基于数据分布特征的线性孪生支持向量机

宋瑞阳[1] 孟 华[1,2] 龙治国[2]
(西南交通大学数学学院 成都 611756)[1] (西南交通大学信息科学与技术学院 成都 611756)[2]

摘 要 孪生支持向量机(TWSVM)目前已在众多领域取得了成功的应用,但标准 TWSVM 模型在处理具有分布特征的数据分类问题时鲁棒性差,尤其当数据的不确定性程度较大时,不考虑样本点分布特征的标准分类模型已不能满足分类准确率的要求。为此,文中提出了基于数据分布特征的加权线性孪生支持向量机(TWSVM-U)模型,它在 TWSVM 的基础上考虑数据的分布特征对分类超平面位置的影响,根据数据在分类超平面法方向的分散程度定量构造距离权重。事实上,TWSVM-U 是 TWSVM 的推广,当训练样本数据不具有分布特征时,TWSVM-U 模型将退化为标准 TWSVM 模型。十折交叉验证的实验结果表明,TWSVM-U 模型在处理波动范围较大的不确定性数据分类问题时比 SVM 和 TWSVM 表现更优。

关键词 二分类,孪生支持向量机,不确定信息,加权距离
中图法分类号 TP301 **文献标识码** A

Linear Twin Support Vector Machine Based on Data Distribution Characteristics

SONG Rui-yang[1] MENG Hua[1,2] LONG Zhi-guo[2]
(School of Mathematics, Southwest Jiaotong University, Chengdu 611756, China)[1]
(School of Information Science and Technology, Southwest Jiaotong University, Chengdu 611756, China)[2]

Abstract Twin Support Vector Machine(TWSVM) have been successfully applied in many fields. However, the standard TWSVM model have poor robustness when dealing with data classification problems involving distribution characteristics, especially when uncertainty in data fluctuates wildly, the standard classification model, which doesn't consider the distribution characteristics, is no longer satisfactory for classification accuracy. Therefore, a weighted linear twin support vector machine model based on data distribution characteristics was proposed in this paper. The new model, denoted by TWSVM-U, further considers the influence of data distribution characteristics on the locations of classification hyperplanes, and constructs distance weights quantitatively according to data dispersity at the normal vector directions of classification hyperplanes. TWSVM-U is a generalization of TWSVM. In fact, when training samples do not have distribution characteristics, TWSVM-U model will degenerate to the standard TWSVM model. Experiments with 10-fold cross validation show that the TWSVM-U model performs better than the SVM and the TWSVM on classification problems with large data fluctuation range.

Keywords Binary classification, Twin support vector machine, Uncertain information, Weighted distance

1 引言

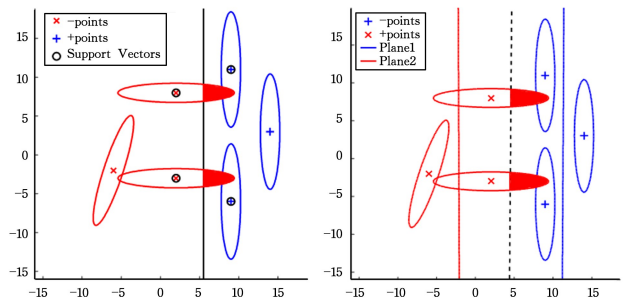
Cortes 等[1]于 1995 年提出了支持向量机(Support Vector Machine,SVM)模型,SVM 以最大化异类数据间的几何间隔为优化目标,通过求解一个凸二次规划问题(Quadratic Programming Problems,QPPs)得到分类超平面。Mangasarian 等[2]在 SVM 的基础上提出了一种基于广义特征值的近似支持向量机(Generalized Eigenvalue Proximal Support Vector Machine,GEPSVM)模型,它构造了两个超平面用以解决二分类问题,要求每一个超平面距离一类数据尽可能近,而距离另一类数据尽可能远。受到 GEPSVM 模型的启发,Jayadeva 等[3]提出了孪生支持向量机(Twin Support Vector Machine,TWSVM)模型,TWSVM 通过计算两个较小规模的二次规划问题得到两个分类超平面,在保证分类准确率的基础上将 SVM 的训练效率提升了 4 倍。目前 TWSVM 已经被广泛应用于模式识别[4-6]、回归[7]、聚类[8]等问题。

