一种改进的最小二乘孪生支持向量机分类算法

储茂祥[1,2],王安娜[1],巩荣芬[1,2]
(1.东北大学信息科学与工程学院,辽宁沈阳 110819;2.辽宁科技大学电子与信息工程学院,辽宁鞍山 114051)

摘 要: 提出了一种新的模式分类器,即广泛权重的最小二乘孪生支持向量机.该支持向量机在正、负两类样本上广泛地增加权重,很好地抑制了交叉噪声样本对数据分类的影响.其次,根据间隔最大化原理,该支持向量机在目标函数上增加了一个正规化项,实现结构风险最小化和避免在求解该目标函数时可能对病态矩阵求逆的处理.同时,提出了利用一种指数函数计算训练样本的密度来获得样本权重值的算法.该算法能够有效缩减计算权重的时间,且具有较强的鲁棒性.实验证明本文提出的广泛权重的最小二乘孪生支持向量机能够实现高精度和高效率的分类效果,而且特别适合于含有交叉噪声样本的数据集分类.

关键词: 模式分类;最小二乘;孪生支持向量机;权重;指数函数
中图分类号: TP18 **文献标识码:** A **文章编号:** 0372-2112 (2014)05-0998-06
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2014.05.026

Improvement on Least Squares Twin Support Vector Machine for Pattern Classification

CHU Mao-xiang[1,2], WANG An-na[1], GONG Rong-fen[1,2]
(1. *College of Information Science and Engineering, Northeastern University, Shenyang, Liaoning* 110819, *China*;
2. *School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan, Liaoning* 114051, *China*)

Abstract: Widely weighted least squares twin support vector machine (WWLSTSVM) is proposed for pattern classification. In WWLSTSVM, weights are widely added on error variables of data samples both in one class and the other. This widely weighted method is especially effective on eliminating the interference of intercrossing noise samples. Moreover, a regularization term is added with the theory of maximizing margin, in which the structural risk is minimized and the possible ill-conditioning is avoided for matrix inversion. Also, an effective weight algorithm with exponential function is proposed to reduce the time complexity of computing weight values and enhance its robustness for cross plane dataset. Comparative experiments show that WWLSTSVM obtains better results on eliminating the interference of noise samples and higher classification accuracy with less computing time in both linear and nonlinear cases compared with the other classifiers.

Key words: pattern classification; least squares; twin support vector machine; weight; exponential function

1 引言

标准的支持向量机(Support Vector Machine, SVM)[1]由Vapnik等人提出,现已广泛应用于模式分类和回归分析等领域[2~4].标准的支持向量机以结构风险最小化为准则,且遵循间隔最大化的思想,其目标是构建一个满足要求的最优分类超平面.随着技术的发展,支持向量机又衍生出新的算法.2007年,Javadeva等人提出了孪生支持向量机(Twin SVM, TSVM)[5].TSVM寻求两个非平行的最优分类面,使得每一个分类面靠近一类样本而远离其它类样本.TSVM适合于交叉分类面数据集的分类,且求解两个相对更小的二次规划问题(Quadratic Programming Problem, QPP),这使得TSVM的速度明显快于标准的支持向量机.邵元海等人在TSVM的基础上提出了孪生边界支持向量机(Twin Bound SVM, TBSVM)[6],它通过在目标函数上增加正规化项来实现结构风险最小化.其他的研究学者又提出了其它改进的TSVM[7~9],这些支持向量机在鲁棒性、有效性和泛化性等方面都有不同程度的提高.

2009年,Kumar和Gopal在TSVM基础上提出了最小二乘孪生支持向量机(Least Squares TSVM, LSTSVM)[10].LSTSVM与TSVM一样适合于交叉分类面

收稿日期:2013-02-26;修回日期:2013-08-21;责任编辑:赵克
基金项目:国家自然科学基金(No.61050006)

数据集的分类.LSTSVM 用近似支持向量机(Proximal SVM,PSVM)的思想来求解两个线性方程,而不是求解 QPPs,这使得 LSTSVM 比 TSVM 的求解速度更快.基于 LSTSVM,陈静等学者提出了加权的最小二乘孪生支持向量机(Weighted LSTSVM, WLSTSVM)[11].WLSTSVM 通过对数据样本增加不同的权重值,来提高算法的鲁棒性.但是,WLSTSVM 对于每一个 QPP 而言,仅对其中一类数据样本的误差量增加了权重,这使得 WLSTSVM 对于含有交叉噪声样本数据集的分类效果不好.而且,WLSTSVM 的权重值是经过对样本数据集训练提取误差信息后计算得到的,这种方法不但会消耗大量的计算时间,而且对交叉分类面数据集不具备鲁棒性.另外,必须指出的是,标准的支持向量机考虑了结构风险最小化原则,而 WLSTSVM 和 LSTSVM 只具备经验风险最小化.

本文基于 LSTSVM 算法提出了广泛权重的最小二乘孪生支持向量机(Widely Weighted LSTSVM, WWLSTSVM).对于每一个 QPP 而言,与 WLSTSVM 相同的是 WWLSTSVM 在一类样本误差量上增加了权重,不同的是在另一类样本误差量上也增加了权重.这一改变有利于抑制交叉噪声样本对分类结果的影响.其次,利用 TBSVM 的思想,对 WWLSTSVM 的目标函数增加正规化项以实现结构风险最小化,而且所增加的正规化项可避免在求解过程中可能对病态矩阵的求逆处理.再者,为 WWLSTSVM 设计了一种有效的权重算法,即通过采用一种指数函数估计训练样本的密度信息来获得相应样本的权重.基于人工数据集的实验表明,WWLSTSVM 能有效地抑制交叉噪声样本对数据分类的影响,而且其权重算法对非交叉分类面数据集和交叉分类面数据集都具有很好的鲁棒性;基于 UCI 数据集的实验表明,WWLSTSVM 具有高的分类精度和计算效率.

