基于 pinball 损失的结构模糊孪生支持向量机

李 凯,李 慧
(河北大学网络空间安全与计算机学院,河北保定 071000)

摘 要: 孪生支持向量机通过求解较小的二次规划问题,提高了分类器的性能,然而,该方法主要利用了类间可分的特性,并使用 hinge 损失函数构建相应的模型,它们并未充分考虑不同类中数据的结构信息以及不同样本对分类的影响,导致该方法对噪声具有较强的敏感性以及重取样的不稳定性. 为了进一步提高孪生支持向量机的性能,基于 pinball 损失函数,将数据集中不同类的结构信息以及不同样本的作用引入到孪生支持向量机中,获得了基于 pinball 损失的结构模糊孪生支持向量机模型,从理论上导出了基于 pinball 损失的结构模糊孪生支持向量机算法 pin-sftsvm,通过选取人工生成数据集与 UCI 标准数据集,对 pin-sftsvm 算法进行了实验,并与 tbsvm、s-tsvm 和 pin-tsvm 算法进行了性能比较,表明了提出算法的有效性.

关键词: 结构信息; pinball 损失; 模糊隶属度; 孪生支持向量机
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2019)10-2221-07
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2019.10.025

Structural Fuzzy Twin Support Vector Machine with Pinball Loss

LI Kai, LI Hui
(*School of Cyber Security and Computer, Hebei University, Baoding, Hebei* 071000, *China*)

Abstract: Twin support vector machine improves the performance of the classifier by solving the smaller quadratic programming problem. However, this method mainly utilizes the separability between classes and constructs the corresponding model using the hinge loss function. Not considering the structural information of the intra-class data and the influence of different samples on the classification, the method has strong sensitivity to noise and instability of resampling. In order to further improve the performance of the twin support vector machine, the structural information of different classes in the data and the effects of different samples are introduced into the twin support vector machine based on the pinball loss function, and the structure fuzzy support vector machine model based on pinball loss is obtained. The structural fuzzy twin support vector machine algorithm pin-sftsvm based on the pinball loss is derived theoretically. The presented algorithm pin-sftsvm is tested by selecting the artificially generated data set and the UCI standard data set, and compared with the tbsvm, s-tsvm and pin-tsvm algorithms. Experimental results show the effectiveness of the proposed algorithm.

Key words: structural information; pinball loss; membership degree of fuzzy; twin support vector machine

1 引言

支持向量机 SVM (Support Vector Machine)是由 Vapnik 等人[1]提出的一种机器学习方法,在模式识别、图像分割等领域获得了较广泛的应用. 之后,人们对支持向量机进行了较深入的研究,并提出了不同的支持向量机算法[2~4],然而,这些方法却存在训练时间长、抗噪性能低、计算复杂性高等缺陷. 为此,人们基于模糊理论或粗糙集理论,提出了模糊支持向量机与粗糙支持向量机[5~9],较好地解决了噪声或异常点对支持向量机的影响. 为了降低支持向量机的计算复杂性,Fung 等人[10]提出了近邻支持向量机 PSVM(Proximal SVM),利用样本到超平面距离的大小决定该样本所属的类别. Mangasarian 等人[11]基于 PSVM 且放松了对两个超平面平行的约束,提出了广义特征值的支持向量机 GEPSVM (Generalized eigenvalues PSVM). 受 GEPSVM 方法的启发,Jayadeva 等人[12]提出了孪生支持向量机 TWSVM (Twin SVM),该方法通过求解两个较小规模的优化问题获得超平面,并且每个优化问题只使用近似一半的样本数据,使得该方法的训练时间被缩短到原 SVM 的 1/4. 研究表明,TWSVM 只利用了经验风险最小化,使得该方法的泛化性能较低且对噪声较为敏感,针对此种情况,人们对 TWSVM 进行了改进,提出了不同的孪

收稿日期:2019-05-16;修回日期:2019-06-15;责任编辑:梅志强
基金项目:河北省自然科学基金(No. F2018201060)

生支持向量机算法[13~18]. 可以看到,对于这些孪生支持向量机方法,为了构造所需要的分类器,它们主要利用了类间样本的可分性,并未考虑类内样本的相似性或类内的结构信息,使得算法的泛化性能并未得到较大的改善. 为此,人们将数据中的局部信息或 k 近邻方法引入到 TWSVM 中,提出了一些改进的孪生支持向量机[19~21]. 最近,人们通过对损失函数的研究,并结合 pinball 函数,提出了 pin-tsvm 算法[22],遗憾的是,此方法并未考虑样本集中的结构信息,以及不同样本对支持向量机的作用,使得该方法对噪声或异常点仍具有较大的敏感性和较低的泛化性能.

为了进一步提高孪生支持向量机的性能,在 pin-tsvm 的基础上,本文进一步研究了基于 pinball 损失的结构模糊孪生支持向量机,将样本集中的结构信息与不同样本的作用引入到 pin-tsvm 中,构造了基于 pinball 损失的结构模糊孪生支持向量机模型,提出了基于 pinball 损失的结构模糊孪生支持向量机算法 pin-sftsvm.

