1) + \xi \geqslant e_2,\ \xi \geqslant 0; \tag{9}$$

$$\min_{w_2,b_2,\eta} \frac{\lambda_3}{2}\hat{\mu}_2 - \lambda_4\bar{\mu}_2 + c_2e_1^{\mathrm{T}}\eta + \frac{1}{2}(Bw_2 + b_2)^{\mathrm{T}}(Bw_2 + b_2),$$
$$\text{s.t. } Aw_2 + e_1b_2 + \eta \geqslant e_1,\ \eta \geqslant 0. \tag{10}$$

引入拉格朗日乘子, 获得问题 (9) 和 (10) 的对偶表示

$$\max_{\alpha} (e_2^{\mathrm{T}} - Q^{\mathrm{T}}P^{-1}G^{\mathrm{T}})\alpha - \frac{1}{2}\alpha^{\mathrm{T}}GP^{-1}G^{\mathrm{T}}\alpha,$$
$$\text{s.t. } 0 \leqslant \alpha \leqslant c_1; \tag{11}$$

$$\max_{\gamma} (e_1^{\mathrm{T}} - Q^{\mathrm{T}}S^{-1}H^{\mathrm{T}})\gamma - \frac{1}{2}\gamma^{\mathrm{T}}HS^{-1}H^{\mathrm{T}}\gamma,$$
$$\text{s.t. } 0 \leqslant \gamma \leqslant c_2. \tag{12}$$

其中

$$M = [X \quad e],\ D = M^{\mathrm{T}}(l - YY^{\mathrm{T}}),$$
$$Q = \frac{\lambda_2}{l}M^{\mathrm{T}}Y,\ P = \frac{\lambda_1}{l^2}DM + H^{\mathrm{T}}H,$$
$$S = \frac{\lambda_1}{l^2}DM + G^{\mathrm{T}}G.$$

这里, 设定 $\lambda_1 = \lambda_3, \lambda_2 = \lambda_4$.

通过求解问题 (11) 和 (12) 可以得到 α 和 γ, 则

$$v_1 = P^{-1}(Q - G^{\mathrm{T}}\alpha),$$
$$v_2 = S^{-1}(Q - H^{\mathrm{T}}\gamma).$$

将 TLDM 扩展到解决非线性分类问题, 引入核函数 K, 获得如下需要求解的二次规划问题:

$$\min_{w_1,b_1,\xi} \frac{\lambda_1}{2}\hat{\mu}_1 - \lambda_2\bar{\mu}_1 + c_1e_2^{\mathrm{T}}\xi + \frac{1}{2}(K(A, X^{\mathrm{T}})w_1 + b_1)^{\mathrm{T}}(K(A, X^{\mathrm{T}})w_1 + b_1),$$
$$\text{s.t. } -(K(B, X^{\mathrm{T}})w_1 + e_2b_1) + \xi \geqslant e_2,\ \xi \geqslant 0; \tag{13}$$

$$\min_{w_2,b_2,\eta} \frac{\lambda_3}{2}\hat{\mu}_2 - \lambda_4\bar{\mu}_2 + c_2e_1^{\mathrm{T}}\eta + \frac{1}{2}(K(B, X^{\mathrm{T}})w_2 + b_2)^{\mathrm{T}}(K(B, X^{\mathrm{T}})w_2 + b_2),$$
$$\text{s.t. } K(A, X^{\mathrm{T}})w_2 + e_1b_2 + \eta \geqslant e_1,\ \eta \geqslant 0. \tag{14}$$

同样, 可以获得二次规划问题 (13) 和 (14) 的对偶问题

$$\max_{\alpha} (e_2^{\mathrm{T}} - Q^{\mathrm{T}}P^{-1}G^{\mathrm{T}})\alpha - \frac{1}{2}\alpha^{\mathrm{T}}GP^{-1}G^{\mathrm{T}}\alpha,$$
$$\text{s.t. } 0 \leqslant \alpha \leqslant c_1; \tag{15}$$

$$\max_{\gamma} (e_1^{\mathrm{T}} - Q^{\mathrm{T}}S^{-1}H^{\mathrm{T}})\gamma - \frac{1}{2}\gamma^{\mathrm{T}}HS^{-1}H^{\mathrm{T}}\gamma,$$
$$\text{s.t. } 0 \leqslant \gamma \leqslant c_2. \tag{16}$$

其中: $G = [K(B, X^{\mathrm{T}}) \quad e_2]$, $H = [K(A, X^{\mathrm{T}}) \quad e_1]$, K 是高斯核函数.

通过求解问题 (15) 和 (16), 可以得到

$$v_1 = [w_1 \quad b_1],\ v_2 = [w_2 \quad b_2].$$

对于新输入 $x \in R^n$, 可通过

$$i = \arg\min_{k=1,2} \frac{|w_k^{\mathrm{T}}K(x^{\mathrm{T}}, X^{\mathrm{T}}) + b_k|}{\sqrt{w_k^{\mathrm{T}}K(X, X^{\mathrm{T}})w_k}}$$

来判断属于哪一类.

3　实验分析

为了验证本文所提出的 TLDM 算法的有效性, 在 7 个 UCI 数据集上与 SVM、TWSVM 和 TBSVM 算法进行实验比较. 表 1 给出了 7 个 UCI 数据集的属性介绍.