2　最小二乘孪生支持向量机

LSTSVM 源于 TSVM,通过求解两个 QPP 来获得一对非平行的分类超平面,适合交叉分类面数据集的分类.LSTSVM 又不同于 TSVM,LSTSVM 求解的是两个线性方程而不是求解两个 QPP,因此其计算速度要比 TSVM 快.下面,对非线性的 LSTSVM 作一描述:

$$\min_{\boldsymbol{u}_1,\gamma_1}\frac{1}{2}\| K(\boldsymbol{A},\boldsymbol{C}')\boldsymbol{u}_1+\boldsymbol{e}_1\gamma_1\|^2+\frac{C_1}{2}\boldsymbol{\xi}_2'\boldsymbol{\xi}_2 \quad (1)$$

$$\text{s.t.}\ -(K(\boldsymbol{B},\boldsymbol{C}')\boldsymbol{u}_1+\boldsymbol{e}_2\gamma_1)=\boldsymbol{e}_2-\boldsymbol{\xi}_2$$

$$\min_{\boldsymbol{u}_2,\gamma_2}\frac{1}{2}\| K(\boldsymbol{B},\boldsymbol{C}')\boldsymbol{u}_2+\boldsymbol{e}_2\gamma_2\|^2+\frac{C_2}{2}\boldsymbol{\xi}_1'\boldsymbol{\xi}_1 \quad (2)$$

$$\text{s.t.}\ (K(\boldsymbol{A},\boldsymbol{C}')\boldsymbol{u}_2+\boldsymbol{e}_1\gamma_2)=\boldsymbol{e}_1-\boldsymbol{\xi}_1$$

其中,$\boldsymbol{A}$ 表示 +1 类样本矩阵,$\boldsymbol{B}$ 表示 −1 类样本矩阵,$\boldsymbol{C}=[\boldsymbol{A}'\ \boldsymbol{B}']'$,$\|\cdot\|$ 表示 L_2 范数,K 表示可选择的核函数,$\boldsymbol{\xi}_1$ 和 $\boldsymbol{\xi}_2$ 表示误差量,$\boldsymbol{e}_1$ 和 $\boldsymbol{e}_2$ 表示全 1 向量,C_1 和 C_2 为惩罚因子.

根据式(1)和式(2),两个非平行的分类超平面 $K(\boldsymbol{x}',\boldsymbol{C}')\boldsymbol{u}_1+\gamma_1=0$ 和 $K(\boldsymbol{x}',\boldsymbol{C}')\boldsymbol{u}_2+\gamma_2=0$ 对应的解如下:

$$\begin{bmatrix}\boldsymbol{u}_1\\ \gamma_1\end{bmatrix}=-\left(\boldsymbol{H}'\boldsymbol{H}+\frac{1}{C_1}\boldsymbol{G}'\boldsymbol{G}\right)^{-1}\boldsymbol{H}'\boldsymbol{e}_2 \quad (3)$$

$$\begin{bmatrix}\boldsymbol{u}_2\\ \gamma_2\end{bmatrix}=\left(\boldsymbol{G}'\boldsymbol{G}+\frac{1}{C_2}\boldsymbol{H}'\boldsymbol{H}\right)^{-1}\boldsymbol{G}'\boldsymbol{e}_1 \quad (4)$$

其中,$\boldsymbol{G}=[K(\boldsymbol{A},\boldsymbol{C}')\ \boldsymbol{e}_1]$,$\boldsymbol{H}=[K(\boldsymbol{B},\boldsymbol{C}')\ \boldsymbol{e}_2]$.式(3)有一逆矩阵项$(\boldsymbol{H}'\boldsymbol{H}+\boldsymbol{G}'\boldsymbol{G}/C_1)^{-1}$,而 $\boldsymbol{H}'\boldsymbol{H}+\boldsymbol{G}'\boldsymbol{G}/C_1$ 可能存在不可逆的情况.为了避免这种情况发生,一般将逆矩阵项改为$(\boldsymbol{H}'\boldsymbol{H}+\boldsymbol{G}'\boldsymbol{G}/C_1+\varepsilon\boldsymbol{I})^{-1}$[10].这里,$\boldsymbol{I}$ 是单位矩阵,ε 是一个很小的正数.

对于线性的 LSTSVM,其分类超平面为 $\boldsymbol{x}'\boldsymbol{w}_1+\gamma_1=0$ 和 $\boldsymbol{x}'\boldsymbol{w}_2+\gamma_2=0$.它可以通过使用等式 $K(\boldsymbol{x}',\boldsymbol{C}')=\boldsymbol{x}'\boldsymbol{C}$、$\boldsymbol{w}_1=\boldsymbol{C}'\boldsymbol{u}_1$ 和 $\boldsymbol{w}_2=\boldsymbol{C}'\boldsymbol{u}_2$ 变换获得.