2 基于 pinball 损失的结构模糊孪生支持向量机

给定数据集 $\boldsymbol{X}=\{(x_i,y_i,s_i)\mid i=1,2,\cdots,l\}$,其中 $x_i\in\mathbf{R}^n, y_i\in\{+1,-1\}, i=1,2,\cdots,l, s_i$ 是样本 x_i 隶属于类 y_i 的程度,φ 是输入空间 $\mathbf{R}^n$ 到特征空间 $\boldsymbol{Z}$ 的映射. 为了充分利用数据集中的先验结构信息,以及不同样本对分类器贡献的大小,本文在孪生支持向量机 pin-tsvm 的基础上,对不同样本赋予相应的权重,并将结构信息引入到该模型中,获得了一种改进的 pin-tsvm 算法,将其记为 pin-sftsvm. 在下面的内容中,分两种情况介绍 pin-sftsvm 算法.

2.1 线性情况

线性 pin-sftsvm 就是在 $\mathbf{R}^n$ 中寻找正超平面 $f_+(x)=\boldsymbol{w}_+^{\mathrm{T}}x+b_+=0$ 和负超平面 $f_-(x)=\boldsymbol{w}_-^{\mathrm{T}}x+b_-=0$,使得正类样本与负类样本分别满足 $\boldsymbol{w}_+^{\mathrm{T}}x_i+b_+\geq 0, i=1,2,\cdots,l_1, \boldsymbol{w}_-^{\mathrm{T}}x_j+b_-\leq 0, j=1,2,\cdots,l_2$,也就是 $1\cdot f_+(x_i)\geq 0, i=1,2,\cdots,l_1$ 与 $(-1)\cdot f_-(x_j)\geq 0, j=1,2,\cdots,l_2$,其中 l_1 与 l_2 分别是正类与负类样本的个数.

在 pin-sftsvm 中,使用的 pinball 损失函数[22]为

$$L_\tau(x,y,f(x))=\begin{cases}0-yf(x), 0-yf(x)\geq 0\\ -\tau(0-yf(x)), 0-yf(x)<0\end{cases}\tag{1}$$

结合此 pinball 损失函数,并将结构信息和样本的权重引入到 TPMSVM[18] 中,得到如下优化问题:

$$\min_{\boldsymbol{w}_+,b_+}-\frac{v_1}{l_2}\sum_{j=1}^{l_2}\frac{|\boldsymbol{w}_+^{\mathrm{T}}x_j^-+b_+|}{\|\boldsymbol{w}_+\|^2}+\frac{C_1}{l_1}\sum_{i=1}^{l_1}s_i^+L_{\tau_1}(x_i^+,y_i,f_+(x_i^+))+\frac{1}{2}C_2\boldsymbol{w}_+^{\mathrm{T}}\boldsymbol{\Sigma}_+\boldsymbol{w}_+\tag{2}$$

$$\min_{\boldsymbol{w}_-,b_-}-\frac{v_2}{l_1}\sum_{i=1}^{l_1}\frac{|\boldsymbol{w}_-^{\mathrm{T}}x_i^+b_-|}{\|\boldsymbol{w}_-\|^2}+\frac{C_3}{l_2}\sum_{j=1}^{l_2}s_j^-L_{\tau_2}(x_j^-,y_j,f_-(x_j^-))+\frac{1}{2}C_4\boldsymbol{w}_-^{\mathrm{T}}\boldsymbol{\Sigma}_-\boldsymbol{w}_-\tag{3}$$

将式(1)分别代入式(2)与(3)中,并通过进一步简化可得优化问题式(4)与(5),

$$\begin{aligned}&\min_{\boldsymbol{w}_+,b_+,\boldsymbol{\xi}}\frac{1}{2}\|\boldsymbol{w}_+\|^2+\frac{v_1}{l_2}\sum_{j=1}^{l_2}(\boldsymbol{w}_+^{\mathrm{T}}x_j^-+b_+)+\frac{C_1}{l_1}\sum_{i=1}^{l_1}s_i^+\xi_i+\frac{1}{2}C_2\boldsymbol{w}_+^{\mathrm{T}}\boldsymbol{\Sigma}_+\boldsymbol{w}_+\\ &\text{s.t. }\boldsymbol{w}_+^{\mathrm{T}}x_i^++b_+\geq 0-\xi_i\\ &\boldsymbol{w}_+^{\mathrm{T}}x_i^++b_+\leq 0+\frac{1}{\tau_1}\xi_i, i=1,2,\cdots,l_1\end{aligned}\tag{4}$$

$$\begin{aligned}&\min_{\boldsymbol{w}_-,b_-,\eta}\frac{1}{2}\|\boldsymbol{w}_-\|^2-\frac{v_2}{l_1}\sum_{i=1}^{l_1}(\boldsymbol{w}_-^{\mathrm{T}}x_i^++b_-)+\frac{C_3}{l_2}\sum_{j=1}^{l_2}s_j^-\eta_j+\frac{1}{2}C_4\boldsymbol{w}_-^{\mathrm{T}}\boldsymbol{\Sigma}_-\boldsymbol{w}_-\\ &\text{s.t. }-(\boldsymbol{w}_-^{\mathrm{T}}x_j^-+b_-)\geq 0-\eta_j\\ &-(\boldsymbol{w}_-^{\mathrm{T}}x_j^-+b_-)\leq 0+\frac{1}{\tau_2}\eta_j,\quad j=1,2,\cdots,l_2\end{aligned}\tag{5}$$