标准 SVM、TWSVM 模型均基于不带分布的训练样本来构建分类器,而实际的样本数据往往由于测量、变换、记录等操作存在一定的偏差。换言之,可以认为每个数据点具有一种分布,而最终用于训练标准模型的样本点只是其分布的大概率采样点或样本均值点,因此标准模型在处理具有分布特征的数据分类问题时存在缺陷。如图 1 所示,每个样本点都为一个二维正态分布的期望值,标准模型不考虑分布信息而只基于期望值得到的线性分类器会将阴影区域内的数据点错分。

不确定信息的研究广泛存在于生物[9]、通信[10]、军事[11]等领域,在监督学习中也有学者对不确定信息的表示与利用进行了研究[12-14]。Lanckriet 等[15]研究了在每一类数据的均值和协方差矩阵已知情况下的二分类问题,他们构造了一个极小极大问题,使测试数据错误分类的最大概率最小化;Peng 等[5]通过考虑每一类样本数据整体的分布特征,提出了基于

本文受 NSFC(61773324),教育部人文社科项目(18XJC72040001),中央高校基本科研业务费专项资金(2682016CX114,2682018CX25)资助。
宋瑞阳(1997—),男,主要研究方向为数据挖掘、机器学习;**孟 华**(1982—),男,博士,主要研究方向为知识表示与推理、机器学习,E-mail:menghua@home.swjtu.edu.cn(通信作者);**龙治国**(1989—),男,博士,主要研究方向为知识表示与推理、机器学习。

马氏距离的孪生支持向量机模型；Tzelepis 等[16]考虑多维正态分布噪声，从积分的角度改进了合页损失，提出了一种线性最大间隔分类器；Chen 等[6]通过结合模糊神经网络和 TWSVM，提出了一种新的模糊双支持向量机模型，减小了具有高度不确定性的样本对分类的影响，提高了泛化能力；Han 等[17]基于机会约束规划和隶属度，提出了一个模糊机会约束最小二乘孪生支持向量机模型，可以有效地度量数据的噪声。上述方法在一定程度上解决了噪声数据和数据集整体分布对分类的影响，但目前关于 TWSVM 如何处理每个样本点都具有分布特征的数据分类问题并没有明确的解决方案，如何基于样本点的分布特征改进 TWSVM 是亟待解决的问题。



(a)SVM (b)TWSVM

图 1 SVM 和 TWSVM 示例图

标准 TWSVM 模型仅基于训练数据的几何特征进行学习，为了提高 TWSVM 模型对带分布数据的分类准确率，本文从样本点在分类超平面法方向的分散程度入手考虑数据的分布特征，通过引入距离权重对 TWSVM 模型进行改进，提出了基于数据分布特征的线性孪生支持向量机(TwinSupport Vector Machine with Distribution Uncertainty，TWSVM-U)模型。TWSVM-U 是 TWSVM 的推广形式，当样本点不具有分布特征时，距离权重退化为 1，TWSVM-U 模型退化为 TWSVM 模型。本文最后通过在构造的说明性数据上的实验和 UCI 数据集上的一系列实验验证了 TWSVM-U 模型在处理具有分布特征的数据分类问题时的优势。

2 标准 TWSVM 模型

标准 TWSVM 模型考虑的是一个软间隔的二分类问题[3,18]，其通过求解两个二次规划问题得到两个分类超平面。模型的基本动机是让超平面距离同类点尽可能近，距离异类点的距离至少为 1。测试样本点的类别由其到两类超平面的距离较小的一类给出。记正负两类分类超平面分别为 $f_1(x)$，$f_2(x)$：

$$f_1(x)=x\cdot w^{(1)}+b^{(1)} \tag{1}$$

$$f_2(x)=x\cdot w^{(2)}+b^{(2)} \tag{2}$$

对于特征空间$\mathbb{R}^n$上给定的 m_1 个+1 类训练样本集：$A=[x_1^{(1)},x_2^{(1)},\cdots,x_{m_1}^{(1)}]^{\mathrm{T}}$，$m_2$ 个−1 类训练样本集：$B=[x_1^{(2)},x_2^{(2)},\cdots,x_{m_2}^{(2)}]^{\mathrm{T}}$。TWSVM 模型可以归结为求解以下一对二次规划问题：