3　广泛权重的最小二乘孪生支持向量机

3.1　线性 WWLSTSVM

假设样本矩阵 $\boldsymbol{A}\in\mathbf{R}^{m_1\times d}$,$\boldsymbol{B}\in\mathbf{R}^{m_2\times d}$.其中,$\boldsymbol{A}=[\boldsymbol{A}_1\ \boldsymbol{A}_2\cdots\ \boldsymbol{A}_{m_1}]'$,$\boldsymbol{B}=[\boldsymbol{B}_1\ \boldsymbol{B}_2\cdots\ \boldsymbol{B}_{m_2}]'$,$\boldsymbol{A}_i\in\mathbf{R}^{d\times 1}$表示 +1 类的数据样本,$\boldsymbol{B}_t\in\mathbf{R}^{d\times 1}$表示 −1 类的数据样本,样本的总数为 $m=m_1+m_2$.则线性的 WWLSTSVM 二次规划问题可表示如下:

$$\begin{aligned}&\min_{w_1,b_1}\frac{C_3}{2}(\|\boldsymbol{w}_1\|^2+b_1^2)+\frac{1}{2}\sum_{i=1}^{m_1}v_{1i}\xi_{1i}^2+\frac{C_1}{2}\sum_{t=1}^{m_2}v_{2t}\xi_{2t}^2\\ &\text{s.t.}\ \boldsymbol{A}_i'\boldsymbol{w}_1+b_1=\xi_{1i},i=1,2,\cdots,m_1\\ &\qquad\boldsymbol{B}_t'\boldsymbol{w}_1+b_1+1=\xi_{2t},t=1,2,\cdots,m_2\end{aligned} \quad (5)$$

$$\begin{aligned}&\min_{w_2,b_2}\frac{C_4}{2}(\|\boldsymbol{w}_2\|^2+b_2^2)+\frac{C_2}{2}\sum_{i=1}^{m_1}v_{1i}\eta_{1i}^2+\frac{1}{2}\sum_{t=1}^{m_2}v_{2t}\eta_{2t}^2\\ &\text{s.t.}\ \boldsymbol{B}_t'\boldsymbol{w}_2+b_2=\eta_{2t},t=1,2,\cdots,m_2\\ &\qquad-(\boldsymbol{A}_i'\boldsymbol{w}_2+b_2)+1=\eta_{1i},i=1,2,\cdots,m_1\end{aligned} \quad (6)$$

其中,C_1、C_2、C_3、C_4 为惩罚因子,v_{1i}和 v_{2t}为权重参数.

显然,式(5)有一个额外的正规化项 $C_3(\|\boldsymbol{w}_1\|^2+b_1^2)/2$.由式(5)可以看出,要实现其结构风险最小化的方法是使两个支持超平面 $\boldsymbol{x}'\boldsymbol{w}_1+b_1=0$ 和 $\boldsymbol{x}'\boldsymbol{w}_1+b_1=-1$ 的间隔最大化[6].间隔最大化等价于正规化项 $C_3(\|\boldsymbol{w}_1\|^2+b_1^2)/2$ 最小化.同时,式(5)分别给误差量 ξ_{1i}和 ξ_{2t}增加了权重 v_{1i}和 v_{2t},且 $v_{1i}\geq 0$、$v_{2t}\geq 0$.与 WLSTSVM 相比,WWLSTSVM 给正负两类所有数据样本的误差量增加了权重,这种广泛实施权重的方法可以有效地去除噪声样本特别是交叉噪声样本对数据分类的影响.这一点将在后面的实验章节进行验证.式(5)中

的惩罚因子 C_1、C_3 可以实现最小化误差量 ξ_{1i}、最小化误差量 ξ_{2t}和最大化间隔三者之间的平衡.

首先,将式(5)的等式约束条件代入目标函数,可得到:

$$\min_{\boldsymbol{w}_1,b_1}\frac{C_3}{2}(\|\boldsymbol{w}_1\|^2+b_1^2)+\frac{1}{2}\sum_{i=1}^{m_1}v_{1i}(\boldsymbol{A}_i'\boldsymbol{w}_1+b_1)^2+\frac{C_1}{2}\sum_{t=1}^{m_2}v_{2t}(\boldsymbol{B}_t'\boldsymbol{w}_1+b_1+1)^2 \tag{7}$$

其次,将式(7)对于变量 $\boldsymbol{w}_1$ 和 b_1 偏导值设为 0,可得到:

$$C_3\boldsymbol{w}_1+\sum_{i=1}^{m_1}v_{1i}\boldsymbol{A}_i(\boldsymbol{A}_i'\boldsymbol{w}_1+b_1)+C_1\sum_{t=1}^{m_2}v_{2t}\boldsymbol{B}_t(\boldsymbol{B}_t'\boldsymbol{w}_1+b_1+1)=\boldsymbol{0} \tag{8}$$

$$C_3b_1+\sum_{i=1}^{m_1}v_{1i}(\boldsymbol{A}_i'\boldsymbol{w}_1+b_1)+C_1\sum_{t=1}^{m_2}v_{2t}(\boldsymbol{B}_t'\boldsymbol{w}_1+b_1+1)=0 \tag{9}$$

然后,将式(8)和(9)合并为矩阵形式:

$$C_3\boldsymbol{I}\begin{bmatrix}\boldsymbol{w}_1\\b_1\end{bmatrix}+\begin{bmatrix}\boldsymbol{A}'\boldsymbol{v}_1\boldsymbol{A}&\boldsymbol{A}'\boldsymbol{v}_1\boldsymbol{e}_1\\\boldsymbol{e}_1'\boldsymbol{v}_1\boldsymbol{A}&\boldsymbol{e}_1'\boldsymbol{v}_1\boldsymbol{e}_1\end{bmatrix}\begin{bmatrix}\boldsymbol{w}_1\\b_1\end{bmatrix}+C_1\begin{bmatrix}\boldsymbol{B}'\boldsymbol{v}_2\boldsymbol{B}&\boldsymbol{B}'\boldsymbol{v}_2\boldsymbol{e}_2\\\boldsymbol{e}_2'\boldsymbol{v}_2\boldsymbol{B}&\boldsymbol{e}_2'\boldsymbol{v}_2\boldsymbol{e}_2\end{bmatrix}\begin{bmatrix}\boldsymbol{w}_1\\b_1\end{bmatrix}=-C_1\begin{bmatrix}\boldsymbol{B}'\boldsymbol{v}_2\boldsymbol{e}_2\\\boldsymbol{e}_2'\boldsymbol{v}_2\boldsymbol{e}_2\end{bmatrix} \tag{10}$$