其中 ξ_i 与 η_j 是松弛变量,$\boldsymbol{\xi}$ 和 $\boldsymbol{\eta}$ 分别是由松弛变量构成的向量,τ_1 与 τ_2 是 pinball 损失函数的参数,$\boldsymbol{\Sigma}_+=\boldsymbol{\Sigma}_1^+\cdots+\boldsymbol{\Sigma}_i^+\cdots+\boldsymbol{\Sigma}_{C_{\mathrm{P}}}^+$ 和 $\boldsymbol{\Sigma}_-=\boldsymbol{\Sigma}_1^-\cdots+\boldsymbol{\Sigma}_j^-\cdots+\boldsymbol{\Sigma}_{C_{\mathrm{N}}}^-$ 是结构信息,$\boldsymbol{\Sigma}_i^+$ 和 $\boldsymbol{\Sigma}_j^-$ 分别是第 i 个簇和第 j 个簇的协方差矩阵,s_i^+ 和 s_j^- 是样本的模糊隶属度,C_{P} 与 C_{N} 分别为正类样本与负类样本的簇数.

为了求解式(4),构造拉格朗日函数

$$\begin{aligned}&L(\boldsymbol{w}_+,b_+,\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{\beta})\\ &=\frac{1}{2}\|\boldsymbol{w}_+\|^2+\frac{v_1}{l_2}\sum_{j=1}^{l_2}(\boldsymbol{w}_+^{\mathrm{T}}x_j^-+b_+)+\frac{C_1}{l_1}\sum_{i=1}^{l_1}s_i^+\xi_i+\\ &\frac{1}{2}C_2\boldsymbol{w}_+^{\mathrm{T}}\boldsymbol{\Sigma}_+\boldsymbol{w}_+-\sum_{i=1}^{l_1}\alpha_i(\boldsymbol{w}_+^{\mathrm{T}}x_i^++b_++\xi_i)+\\ &\sum_{i=1}^{l_1}\beta_i(\boldsymbol{w}_+^{\mathrm{T}}x_i^++b_+-\frac{1}{\tau_1}\xi_i)\end{aligned}$$

其中 $\boldsymbol{\alpha}\geq 0, \boldsymbol{\beta}\geq 0$ 是拉格朗日乘子构成的向量.

将 $L(\boldsymbol{w}_+,b_+,\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{\beta})$ 分别对 $\boldsymbol{w}_+,b_+,\xi_i$ 求偏导数并利用 KKT 条件,则得

$$\frac{\partial L}{\partial w_+}=\boldsymbol{w}_+(\boldsymbol{I}+C_2\boldsymbol{\Sigma}_+)+\frac{v_1}{l_2}\sum_{j=1}^{l_2}x_j^--\sum_{i=1}^{l_1}(\alpha_i-\beta_i)x_i^+=0\tag{6}$$

$$\frac{\partial L}{\partial b_+}=v_1-\sum_{i=1}^{l_1}(\alpha_i-\beta_i)=0\tag{7}$$

$$\frac{\partial L}{\partial \xi_i}=\frac{C_1}{l_1}s_i^+-\alpha_i-\frac{1}{\tau_1}\beta_i=0\tag{8}$$

$$\alpha_i(\boldsymbol{w}_+^{\mathrm{T}}x_i^++b_++\xi_i)=0\tag{9}$$

$$\beta_i(\boldsymbol{w}_+^{\mathrm{T}}x_i^+ + b_+ - \frac{1}{\tau_1}\xi_i) = 0 \tag{10}$$

从而得到

$$\boldsymbol{w}_+ = (\boldsymbol{I} + C_2\boldsymbol{\Sigma}_+)^{-1}(\sum_{i=1}^{l_1}(\alpha_i - \beta_i)x_i^+ - \frac{v_1}{l_2}\sum_{j=1}^{l_2}x_j^-)$$

$$v_1 = \sum_{i=1}^{l_1}(\alpha_i - \beta_i)$$

$$\frac{C_1}{l_1}s_i^+ = \alpha_i + \frac{1}{\tau_1}\beta_i$$

利用式(6)至(10),可以得到式(4)的对偶问题

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} -\frac{1}{2}(\boldsymbol{I} + C_2\boldsymbol{\Sigma}_+)^{-1}\sum_{i=1}^{l_1}\sum_{j=1}^{l_1}(\alpha_i - \beta_i)(x_i^+ \cdot x_j^+)(\alpha_j - \beta_j) + (\boldsymbol{I} + C_2\boldsymbol{\Sigma}_+)^{-1}\frac{v_1}{l_2}\sum_{i=1}^{l_1}\sum_{j=1}^{l_2}(\alpha_i - \beta_i)(x_i^+ \cdot x_j^-)$$