$$\min_{w^{(1)},b^{(1)},p^{(2)}}\frac{1}{2}\|Aw^{(1)}+e_1b^{(1)}\|^2+c_1e_2^{\mathrm{T}}p^{(2)}$$
$$\text{s.t. } -(Bw^{(1)}+e_2b^{(1)})+p^{(2)}\geqslant e_2 \tag{3}$$
$$p^{(2)}\geqslant 0$$

$$\min_{w^{(2)},b^{(2)},p^{(1)}}\frac{1}{2}\|Bw^{(2)}+e_2b^{(2)}\|^2+c_2e_1^{\mathrm{T}}p^{(1)}$$
$$\text{s.t. } Aw^{(2)}+e_1b^{(2)}+p^{(1)}\geqslant e_1 \tag{4}$$
$$p^{(1)}\geqslant 0$$

式(3)和式(4)中，$w^{(1)}\in\mathbb{R}^n$，$w^{(2)}\in\mathbb{R}^n$，$b^{(1)}\in\mathbb{R}$，$b^{(2)}\in\mathbb{R}$，$p^{(1)}\in\mathbb{R}^{m_1}$，$p^{(2)}\in\mathbb{R}^{m_2}$，$c_1>0$，$c_2>0$，$e_1\in\mathbb{R}^{m_1}$，$e_2\in R^{m_2}$ 是分量全为 1 的列向量。

3 基于多维正态分布的 TWSVM-U 模型

图 1(b)中，虚线为分类中线，因此阴影部分全部被判为正类，然而从分布特性上来看，负类有更大的概率在此处取值。这些点被错分的直接原因是求解超平面时只考虑了每个分布的均值点或大概率采样点的几何间隔，未考虑数据的分布特征。为了提高正确率，一个简单的办法是对每个分布做高密度采样，并基于采样点求解 TWSVM 模型以适当考虑方差信息。这样做一方面人为增大了计算复杂度，另一方面由于 TWSVM 模型本身带有回归模型的特质，最终得到的分类超平面与仅基于样本均值求得的分类超平面并无太大差异，因此这一方法不是理想的选择。本文从协方差矩阵的角度考虑样本点的分布特征，提出了 TWSVM-U 模型。

3.1 基于协方差矩阵的权重构造方法

本文假设样本点满足多维正态分布，将样本点简记为(x_i,y_i,Σ_i)。首先不考虑样本点的分布特征，只关注其几何特征得到 TWSVM 的两个分类超平面 $f_1(x)$，$f_2(x)$。然后基于数据的分布特征修正两个分类超平面以提高分类准确率。对于正态分布而言，其分布信息由均值和协方差矩阵唯一确定，协方差矩阵的特征向量表示了数据的分散方向，特征值表示了分散程度。为了结合数据的分布特征给样本点距离赋予权重并通过求解加权模型来提高分类准确率，本文考虑了以下两方面内容。

1)注意到真正容易错分的点是到两个超平面距离相差较小的点，即两个分类超平面之间的点，而超平面外侧的点虽然也受分布影响，但不容易错分。本文通过对易错分点提高权重、对不易错分点降低权重来对距离进行修正。

2)图 1(b)中，当考虑分布信息后，为使得分类准确率提高，应当让负类超平面 $f_2(x)$右移。这主要是由于负类样本点在 $f_2(x)$的法方向的分散性大，更容易产生错分点，因此权重应当与数据在超平面法向量方向的分散程度有关。