最后,定义两个矩阵 $\boldsymbol{E}$ 和 $\boldsymbol{F}$,并求解 $\boldsymbol{w}_1$ 和 b_1 如下:

$$\boldsymbol{E}=[\boldsymbol{p}_1\boldsymbol{A}\quad\boldsymbol{p}_1\boldsymbol{e}_1],\boldsymbol{F}=[\boldsymbol{p}_2\boldsymbol{B}\quad\boldsymbol{p}_2\boldsymbol{e}_2] \tag{11}$$

$$C_3\boldsymbol{I}\begin{bmatrix}\boldsymbol{w}_1\\b_1\end{bmatrix}+C_1\begin{bmatrix}(\boldsymbol{p}_2\boldsymbol{B})'\\(\boldsymbol{p}_2\boldsymbol{e}_2)'\end{bmatrix}[\boldsymbol{p}_2\boldsymbol{A}\quad\boldsymbol{p}_2\boldsymbol{e}_2]\begin{bmatrix}\boldsymbol{w}_1\\b_1\end{bmatrix}+\begin{bmatrix}(\boldsymbol{p}_1\boldsymbol{A})'\\(\boldsymbol{p}_1\boldsymbol{e}_1)'\end{bmatrix}[\boldsymbol{p}_1\boldsymbol{A}\quad\boldsymbol{p}_1\boldsymbol{e}_1]\begin{bmatrix}\boldsymbol{w}_1\\b_1\end{bmatrix}=-C_1\begin{bmatrix}\boldsymbol{B}'\boldsymbol{v}_2\boldsymbol{e}_2\\\boldsymbol{e}_2'\boldsymbol{v}_2\boldsymbol{e}_2\end{bmatrix} \tag{12}$$

$$\begin{bmatrix}\boldsymbol{w}_1\\b_1\end{bmatrix}=-(\frac{1}{C_1}\boldsymbol{E}'\boldsymbol{E}+\boldsymbol{F}'\boldsymbol{F}+\frac{C_3}{C_1}\boldsymbol{I})^{-1}\boldsymbol{F}'\boldsymbol{p}_2\boldsymbol{e}_2 \tag{13}$$

上述公式中,$\boldsymbol{e}_1\in\mathbf{R}^{m_1\times1}$和 $\boldsymbol{e}_2\in\mathbf{R}^{m_2\times1}$是全 1 向量,$\boldsymbol{I}$ 是$(d+1)\times(d+1)$维单位矩阵,$\boldsymbol{v}_1\in\mathbf{R}^{m_1\times m_1}$ 和 $\boldsymbol{v}_2\in\mathbf{R}^{m_2\times m_2}$是分别以$(\boldsymbol{v}_1)_{ii}$和$(\boldsymbol{v}_2)_{tt}$为对角元素的对角矩阵,$\boldsymbol{v}_1=\boldsymbol{p}_1'\boldsymbol{p}_1$,$\boldsymbol{v}_2=\boldsymbol{p}_2'\boldsymbol{p}_2$,$\boldsymbol{p}_1$ 和 $\boldsymbol{p}_2$ 是对角阵.可以看出,式(13)增加了一项 $C_3\boldsymbol{I}/C_1$.该项可以在不人为增加 $\varepsilon\boldsymbol{I}$ 项的情况下,有效的避免对病态矩阵的求逆处理.

同理,可求得式(6)的解:

$$\begin{bmatrix}\boldsymbol{w}_2\\b_2\end{bmatrix}=(\boldsymbol{E}'\boldsymbol{E}+\frac{1}{C_2}\boldsymbol{F}'\boldsymbol{F}+\frac{C_4}{C_2}\boldsymbol{I})^{-1}\boldsymbol{E}'\boldsymbol{p}_1\boldsymbol{e}_1 \tag{14}$$

最终,式(13)和(14)确定了两个最优的非平行分类超平面:

$$\boldsymbol{x}'\boldsymbol{w}_1+b_1=0,\quad\boldsymbol{x}'\boldsymbol{w}_2+b_2=0 \tag{15}$$

一个新的未标签数据样本 $\boldsymbol{x}\in\mathbf{R}^d$ 属于 +1 类还是 -1 类,取决于样本 $\boldsymbol{x}$ 距离哪一个分类超平面的垂直距离更近.其公式如(16)所示.其中,$|\cdot|$表示求绝对值.

$$\boldsymbol{x}\in\begin{cases}+1\text{类},\dfrac{|\boldsymbol{x}'\boldsymbol{w}_1+b_1|}{\|\boldsymbol{w}_1\|}<\dfrac{|\boldsymbol{x}'\boldsymbol{w}_2+b_2|}{\|\boldsymbol{w}_2\|}\\-1\text{类},\dfrac{|\boldsymbol{x}'\boldsymbol{w}_1+b_1|}{\|\boldsymbol{w}_1\|}>\dfrac{|\boldsymbol{x}'\boldsymbol{w}_2+b_2|}{\|\boldsymbol{w}_2\|}\end{cases} \tag{16}$$

3.2 非线性 WWLSTSVM

同样的,可将非线性 LSTSVM 扩展为非线性 WWLSTSVM.非线性 WWLSTSVM 构造的两个非平行分类超平面可表示为:

$$K(\boldsymbol{x}',\boldsymbol{C}')\boldsymbol{u}_1+\gamma_1=0,K(\boldsymbol{x}',\boldsymbol{C}')\boldsymbol{u}_2+\gamma_2=0 \tag{17}$$