$$\text{s.t.}\ \sum_{i=1}^{l_1}(\alpha_i - \beta_i) = v_1$$

$$\alpha_i + \frac{1}{\tau_1}\beta_i = \frac{C_1}{l_1}s_i^+$$

$$0 \le \alpha_i \le \frac{C_1}{l_1}s_i^+, 0 \le \beta_i \le \frac{C_1}{l_1}s_i^+, i = 1,2,\cdots,l_1 \tag{11}$$

按照同样方法,可以获得式(5)的对偶问题式(12),

$$\max_{\boldsymbol{\gamma},\boldsymbol{\rho}} -\frac{1}{2}(\boldsymbol{I} + C_4\boldsymbol{\Sigma}_-)^{-1}\sum_{i=1}^{l_2}\sum_{j=1}^{l_2}(\gamma_i - \rho_i)(x_i^- \cdot x_j^-)(\gamma_j - \rho_j) + (\boldsymbol{I} + C_4\boldsymbol{\Sigma}_-)^{-1}\frac{v_2}{l_1}\sum_{i=1}^{l_1}\sum_{j=1}^{l_2}(\gamma_j - \rho_j)(x_i^+ \cdot x_j^-)$$

$$\text{s.t.}\ \sum_{j=1}^{l_2}(\gamma_j - \rho_j) = v_2$$

$$-\gamma_j - \frac{1}{\tau_2}\rho_j = \frac{C_3}{l_2}s_j^-$$

$$0 \le \gamma_j \le \frac{C_3}{l_2}s_j^-, 0 \le \rho_j \le \frac{C_3}{l_2}s_j^-, j = 1,2,\cdots,l_2 \tag{12}$$

通过求解优化问题式(11)与(12),分别得到乘子 $\boldsymbol{\alpha}$ 与 $\boldsymbol{\beta}$ 以及 $\boldsymbol{\gamma}$ 和 $\boldsymbol{\rho}$,从而得到

$$\boldsymbol{w}_+ = (\boldsymbol{I} + C_2\boldsymbol{\Sigma}_+)^{-1}(\sum_{i=1}^{l_1}(\alpha_i - \beta_i)x_i^+ - \frac{v_1}{l_2}\sum_{j=1}^{l_2}x_j^-),$$

$$\boldsymbol{w}_- = (\boldsymbol{I} + C_4\boldsymbol{\Sigma}_-)^{-1}(\frac{v_2}{l_1}\sum_{i=1}^{l_1}x_i^+ - \sum_{j=1}^{l_2}(\gamma_j - \rho_j)x_j^-),$$

$$b_+ = -\frac{1}{|N_+|}\sum_{i\in N_+}\boldsymbol{w}_+^{\mathrm{T}}x_i^+, b_- = -\frac{1}{|N_-|}\sum_{j\in N_-}\boldsymbol{w}_-^{\mathrm{T}}x_j^-,$$

其中 $N_+ = \{i \mid \alpha_i > 0, \beta_i > 0\}$, $N_- = \{j \mid \gamma_j > 0, \rho_j > 0\}$, $|N_+|$和$|N_-|$分别为集合 N_+ 与 N_- 的基数. 算法 pin-sftsvm 的决策函数为

$$g(x) = \mathrm{sign}\left(\frac{\boldsymbol{w}_+ \cdot x + b_+}{\|\boldsymbol{w}_+\|} + \frac{\boldsymbol{w}_- \cdot x + b_-}{\|\boldsymbol{w}_-\|}\right).$$

2.2　非线性情况

针对非线性可分问题,通过引入核函数,将线性 pin-sftsvm 扩展到非线性情况. 假设 φ 是一个从输入空间 $\mathbf{R}^n$ 到特征空间 $\boldsymbol{Z}$ 的映射,则非线性 pin-sftsvm 的优化问题分别为

$$\min_{\boldsymbol{w}_+,b_+,\boldsymbol{\xi}} \frac{1}{2}\|\boldsymbol{w}_+\|^2 + \frac{v_1}{l_2}\sum_{j=1}^{l_2}(\boldsymbol{w}_+^{\mathrm{T}}\varphi(x_j^-) + b_+) + \frac{C_1}{l_1}\sum_{i=1}^{l_1}s_i^+\xi_i + \frac{1}{2}C_2\boldsymbol{w}_+^{\mathrm{T}}\boldsymbol{\Sigma}_+\boldsymbol{w}_+$$

$$\text{s.t.}\ \boldsymbol{w}_+^{\mathrm{T}}\varphi(x_i^+) + b_+ \ge 0 - \xi_i$$

$$\boldsymbol{w}_+^{\mathrm{T}}\varphi(x_i^+) + b_+ \le 0 + \frac{1}{\tau_1}\xi_i, i = 1,2,\cdots,l_1 \tag{13}$$