综上所述，易错分点的权重应当增高，不易错分点的权重应当降低，同时在分类面法方向分散性大的点权重增高，分散性小的点权重降低。

以分类超平面 $x\cdot w+b=0$ 为例，对于任意的此类样本点(x_i,y_i,Σ_i)，其到超平面的欧氏平方距离 $d_e^2=\frac{(x_i\cdot w+b)^2}{w^{\mathrm{T}}w}$，马氏平方距离 $d_m^2=\frac{(x_i\cdot w+b)^2}{w^{\mathrm{T}}\Sigma_iw}$。其中，马氏平方距离的分母 $w^{\mathrm{T}}\Sigma_iw$ 刻画了该数据点在 w 方向上的分散程度。如果 $w^{\mathrm{T}}\Sigma_iw$ 越大，则该数据点的分布越容易产生远离超平面的点，即易错分。本文采用如下方式对距离赋权：

$$a_i=\begin{cases}\dfrac{w^{\mathrm{T}}w+w^{\mathrm{T}}\Sigma_iw}{w^{\mathrm{T}}w}, & x_i \text{ 容易错分}\\ \dfrac{w^{\mathrm{T}}w}{w^{\mathrm{T}}w+w^{\mathrm{T}}\Sigma_iw}, & x_i \text{ 不易错分}\end{cases} \tag{5}$$

可以看出，权重的选择满足上述要求，且当样本点不具有分布信息时，权重退化为 1。

对于超平面完全穿过训练样本点均值 x_i 的情况，超平面两侧均没有样本点，此时同类样本点到超平面的距离和为 0，因此考虑其权重也无法使超平面移动。为了避免这种现象的发生且进一步提高分布特征对模型的影响程度，本文为每个

点(x_i,y_i,Σ_i)增加额外两个点(x_i^+,y_i,Σ_i)，(x_i^-,y_i,Σ_i)作为模型的训练样本点，其中$x_i^+=x_i+w_e^{\mathrm{T}}\Sigma_i w_e\cdot w_e$，$x_i^-=x_i-w_e^{\mathrm{T}}\Sigma_i w_e\cdot w_e$，$w_e$是超平面单位化的法向量。当$x_i$不具有分布信息时，$x_i^+$和$x_i^-$为$x_i$，即此时仍发生退化。$x_i^+$和$x_i^-$的示例如图 2 所示。

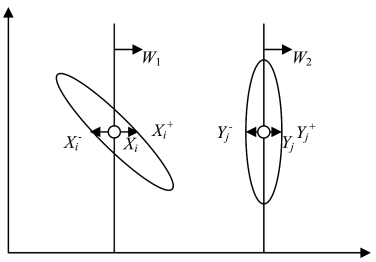


图 2　训练样本点示意图

3.2　TWSVM-U 模型

将引入新的训练样本后的训练数据集仍记为 A 和 B，通过上文的讨论，考虑权重后的 TWSVM-U 模型可以归结为求解以下一对二次规划问题：

$$\min_{w^{(1)},b^{(1)},p^{(2)}}\ \frac{1}{2}\ \|a^{(1)}(Aw^{(1)}+e_1b^{(1)})\|^2+c_1e_2^{\mathrm{T}}p^{(2)}$$
$$\text{s.t. } -(Bw^{(1)}+e_2b^{(1)})+p^{(2)}\geqslant e_2 \tag{6}$$
$$p^{(2)}\geqslant 0$$

$$\min_{w^{(2)},b^{(2)},p^{(1)}}\ \frac{1}{2}\|a^{(2)}(Bw^{(2)}+e_2b^{(2)})\|^2+c_2e_1^{\mathrm{T}}p^{(1)}$$
$$\text{s.t. } Aw^{(2)}+e_1b^{(2)}+p^{(1)}\geqslant e_1 \tag{7}$$
$$p^{(1)}\geqslant 0$$

其中，$a^{(1)}$和$a^{(2)}$分别是维度为 m_1 和 m_2 的对角矩阵，对角线元素为相对应训练样本到同类超平面的权重（见式(5)）。应用拉格朗日对偶性[19]，通过求解对偶问题得到原问题的最优解，式(6)的拉格朗日函数为：

$$L(w^{(1)},b^{(1)},p^{(2)},\alpha,\beta)=$$
$$\frac{1}{2}(a^{(1)}(Aw^{(1)}+e_1b^{(1)}))^{\mathrm{T}}(a^{(1)}(Aw^{(1)}+e_1b^{(1)}))+$$
$$c_1e_2^{\mathrm{T}}p^{(2)}-\alpha^{\mathrm{T}}(-(Bw^{(1)}+e_2b^{(1)})+p^{(2)}-e_2)-\beta^{\mathrm{T}}p^{(2)} \tag{8}$$