其中,$\boldsymbol{C}=[\boldsymbol{A}'\quad\boldsymbol{B}']'$,$K$ 为可选择的核函数.非线性 WWLSTSVM 二次规划问题构造为:

$$\begin{aligned}&\min_{\boldsymbol{u}_1,\gamma_1}\frac{C_3}{2}(\|\boldsymbol{u}_1\|^2+\gamma_1^2)+\frac{1}{2}\sum_{i=1}^{m_1}v_{1i}\xi_{1i}^2+\frac{C_1}{2}\sum_{t=1}^{m_2}v_{2t}\xi_{2t}^2\\&\text{s.t. }K(\boldsymbol{A}_i',\boldsymbol{C}')\boldsymbol{u}_1+\gamma_1=\xi_{1i},i=1,2,\cdots,m_1\\&\quad K(\boldsymbol{B}_t',\boldsymbol{C}')\boldsymbol{u}_1+\gamma_1+1=\xi_{2t},t=1,2,\cdots,m_2\end{aligned} \tag{18}$$

$$\begin{aligned}&\min_{\boldsymbol{u}_2,\gamma_2}\frac{C_4}{2}(\|\boldsymbol{u}_2\|^2+\gamma_2^2)+\frac{C_2}{2}\sum_{i=1}^{m_1}v_{1i}\eta_{1i}^2+\frac{1}{2}\sum_{t=1}^{m_2}v_{2t}\eta_{2t}^2\\&\text{s.t. }K(\boldsymbol{B}_t',\boldsymbol{C}')\boldsymbol{u}_2+\gamma_2=\eta_{2t},t=1,2,\cdots,m_2\\&\quad-(K(\boldsymbol{A}_i',\boldsymbol{C}')\boldsymbol{u}_2+\gamma_2)+1=\eta_{1i},i=1,2,\cdots,m_1\end{aligned} \tag{19}$$

经过一系列的推导,式(18)和(19)的解如下:

$$\begin{bmatrix}\boldsymbol{u}_1\\\gamma_1\end{bmatrix}=-(\frac{1}{C_1}\boldsymbol{G}'\boldsymbol{G}+\boldsymbol{H}'\boldsymbol{H}+\frac{C_3}{C_1}\boldsymbol{I})^{-1}\boldsymbol{H}'\boldsymbol{p}_2\boldsymbol{e}_2 \tag{20}$$

$$\begin{bmatrix}\boldsymbol{u}_2\\\gamma_2\end{bmatrix}=(\boldsymbol{G}'\boldsymbol{G}+\frac{1}{C_2}\boldsymbol{H}'\boldsymbol{H}+\frac{C_4}{C_2}\boldsymbol{I})^{-1}\boldsymbol{G}'\boldsymbol{p}_1\boldsymbol{e}_1 \tag{21}$$

其中:

$$\begin{aligned}\boldsymbol{G}&=[\boldsymbol{p}_1K(\boldsymbol{A},\boldsymbol{C}')\ \boldsymbol{p}_1\boldsymbol{e}_1]\\\boldsymbol{H}&=[\boldsymbol{p}_2K(\boldsymbol{B},\boldsymbol{C}')\ \boldsymbol{p}_2\boldsymbol{e}_2]\end{aligned} \tag{22}$$

一个新的未标签数据样本 $\boldsymbol{x}\in\mathbf{R}^d$ 属于 +1 类还是 -1 类,取决于样本 $\boldsymbol{x}$ 距离哪一个分类超平面的距离更近,其公式如下:

$$\boldsymbol{x}\in\begin{cases}+1\text{类},\dfrac{|K(\boldsymbol{x}',\boldsymbol{C}')\boldsymbol{u}_1+\gamma_1|}{\|\boldsymbol{u}_1\|}<\dfrac{|K(\boldsymbol{x}',\boldsymbol{C}')\boldsymbol{u}_2+\gamma_2|}{\|\boldsymbol{u}_2\|}\\-1\text{类},\dfrac{|K(\boldsymbol{x}',\boldsymbol{C}')\boldsymbol{u}_1+\gamma_1|}{\|\boldsymbol{u}_1\|}>\dfrac{|K(\boldsymbol{x}',\boldsymbol{C}')\boldsymbol{u}_2+\gamma_2|}{\|\boldsymbol{u}_2\|}\end{cases} \tag{23}$$

4 WWLSTSVM 算法

4.1 权重算法

在 WWLSTSVM 公式中,参数 $\boldsymbol{v}_1$ 和 $\boldsymbol{v}_2$ 决定了分类的可靠性和准确性.最优的权重值可以有效地去除噪声样本特别是交叉噪声样本对数据分类的影响.WLSTSVM 提出了一种根据误差信息计算权重的算

法[11,12],但是它耗时长,而且对交叉分类面的数据集而言鲁棒性不够好.为了解决这些问题,本文提出了一种新的权重算法,它对训练样本的密度进行估计,并将此估计信息作为样本的权重.

核密度估计(Kernel Density Estimation,KDE)是从统计学领域发展而来.样本的 KDE 信息能够反映其在数据集中的重要性[8],即样本的密度越大,该样本在数据集中越重要.对于样本 $\boldsymbol{x}$ 而言,它的密度估计值可以通过计算样本 $\boldsymbol{x}$ 与其邻域样本的相似性而得到.样本 $\boldsymbol{x}$ 的密度值越大,则样本 $\boldsymbol{x}$ 与它的邻域样本就越相似.一般,可以使用高斯核函数来计算两个数据样本之间的距离来估计它们的相似性.文献[13]提到的一种指数函数也可以用于计算两个样本之间的距离,其表达式如下:

$$\mu(\boldsymbol{x}_i)=(\exp(d_p(\boldsymbol{x}_i,\boldsymbol{x}_j)^{\lambda}))^{-1/\beta} \tag{24}$$

其中,参数 λ 为正的常数,参数 β 为距离阈值. $d_p(\boldsymbol{x}_i,\boldsymbol{x}_j)$是样本 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 间的闵科夫斯基(Minkowski)距离.该指数函数可以很好的估计样本间的相似性[13,14].在文献[14]中,该指数函数被很好的应用于计算两个像素点之间的相似性,且 $d_p(\boldsymbol{x}_i,\boldsymbol{x}_j)$被定义为欧氏距离,$\lambda$ 被设为 1,β 被定义为一个可调参数.