$$\min_{\boldsymbol{w}_-,b_-,\boldsymbol{\eta}} \frac{1}{2}\|\boldsymbol{w}_-\|^2 - \frac{v_2}{l_1}\sum_{i=1}^{l_1}(\boldsymbol{w}_-^{\mathrm{T}}\varphi(x_i^+) + b_-) + \frac{C_3}{l_2}\sum_{j=1}^{l_2}s_j^-\eta_j + \frac{1}{2}C_4\boldsymbol{w}_-^{\mathrm{T}}\boldsymbol{\Sigma}_-\boldsymbol{w}_-$$

$$\text{s.t.}\ -(\boldsymbol{w}_-^{\mathrm{T}}\varphi(x_j^-) + b_-) \ge 0 - \eta_j$$

$$-(\boldsymbol{w}_-^{\mathrm{T}}\varphi(x_j^-) + b_-) \le 0 + \frac{1}{\tau_2}\eta_j, j = 1,2,\cdots,l_2 \tag{14}$$

按照 2.1 节给出的方法,分别获得式(13)与(14)的对偶问题式(15)与(16),即

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} -\frac{1}{2}(\boldsymbol{I} + C_2\boldsymbol{\Sigma}_+)^{-1}\sum_{i=1}^{l_1}\sum_{j=1}^{l_1}(\alpha_i - \beta_i)K(x_i^+,x_j^+)(\alpha_j - \beta_j) + (\boldsymbol{I} + C_2\boldsymbol{\Sigma}_+)^{-1}\frac{v_1}{l_2}\sum_{i=1}^{l_1}\sum_{j=1}^{l_2}K(x_i^+,x_j^-)(\alpha_i - \beta_i)$$

$$\text{s.t.}\ \sum_{i=1}^{l_1}(\alpha_i - \beta_i) = v_1$$

$$\alpha_i + \frac{1}{\tau_1}\beta_i = \frac{C_1}{l_1}s_i^+$$

$$0 \le \alpha_i \le \frac{C_1}{l_1}s_i^+, 0 \le \beta_i \le \frac{C_1}{l_1}s_i^+, i = 1,2,\cdots,l_1 \tag{15}$$

$$\max_{\boldsymbol{\gamma},\boldsymbol{\rho}} -\frac{1}{2}(\boldsymbol{I} + C_4\boldsymbol{\Sigma}_-)^{-1}\sum_{i=1}^{l_2}\sum_{j=1}^{l_2}(\gamma_i - \rho_i)K(x_i^-,x_j^-)(\gamma_j - \rho_j) + (\boldsymbol{I} + C_4\boldsymbol{\Sigma}_-)^{-1}\frac{v_2}{l_1}\sum_{i=1}^{l_1}\sum_{j=1}^{l_2}K(x_i^+ \cdot x_j^-)(\gamma_j - \rho_j)$$

$$\text{s.t.}\ \sum_{j=1}^{l_2}(\gamma_j - \rho_j) = v_2$$

$$-\gamma_j - \frac{1}{\tau_2}\rho_j = \frac{C_3}{l_2}s_j^-$$

$$0 \le \gamma_j \le \frac{C_3}{l_2}s_j^-, 0 \le \rho_j \le \frac{C_3}{l_2}s_j^-, j = 1,2,\cdots,l_2 \tag{16}$$

其中 $K(x_i,x_j)$ 是核函数. 通过求解优化问题式(15)与(16),进而得到非线性可分情况的决策函数.

3　实验研究

3.1　实验数据与方法

为了验证提出方法的有效性,我们选取了 UCI 数据库[23]的 7 个标准数据集以及人工生成数据集进行了实验,其中 UCI 标准数据集为两类数据,主要包括 cancer, ionosphere, liver-disorders, fertility, australian, planning relax

和 bankonote. 人工生成数据集为 data1,data2,data3 和 data4,它们使用了不同均值和方差的高斯分布生成的两类数据集,分别含有 200,300,300 和 250 个样本,其中 data1 与 data2 中的每类样本分别由 2 个簇和 3 个簇构成,而 data3 与 data4 中的每类样本由不同个数的簇构成. 为了验证提出算法的抗噪性能,对 UCI 数据集分别加入了 5% 与 10% 的噪声;对人工生成数据集 data1,data2,data3 和 data4 分别加入了 10% 的噪声,并将其分别记为 data1-n,data2-n,data3-n 与 data4-n. 实验中使用的核函数为 $K(x,y)=\exp(-\|x-y\|^2/(2\sigma^2))$,评价指标分别为正确率和标准偏差,实验中使用了五重交叉验证,获得的正确率为 10 次实验的平均值,损失函数的参数取值范围是 0 ~ 1,其它参数取值范围为 10^{-5} ~ 10^7.