其中，$\alpha=(\alpha_1,\alpha_2,\cdots,\alpha_{m2})^{\mathrm{T}}$，$\beta=(\beta_1,\beta_2,\cdots,\beta_{m2})^{\mathrm{T}}$是拉格朗日乘子向量。由 K.K.T 条件可得：

$$(a^{(1)}A)^{\mathrm{T}}(a^{(1)}(Aw^{(1)}+e_1b^{(1)}))+B^{\mathrm{T}}\alpha=0 \tag{9}$$
$$(a^{(1)}e_1)^{\mathrm{T}}(a^{(1)}(Aw^{(1)}+e_1b^{(1)}))+e_2{}^{\mathrm{T}}\alpha=0 \tag{10}$$
$$c_1e_2-\alpha-\beta=0 \tag{11}$$
$$-(Bw^{(1)}+e_2b^{(1)})+p^{(2)}\geqslant e_2 \tag{12}$$
$$\alpha^{\mathrm{T}}(-(Bw^{(1)}+e_2b^{(1)})+p^{(2)}-e_2)=0 \tag{13}$$
$$\beta^{\mathrm{T}}p^{(2)}=0 \tag{14}$$
$$\alpha\geqslant 0,\beta\geqslant 0,p^{(2)}\geqslant 0 \tag{15}$$

将式(9)、式(10)改写成下述矩阵形式：

$$\begin{bmatrix}(a^{(1)}A)^{\mathrm{T}}\\(a^{(1)}e_1)^{\mathrm{T}}\end{bmatrix}[(a^{(1)}A\ a^{(1)}e_1)]\begin{bmatrix}w^{(1)}\\b^{(1)}\end{bmatrix}+\begin{bmatrix}B^{\mathrm{T}}\\e_2^{\mathrm{T}}\end{bmatrix}\alpha=0$$

记 $H=[a^{(1)}A\ a^{(1)}e_1]$，$H^{\mathrm{T}}=\begin{bmatrix}(a^{(1)}A)^{\mathrm{T}}\\(a^{(1)}e_1)^{\mathrm{T}}\end{bmatrix}$，$G=[B\ e_2]$，$G^{\mathrm{T}}=\begin{bmatrix}B^T\\e_2^{\mathrm{T}}\end{bmatrix}$，$U=\begin{bmatrix}w^{(1)}\\b^{(1)}\end{bmatrix}$。

则上述矩阵形式可以写为：

$$H^{\mathrm{T}}HU+G^{\mathrm{T}}\alpha=0 \tag{16}$$

当 $H^{\mathrm{T}}H$ 可逆时：

$$U=-(H^{\mathrm{T}}H)^{-1}G^{\mathrm{T}}\alpha \tag{17}$$

当 H^TH 不可逆时，加入正则化项，此时：

$$U=-(H^{\mathrm{T}}H+\varepsilon I)^{-1}G^{\mathrm{T}}\alpha \tag{18}$$

由式(11)，式(15)可得：

$$0\leqslant\alpha\leqslant c_1 \tag{19}$$

结合 K.K.T 条件可将式(6)转化为：

$$\max_{\alpha}\ \alpha^{\mathrm{T}}e_2-\frac{1}{2}\alpha^{\mathrm{T}}G(H^{\mathrm{T}}H+\varepsilon I)^{-1}G^{\mathrm{T}}\alpha$$
$$\text{s.t. } 0\leqslant\alpha\leqslant c_1 \tag{20}$$

同样可以得到式(7)的对偶问题为：

$$\max_{\gamma}\ \gamma^{\mathrm{T}}e_1-\frac{1}{2}\gamma^{\mathrm{T}}P(Q^{\mathrm{T}}Q+\varepsilon I)^{-1}P^{\mathrm{T}}\gamma$$
$$\text{s.t. } 0\leqslant\gamma\leqslant c_2 \tag{21}$$