本文利用式(24)来估计样本的密度,并认为样本的密度越大,则样本在数据集的权重越大.假设训练样本矩阵 $\boldsymbol{A}=[\boldsymbol{A}_1 \quad \boldsymbol{A}_2 \quad \cdots \quad \boldsymbol{A}_{m_1}]'$,$\Omega_i$ 是以 $\boldsymbol{A}_i$ 为中心、r 为半径的邻域.根据式(24),可以计算数据样本 $\boldsymbol{A}_i$ 在邻域 Ω_i 的密度,并将此密度值作为样本 $\boldsymbol{A}_i$ 的权重值,其公式如下:

$$v_{1i}=\sum_{\boldsymbol{A}_j\in\Omega_i}(\exp(\|\boldsymbol{A}_j-\boldsymbol{A}_i\|))^{-1/r},\ i=1,2,\cdots,m_1 \tag{25}$$

其中,$\|\cdot\|$代表 L_2 范数(欧式距离),λ 和 β 分别定义为 1 和 r.数据样本 $\boldsymbol{A}_i$ 的密度与邻域 Ω_i 内的所有样本都相关.邻域 Ω_i 内的样本 $\boldsymbol{A}_j$ 与中心样本 $\boldsymbol{A}_i$ 越近,则样本 $\boldsymbol{A}_j$ 对中心样本 $\boldsymbol{A}_i$ 的密度贡献就越大.显然,样本 $\boldsymbol{A}_i$ 的密度值最小为 1.同时,半径 r 会影响样本的密度.半径 r 越小,每一个数据样本就越独立,相反,数据样本的密度值差异就越小.因此,半径 r 应根据实际问题来决定,一般可从某个区间搜索最优的 r 值.

同样,对训练样本矩阵 $\boldsymbol{B}=[\boldsymbol{B}_1 \quad \boldsymbol{B}_2 \quad \cdots \quad \boldsymbol{B}_{m_2}]'$,定义 Ω_t 是以 $\boldsymbol{B}_t$ 为中心、r 为半径的邻域.数据样本 $\boldsymbol{B}_t$ 的权重值计算如下:

$$v_{2t}=\sum_{\boldsymbol{B}_j\in\Omega_t}(1+\exp(\|\boldsymbol{B}_j-\boldsymbol{B}_t\|))^{-1/r},\ t=1,2,\cdots,m_2 \tag{26}$$

4.2　线性和非线性 WWLSTSVM 算法

根据本文第三节的内容,可将线性分类器 WWLSTSVM 的算法归纳为:

第一步:选择合适的参数 r、C_1、C_2、C_3 和 C_4.

第二步:用式(25)和(26)计算权重参数 v_{1i}和 v_{2t}.

第三步:计算 $\boldsymbol{p}_1$ 和 $\boldsymbol{p}_2$,并由式(11)定义 $\boldsymbol{E}$ 和 $\boldsymbol{F}$.

第四步:利用式(13)和(14)计算并确定两个非平行的分类超平面.

第五步:根据式(16)对新的未标签数据样本 $\boldsymbol{x}\in\mathbf{R}^d$ 进行分类.

类似的,非线性分类器 WWLSTSVM 的算法可归纳为:

第一步:选择合适的参数 r、C_1、C_2、C_3、C_4 以及核函数 K.

第二步:用式(25)和(26)计算权重参数 v_{1i}和 v_{2t}.

第三步:计算 $\boldsymbol{p}_1$ 和 $\boldsymbol{p}_2$,并由式(22)定义 $\boldsymbol{G}$ 和 $\boldsymbol{H}$.

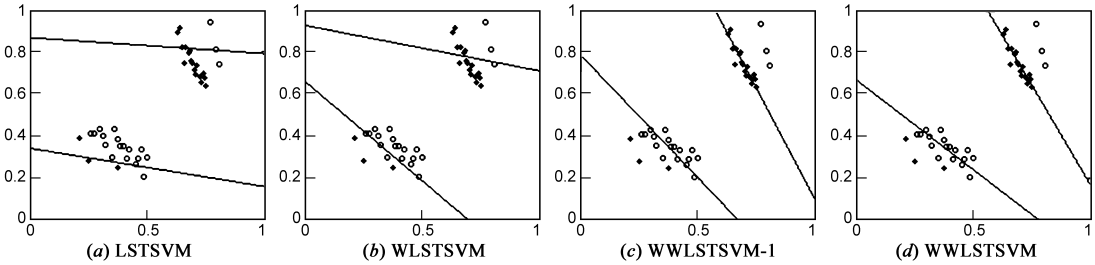
第四步:利用式(20)和(21)计算并确定两个非平行的分类超平面.

第五步:根据式(23)对新的未标签数据样本 $\boldsymbol{x}\in\mathbf{R}^d$ 进行分类.