3.2 实验结果与分析

3.2.1 人工数据集

在本小节,主要针对提出的算法 pin-sftsvm 与文献[22]中算法 pin-tsvm 在人工生成数据集 data1,data2,data3,data4 以及加入噪声的数据集 data1-n,data2-n,data3-n,data4-n 进行了实验. 为了表明不同结构信息的获取方法对 pin-sftsvm 的影响,实验中分别选取层次聚类,k 均值聚类,模糊 c 均值聚类与类内离散度获取结构信息;对于不同样本的作用,即样本的权重,我们使用了模糊 c 均值聚类方法获取,实验结果如图 1 所示. 由图 1(*a*)与(*b*)看到,不同的结构信息获取方法对算法的性能具有一定的影响,特别是当使用层次聚类,k 均值聚类与模糊 c 均值聚类获取结构信息后,pin-sftsvm 的性能优于使用类内离散度获取结构信息后的算法性能.

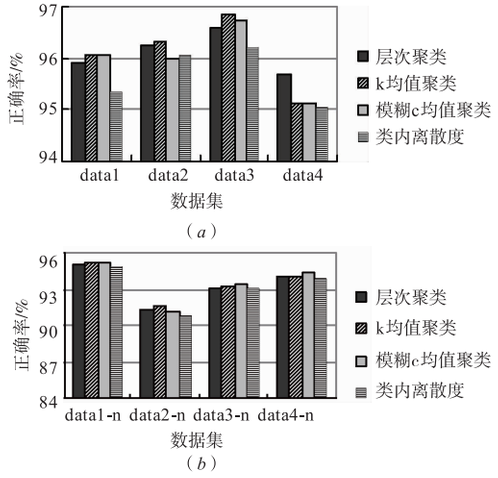

图1 人工生成数据集的实验结果

同时与算法 pin-tsvm[22] 进行了比较,实验结果如图 2 所示,其中 pin-sftsvm 算法的正确率是对图 1 中数据平均后获得的结果,可以看到,针对提出的算法 pin-sftsvm,由于加入了结构信息并考虑了不同样本的作用,因此,算法 pin-sftsvm 的性能优于 pin-tsvm.

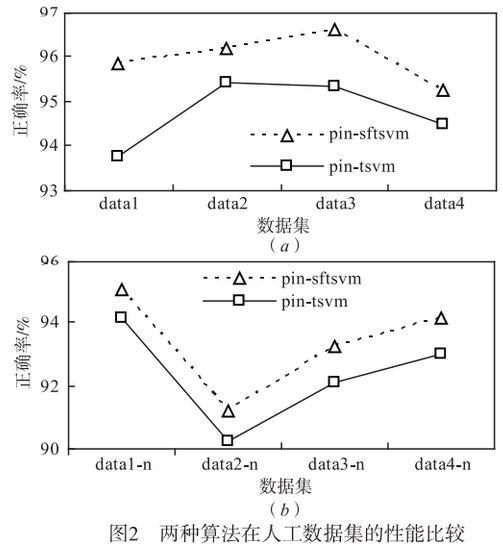

图2 两种算法在人工数据集的性能比较

另外,针对算法中损失函数的参数 τ_1 与 τ_2,我们实验研究了它们取不同值时 pin-sftsvm 的平均正确率及标准差,实验结果发现,使用聚类方法获取结构信息对算法具有较小的影响,而使用类内离散度方法获取的结构信息对算法具有较大的影响. 但总体看来,参数 τ_1 与 τ_2 的不同取值对算法的影响较小,表明了 pin-sftsvm 算法的性能是较稳定的.

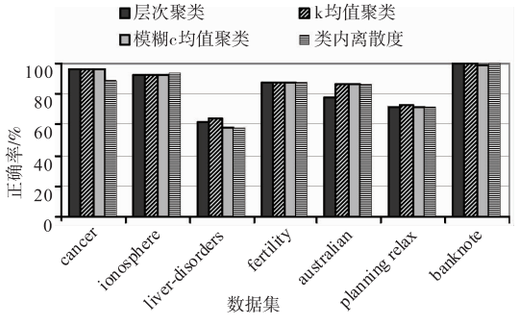

图3 pin-sftsvm在UCI数据集的实验结果

3.2.2 UCI 数据集

为了表明提出的算法 pin-sftsvm 在真实数据集上的性能,首先,针对 UCI 数据库中的 7 个数据集,使用不同的获取结构信息方法对算法进行了实验,实验结果如图 3 所示. 然后,针对 tbsvm[13],s-tsvm[20] 和 pin-tsvm[22] 算法对 UCI 中的 7 个数据集进行了实验,并与提出的算法 pin-sftsvm 进行了比较,实验结果如表 1 所示,其中 s-tsvm 与 pin-sftsvm 算法中结构信息的获取分别采用了层次聚类和 k 均值聚类. 可以看到,提出的算法 pin-sftsvm 在选择的 7 个数据集上的正确率优于算法 pin-tsvm,特别是,当使用 k 均值聚类获取结构信息时,pin-sftsvm 的正确率较为明显;与 tbsvm 和 s-tsvm 相比较,提出的算法也获得了较好的结果.