表 1　数据集属性

数据集名称	数据量	特征维度
Australian	690	14
CMC	1 473	9
Pima-Indian	768	8
Sonar	208	60
Heart-Statlog	270	14
Votes	435	16
BUPA liver	345	6

所有的算法都在 Matlab R2012b 上实现, SVM 在 LibSVM 工具箱[10]上实现. 采用逐次超松弛技术对 TWSVM、TBSVM 和 TLDM 算法中的 QPP 问题进行求解. 通过分析可以获得, SVM 的算法复杂度为 $O(m^3)$, TSVM、TBSVM 和 TLDM 的算法复杂度为 $O(m^3/4)$, m 为训练集样本总数, 这里假设正负类的训练样本数相同.

为了评估比较 4 个算法的性能, 本文采用如下的 Accuracy 计算公式进行评价:

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{FP} + \text{TN} + \text{FN}).$$

其中: TP 为正类预测预测为正类数; TN 为负类预测预测为负类数; FP 为负类预测预测为正类数; FN 为负类预测预测为负类数. Accuracy 通过 10 次实验计算平均值获得.

本文采用5折交叉验证算法进行参数的选择. 文中采用高斯核函数$K(x_i,x_j)=\exp(-q\|x_i-x_j\|^2)$, q从$\{2^i|i=-8,-7,\cdots,8\}$中选取. 设定$c_1=c_2, c_3=c_4, \lambda_1=\lambda_3, \lambda_2=\lambda_4$. 对于TLDM算法, 参数$\lambda_1$和$\lambda_2$从$\{2^i|i=-8,-7,\cdots,8\}$中选取. 对于TWSVM、TBSVM和TLDM算法, 参数c_1、c_2、c_3和c_4从$\{2^i|i=-8,-7,\cdots,8\}$中选取. 对于SVM算法, 参数C从$\{2^i|i=-8,-7,\cdots,8\}$中选取. 表2给出了4个算法的性能比较.

表 2 算法性能比较(Accuracy) %

数据集名称	SVM	TWSVM	TBSVM	TLDM
Australian	88.62	86.13	87.20	88.92
CMC	66.98	68.10	73.12	73.80
Pima-Indian	78.32	76.68	77.10	78.24
Sonar	80.10	77.02	79.11	78.33
Heart-Statlog	83.22	84.90	85.04	85.68
Votes	95.75	95.85	96.33	97.02
BUPA liver	68.12	66.68	70.20	70.98

表2的实验结果表明, 在除了Sonar外的6个UCI标准数据集上, 本文所提出的TLDM算法相比于SVM、TWSVM以及TBSVM算法有更好的性能. TLDM在4个算法中有最高的Accuracy. 实验结果表明, 间隔分布通过间隔均值和间隔方差对孪生支持向量机算法有着相当重要的影响. 综上所述, 本文所提出的TLDM算法表现出了更好的泛化性, 验证了该算法的有效性.

4 结 论

本文提出了一种新的孪生大间隔分布机算法(TLDM), 该算法增加了间隔分布对于训练模型的影响. 最近的理论研究表明, 间隔分布对于模型的泛化性能有着非常重要的影响. 间隔分布通过间隔均值和间隔方差两个数据特性来体现. TLDM算法在标准TWSVM算法优化目标函数的基础上同时最大化间隔均值并最小化间隔方差. 此外, 该算法采用了连续超松弛算法进行了优化问题的求解, 从一定程度上提高了训练的速度. 选取了7个UCI数据集, 将本文提出的算法与SVM、TWSVM和TBSVM算法进行实验. 实验结果表明, 本文提出的TLDM算法具有更好的泛化性能.

参考文献(References)

[1] Cortes C, Vapnik V. Support vector networks[J]. Machine Learning, 1995, 20(3): 273-297.

[2] Vapnik V N. The nature of statistical learning theory[M]. New York: Springer-Verlag, 1995: 93-110.

[3] Vapnik V N. Statistical learning theory[M]. New York: Wiley, 1998: 145-254.

[4] Mangasarian O L, Wild E W. Multisurface proximal support vector classification via generalized eigenvalues[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2006, 28(1): 69-74.

[5] Jayadeva, Khemchandani R, Chandra S. Twin support vector machines for pattern classification[J]. IEEE Trans on Pattern Analysis and Machine Intellegence, 2007, 29(5): 905-910.

[6] Yuan-Hai Shao, Chun-Hua Zhang, Xiao-Bo Wang, et al. Improvements on twin support vector mahines[J]. IEEE trans on Neural Networks, 2011, 22(6): 962-968.

[7] Qi Zhiquan, Tian Yingjie, Shi Yong. Robust twin support vector machine for pattern classification[J]. Pattern Recognition, 2013, 46(1): 305-316.

[8] Kumar M A, Gopal M. Least squares twin support vector machines for pattern classification[J]. Expert Systems with Applications, 2009, 36(4): 7535-7543.

[9] Gao Wei, Zhou ZhiHua. On the doubt about margin explanation of boosting[J]. Artificial Intelligence, 2013, 203(5): 1-18.

[10] Chih-Chung Chang, Chih-Jen Lin. LIBSVM: A library for support vector machines[J]. ACM Trans on Intelligent Systems and Technology, 2011, 2(27): 1-27.

(责任编辑：齐 霁)