5　实验及分析

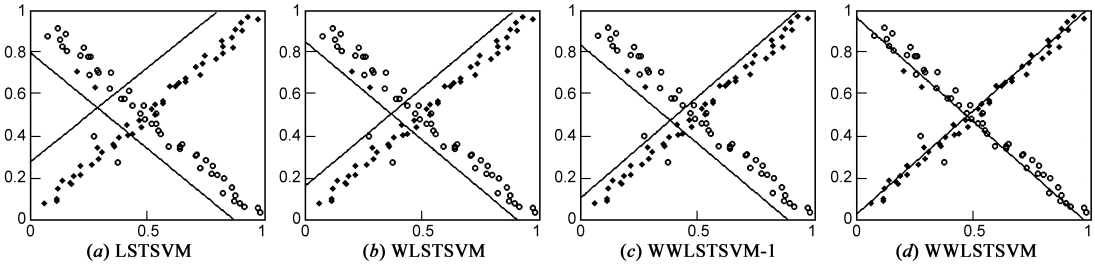
为了验证 WWLSTSVM 的性能,本文既在二维的人工数据集上进行了算法抑制交叉噪声样本干扰的对比实验,又在 UCI 标准数据集上进行了算法精度与效率的测试实验.

首先,本文针对含有交叉噪声样本的数据集进行分类实验,以此来验证算法抑制交叉噪声样本的有效性.假设有两个被交叉噪声样本污染的二维数据集,一个是非交叉分类面数据集,一个是交叉分类面数据集.分别用 LSTSVM、WLSTSVM、文献[11]提出的权重算法的 WWLSTSVM(WWLSTSVM-1)以及本文提出的新权重算法的 WWLSTSVM 进行分类实验,且实验的参数$\{C_i\mid i=1,2,3,4\}$和 r 均做了优化选择.其实验结果如图 1 和图 2 所示.从图 1 不难看出,LSTSVM 受交叉噪声样本的影响很大.对于 WLSTSVM,一类分类面的效果较好,而另一类的分类面发生了偏移,这是因为 WLSTSVM 只给一类样本增加了权重,没有考虑另一类样本的权重.对于 WWLSTSVM-1 和 WWLSTSVM,它们都具有很好的分类效果,这是因为它们对所有噪声样本的误差量赋予了小的权重值,而对所有的非噪声样本赋予了大的权重值.从图 2 可以看出,LSTSVM 和 WLSTSVM 分类器在交叉分类面数据集上受交叉噪声样本的影响很大.WWLSTSVM-1 在交叉分类面数据集上的分类结果也不理想,这是因为文献[11]提出的权重算法对交叉分类面数据集缺乏鲁棒性.本文提出的新权重算法的 WWLSTSVM 分类效果则较为理想.



(*a*) LSTSVM (*b*) WLSTSVM (*c*) WWLSTSVM-1 (*d*) WWLSTSVM

图1 四种分类器的分类结果:针对具有交叉噪声样本的非交叉分类面数据集



(*a*) LSTSVM (*b*) WLSTSVM (*c*) WWLSTSVM-1 (*d*) WWLSTSVM

图2 四种分类器的分类结果:针对具有交叉噪声样本的交叉分类面数据集

其次,为了验证 WWLSTSVM 的分类性能,本文从标准 UCI 机器学习库中选取了 8 个数据集,并分别对线性和非线性两种类型的 WWLSTSVM、WWLSTSVM-1、WLSTSVM、LSTSVM、PSVM、TSVM 进行分类精度与计算效率的测试实验.在实验的过程中,所有数据集的数据样本被归一化到[0,1]区间,并将数据集 90% 的数据样本划分为训练集,10% 的数据样本划分为测试集.非线性情况均采用高斯核函数,核参数从$[2^{-8},2^{4}]$区间选取.参数$\{C_i \mid i=1,2,3,4\}$在区间$[2^{-20},2^{4}]$中选取,参数 r 在区间[0.1,0.4]中选取.所有参数的选取方法均采用十字交叉验证法[5,10].最终所有分类器的分类测试结果如表 1 和表 2 所示.

表 1 线性分类器的分类精度与计算时间结果

UCI 数据集	WWLSTSVM		WWLSTSVM-1		WLSTSVM		LSTSVM		PSVM		TSVM	
	精度(%)	时间(s)	精度(%)	时间(s)	精度(%)	时间(s)	精度(%)	时间(s)	精度(%)	时间(s)	精度(%)	时间(s)
Heart-statlog	89.35	6.48e-4	88.37	8.02e-4	88.33	6.83e-4	87.29	4.48e-4	87.29	4.46e-4	87.66	0.4523
Bupaliver	71.29	6.53e-4	70.85	8.16e-4	70.54	6.99e-4	70.36	4.52e-4	70.15	4.51e-4	70.50	0.8411
Ionosphere	89.68	0.0016	87.58	0.0024	87.50	0.0019	84.38	0.0010	84.05	0.0009	83.23	0.9690
Votes	94.37	0.0020	93.26	0.0030	93.86	0.0024	93.45	0.0014	93.25	0.0012	94.06	1.8510
Australian	87.91	0.0026	87.12	0.0039	86.36	0.0031	86.36	0.0018	85.78	0.0015	86.09	6.9075
Pima-Indian	78.72	0.0039	77.29	0.0056	77.29	0.0045	75.94	0.0029	74.04	0.0024	74.97	7.2816
German	76.43	0.0084	76.07	0.0105	76.13	0.0093	76.07	0.0056	75.76	0.0048	74.52	8.2815
CMC	72.84	0.0179	70.64	0.0241	69.36	0.0198	68.84	0.0128	68.98	0.0116	68.84	8.6860