表 1　不同算法在 UCI 数据集的正确率(%)与标准差

数据集	tbsvm	s-tsvm(层次聚类)	s-tsvm(k 均值聚类)	pin-tsvm	pin-sftsvm(层次聚类)	pin-sftsvm(k 均值聚类)
cancer	96.1633 ±0.1714	96.0785 ±0.1838	96.0912 ±0.1344	95.943 ±0.1988	96.2511 ±0.3418	96.3088 ±0.1848
ionosphere	91.8233 ±2.142	90.9976 ±3.3477	91.6889 ±2.9905	79.8946 ±1.417	92.7066 ±0.4433	92.3924 ±0.599
liver-disorders	60.6957 ±2.1229	60.3768 ±2.7377	60 ±1.5823	58.6377 ±4.1845	61.3333 ±0.8906	64.5797 ±2.0115
fertility	88 ±0	88 ±0	88 ±0	87.8 ±0.7483	88.1 ±0.3	88 ±0
australian	82.1449 ±0.5606	82.7826 ±0.5643	82.8696 ±0.8495	78.4493 ±0.6982	78.1014 ±0.7054	86.9565 ±0.5345
planning relax	73.6579 ±0.0673	73.6433 ±0.0719	73.5497 ±0.3362	71.7544 ±0.4125	71.4415 ±0.074	73.1023 ±0.8792
banknote	99.8542 ±0.0003	99.8251 ±0.0484	99.8104 ±0.0743	92.9292 ±0.6161	99.8253 ±0.1088	99.8908 ±0.0586

为了表明算法 pin-sftsvm 的抗噪性能,对加入 5% 和 10% 的高斯噪声的标准数据集进行了实验,并与 tbsvm[13],s-tsvm[20] 和 pin-tsvm[22] 进行了比较,其中噪声服从均值为 0,方差为 1 的高斯分布,结构信息采用 k 均值聚类法获取,实验结果如表 2 所示. 可以看到,对于加入不同比例的噪声,算法 pin-sftsvm 的性能优于算法 tbsvm,s-tsvm 和 pin-tsvm,且噪声对算法的影响较小;由标准差可以知道,提出的算法 pin-sftsvm 其结果更稳定.

表 2　不同算法在加入噪声的数据集的正确率(%)与标准差

数据集	噪声比例	tbsvm	s-tsvm	pin-tsvm	pin-sftsvm
cancer	5%	93.6803 ±0.1812	93.7380 ±0.1797	92.9129 ±0.2314	94.1267 ±0.2201
	10%	92.0364 ±0.2292	92.1028 ±0.2605	90.4408 ±0.329	92.1177 ±0.1775
ionosphere	5%	89.9467 ±0.9162	84.6629 ±5.1331	74.6059 ±1.1944	90.4617 ±0.8065
	10%	86.4514 ±0.5147	86.4347 ±0.5585	80.4863 ±3.7064	87.5956 ±0.8262
liver-disorders	5%	61.4032 ±2.4328	59.9054 ±1.4089	60.4354 ±3.7692	62.4895 ±1.2384
	10%	60.6316 ±1.9129	59.1842 ±2.3015	58.6579 ±3.4078	61.7368 ±0.5289
fertility	5%	87.9048 ±0.6098	87.6190 ±0	87.8095 ±0.5714	88.0952 ±0.4762
	10%	88.1818 ±0	88.1818 ±0	87.9091 ±0.4166	88.1818 ±0
australian	5%	80.9261 ±0.8806	81.6171 ±1.1901	76.9180 ±0.5189	83.7561 ±0.4894
	10%	78.8546 ±0.7	78.9051 ±0.6081	74.5064 ±0.6307	82.0958 ±0.3908
planning relax	5%	73.3368 ±0.337	71.3816 ±0.0636	71.8263 ±0.4848	71.4158 ±0.0404
	10%	73.4000 ±0.3	73.3000 ±0.6	73.2500 ±0.4031	71.9000 ±0.5385
banknote	5%	96.6481 ±0.274	96.3777 ±0.2297	89.6954 ±0.5342	97.4739 ±0.0831
	10%	93.9299 ±0.1398	93.7444 ±0.0676	89.1778 ±0.3991	95.4868 ±0.1929

同时,我们也给出了显著程度为 0.05 的双边 t 检验的实验结果,如表 3 所示,其中 Win,Tie 与 Loss 分别表示 pin-sftsvm 算法的性能优于、打平以及劣于另一算法的数据集个数,算法 1 为本文提出的算法,算法 2 为 tbsvm,s-tsvm 和 pin-tsvm. 可以看到,在选择的 7 个数据集,不论数据中是否加入噪声,提出的算法 pin-sftsvm 与算法 tbsvm,s-tsvm 和 pin-tsvm 相比较都表现出较好的分类性能.

综上所述,当加入结构信息并考虑不同样本的作用后,提出的算法 pin-sftsvm 在选取的数据集上不仅获得了较好的分类正确率,较好的抗噪性与稳定性,而且其性能优于算法 tbsvm,s-tsvm 和 pin-tsvm.