表 2 非线性分类器的分类精度与计算时间结果

UCI 数据集	WWLSTSVM		WWLSTSVM-1		WLSTSVM		LSTSVM		PSVM		TSVM	
	精度(%)	时间(s)	精度(%)	时间(s)	精度(%)	时间(s)	精度(%)	时间(s)	精度(%)	时间(s)	精度(%)	时间(s)
Heart-statlog	89.65	0.0256	88.98	0.0426	88.75	0.0364	87.92	0.0246	75.74	0.0239	87.18	1.1301
Bupaliver	76.21	0.0438	75.08	0.0718	74.91	0.0678	73.39	0.0429	73.39	0.0405	74.84	2.7005
Ionosphere	93.75	0.0636	92.85	0.0986	92.05	0.0838	90.06	0.0538	89.17	0.0502	90.06	5.5762
Votes	94.75	0.1227	93.56	0.1965	93.75	0.1625	93.50	0.1128	92.50	0.0988	92.50	7.5406
Australian	84.64	0.2297	82.74	0.4297	82.63	0.3445	81.80	0.2121	78.55	0.1918	80.15	33.432
Pima-Indian	80.65	0.2965	79.85	0.5105	79.37	0.4273	76.42	0.2786	77.80	0.2435	76.33	41.923
German	77.47	0.6108	77.04	0.9558	76.37	0.8258	75.71	0.5722	75.12	0.5262	74.56	46.036
CMC	75.36	1.5506	74.86	2.3206	74.42	2.0134	74.42	1.5009	73.91	1.4326	73.95	58.368

从表 1 和表 2 不难看出,本文提出的线性和非线性 WWLSTSVM 的分类精度要好于 WWLSTSVM-1、WLSTSVM、LSTSVM、PSVM、TSVM.以表 2 中的 Ionosphere 数据集为例,WWLSTSVM 的分类精度是 93.75%,而 WWLSTSVM-1、WLSTSVM、LSTSVM、PSVM、TSVM 的分类精度分别是 92.85%、92.05%、90.06%、89.17%、90.06%.这

表明本文提出的分类器 WWLSTSVM 提高了分类精度.从表 1 和表 2 中也可以看出各个分类器计算耗费的时间.由于本文提出的新权重算法降低了时间计算的复杂度,所以 WWLSTSVM 计算耗费的时间明显少于 WWLSTSVM-1 和 WLSTSVM.同时,TSVM 比 WWLSTSVM 耗费的计算时间要长的多,这是因为 TSVM 需要求解两个具有不等式约束的 QPP,而 WWLSTSVM 求解的是两个线性方程.但是,WWLSTSVM 比 LSTSVM 和 PSVM 两个分类器耗费的计算时间要长一些,这是因为 WWLSTSVM 花费了时间计算样本的权重来提高分类精度.

6　结论

本文在 LSTSVM 基础上提出了广泛权重的最小二乘孪生支持向量机(WWLSTSVM).WWLSTSVM 通过给所有的数据样本赋予权重值来抑制交叉噪声样本对分类结果的影响.同时,根据间隔最大化思想在目标函数上增加了正规化项,既实现了结构风险最小化,又避免了在求解过程中可能对病态矩阵求逆的处理.此外,本文还提出了一种新的权重算法,它使用一种指数函数计算训练样本的密度值,并将此密度作为样本的权重值.最终,在二维人工数据集上的实验表明,WWLSTSVM 可以有效地去除交叉噪声样本对分类结果的影响,并且对交叉分类面数据集具备好的鲁棒性.在 UCI 数据集上的实验表明,WWLSTSVM 算法具有较高的分类精度和计算效率.下一步的研究,可考虑将 WWLSTSVM 扩展到多类别分类中.

参考文献

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

[2] Bayro-Corrochano E J, Arana-Daniel N. Clifford support vector machines for classification, regression, and recurrence[J]. IEEE Trans on Neural Networks, 2010, 21(11): 1731 – 1746.

[3] Ertekin S, et al. Nonconvex online support vector machines[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2011, 33(2): 368 – 381.

[4] Zhang Y S, et al. Adaptive resource allocation with SVM-based multi-hop video packet delay bound violation modeling[J]. Chinese Journal of Electronics, 2011, 20(2): 261 – 267.

[5] Jayadeva, et al. Twin support vector machines for pattern classification[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2007, 29(5): 905 – 910.

[6] Shao Y H, et al. Improvements on twin support vector machines[J]. IEEE Trans on Neural Networks, 2011, 22(6): 962 – 968.

[7] Qi Z Q, Tian Y J, et al. Robust twin support vector machine for pattern classification[J]. Pattern classification, 2013, 46(1): 305 – 316.

[8] Peng X J, Xu Dong. Bi-density twin support vector machines for pattern recognition[J]. Neurocomputing, 2013, 99(1): 134 – 143.

[9] Ye Q L, Zhao C X, et al. Weighted twin support vector machines with local information and its application[J]. Neural Networks, 2012, 35(11): 31 – 39.

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

[11] Chen J, Ji G R. Weighted least squares twin support vector machines for pattern classification[A]. The 2nd International Conference on Computer and Automation Engineering[C]. Singapore: IEEE. 2010. 242 – 246.

[12] Suykens J A K, et al. Weighted least squares support vector machines: Robustness and sparse approximation[J]. Neurocomputing, 2002, 48(1 – 4): 85 – 105.

[13] Plataniotis K N, et al. Adaptive fuzzy systems for multichannel signal processing[J]. Proceedings of the IEEE, 1999, 87(9): 1601 – 1622.

[14] Jin L H, et al. Improved bilateral filter for suppressing mixed noise in color images[J]. Digital Signal Processing, 2012, 22(6): 903 – 912.

作者简介

储茂祥　男,1978 年生于安徽桐城.东北大学信息科学与工程学院博士研究生.辽宁科技大学电子与信息工程学院讲师.研究方向为模式识别、智能系统.

E-mail: chu52_2004@163.com

王安娜　女,1956 年生于辽宁鞍山.东北大学信息科学与工程学院博士生导师、教授.研究方向为模式识别、电力电子系统、复杂工业过程故障诊断.