表 3 不同算法在显著性水平为 0.05 的双边 t 检验结果

算法 1	算法 2		Win	Tie	Loss
pin-sftsvm	tbsvm	不加噪声	4	3	0
pin-sftsvm	s-tsvm		4	3	0
pin-sftsvm	pin-tsvm		6	1	0
pin-sftsvm	tbsvm	5% 噪声	6	0	1
pin-sftsvm	s-tsvm		6	1	0
pin-sftsvm	pin-tsvm		5	2	0
pin-sftsvm	tbsvm	10% 噪声	4	2	1
pin-sftsvm	s-tsvm		4	2	1
pin-sftsvm	pin-tsvm		5	1	1

4 结论

本文基于 pinball 损失函数对结构模糊孪生支持向量机进行了研究,提出了基于 pinball 损失的结构模糊孪生支持向量机算法 pin-sftsvm,并通过人工生成数据集与 UCI 标准数据集进行了实验,同时与 tbsvm,s-tsvm 和 pin-tsvm 算法进行了实验比较,表明了在孪生支持向量机中,利用 pinball 损失函数,并引入样本集中的结构信息与不同样本的作用,进一步提高了算法的分类正确率,抗噪性能以及分类的稳定性.

参考文献

[1] Vapnik V N. The Nature of Statistical Learning Theory [M]. New York:Springer-Verlag,2000.

[2] Scholkopf B,Smola A J,Williamson R C,et al. New support vector algorithms[J]. Neural Computation,2000,12(5):1207 - 1245.

[3] 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.

[4] Ding SF,Shi Z Z,Tao D C,et al. Recent advances in support vector machines[J]. Neurocomputing,2016,211(c):1 - 3.

[5] 方佳艳,刘峤,吴德,秦志光. 基于模糊 C-均值的相似性特征转换光滑支持向量机[J]. 电子学报,2018,46(11):2714 - 2724.
FANG Jiayan,LIU Qiao,WU De,QIN Zhiguang. Smooth support vector machine with similarity-based feature transformation technique and fuzzy c-means clustering[J]. Acta Electronica Sinica,2018,46(11):2714 - 2724. (in Chinese)

[6] Lin C F,Wang S D. Fuzzy support vector machines[J]. IEEE Transaction on Neural Networks,2002,13(2):464 - 471.

[7] Yang X W,Zhang G Q,Lu J. 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.

[8] Xu Y T. A rough margin-based linear υ support vector regression[J]. Statistics and Probability Letters,2012,82(3):528 - 534.

[9] 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.

[10] Fung G,Mangasarian O L. Proximal support vector machine classifiers[A]. Proceedings of the 7th International Conference on Knowledge and Data Discovery[C]. New York:ACM,2001. 77 - 86.

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

[12] 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.

[13] 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.

[14] Huang H J,Wei X X,Zhou Y Q. Twin support vector machines:A survey [J]. Neurocomputing,2018,300:34 - 43.

[15] Ding S F,Zhang N,Zhang X K,et al. Twin support vector machine:theory,algorithm and applications[J]. Neural Computing and Applications,2017,28(11):3119 - 3130.

[16] Ding S F,Yu J Z,Qi B J,et al. An overview on twin support vector machines[J]. Artificial Intelligence Revie,2014,42(2):245 - 252.

[17] Tomar D,Agarwal S. Twin support vector machine:a review from 2007 to 2014[J]. Egyptian Informatics Journal,2015,16(1):55 - 69.

[18] Peng X J. TPMSVM:A novel twin parametric-margin support vector machine for pattern recognition[J]. Pattern Recognition,2011,44(10):2678 - 2692.

[19] Xu Y T,Pan X L,Zhou Z J. Structural least square twin support vector machine for classification[J]. Applied Intelligence,2015,42(3):527 - 536.

[20] Qi Z Q,Tian Y J,Shi Y. Structural twin support vector machine for classification[J]. Knowledge-Based Systems,2013,43(1):74 - 81.

[21] 陈素根,吴小俊. 改进的投影孪生支持向量机[J]. 电子学报,2017,45(2):408 - 416.
CHEN Sugen,WU Xiaoju. Improved projection twin sup-

portvector machine[J]. Acta Electronica Sinica,2017,45(2):408 -416. (in Chinese)

[22] Xu Y T,Yang Z J,Pan X L. A novel twin support-vector machine with pinball loss[J]. IEEE Transactions on Neural Networks and Learning Systems,2017,28(2):359 -370.

[23] Blake C,Merz C J. UCI Repository for machine learning databases[DB/OL]. http://www.ics.uci.edu/mlearn/MLRepository.html,1998.

作者简介

李　凯　男,1963 年生于河北保定. 2005 年毕业于北京交通大学计算机与信息技术学院,并获得工学博士学位. 主要从事机器学习,模式识别,数据挖掘等方面研究.

E-mail:likai@hbu.edu.cn

李　慧　女,1993 年生于河北石家庄. 硕士研究生. 研究方向为机器学习与数据挖掘