其中，

$$P=[A\ e_1],P^{\mathrm{T}}=\begin{bmatrix}A^{\mathrm{T}}\\e_1^{\mathrm{T}}\end{bmatrix}$$

$$Q=[a^{(2)}B\ a^{(2)}e_2],Q^{\mathrm{T}}=\begin{bmatrix}(a^{(2)}B)^{\mathrm{T}}\\a^{(2)}e_2{}^{\mathrm{T}}\end{bmatrix}$$

$$V=\begin{bmatrix}w^{(2)}\\b^{(2)}\end{bmatrix}=-(Q^{\mathrm{T}}Q)^{-1}P^{\mathrm{T}}\gamma \tag{22}$$

通过求解式(20)、式(21)可得 α,γ，再由式(17)、式(22)可求出超平面参数 $w^{(1)},w^{(2)},b^{(1)},b^{(2)}$。对于测试数据 $x\in\mathbb{R}^n$，x 的类别由下式给出：

$$\mathrm{class}(x)=\arg\min_{k=1,2}(\mathrm{dist}_k(x))$$

其中，$\mathrm{dist}_k(x)=\dfrac{|x\cdot w^{(k)}+b^{(k)}|}{\|w^{(k)}\|}$。

3.3　TWSVM-U 求解流程

基于训练样本均值做一次 TWSVM 分类后，引入新的训练样本并计算权重来求解 TWSVM-U 模型，得到修正后的分类超平面。具体求解流程如算法 1 所示，其中 A^- 和 B^- 分别表示正类和负类的易错分点集。

算法 1　Solution flow for solving TWSVM-U

Inputs: A, B, c_1, c_2

1. Calculate (w_1, w_2, b_1, b_2) by TWSVM
2. Update A, B with new train data
3. **for** x_i in A **do**
   if x_i in A^-
   $$a_{ii}^{(1)}=\frac{w_1^{\mathrm{T}}w_1+w_1^{\mathrm{T}}\Sigma_i w_1}{w_1^{\mathrm{T}}w_1}$$
   else
   $$a_{ii}^{(1)}=\frac{w_1^{\mathrm{T}}w_1}{w_1^{\mathrm{T}}w_1+w_1^{\mathrm{T}}\Sigma_i w_1}$$
4. **end for**
5. **for** y_j in B **do**
   if y_j in B^-
   $$a_{jj}^{(2)}=\frac{w_2^{\mathrm{T}}w_2+w_2^{\mathrm{T}}\Sigma_j w_2}{w_2^{\mathrm{T}}w_2}$$
   else
   $$a_{jj}^{(2)}=\frac{w_2^{\mathrm{T}}w_2}{w_2^{\mathrm{T}}w_2+w_2^{\mathrm{T}}\Sigma_j w_2}$$
6. **end for**
7. Calculate (w_1', w_2', b_1', b_2') by TWSVM-U
8. Outputs: w_1', w_2', b_1', b_2'

4　实验验证

为验证 TWSVM-U 模型处理具有分布特征的数据分类问题的有效性，本文先后在说明性数据和引入正态分布后的 6 组 UCI[20] 数据集上分别对 SVM，TWSVM 和 TWSVM-U

进行了十折交叉验证的对比测试。

实验环境如下:处理器为 Intel i5-6200U,内存为 4 GB,编程环境为 MATLAB R2016a。

4.1 说明性数据实验

4.1.1 数据描述

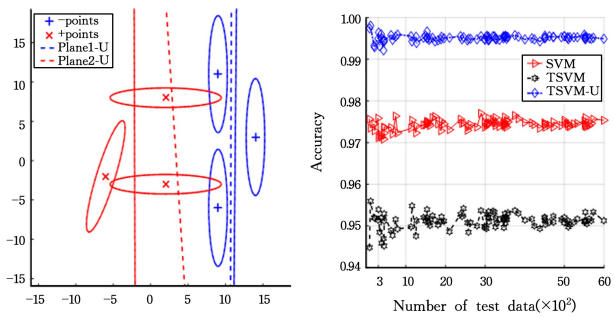
说明性数据示例如图 3(a)所示,通过协方差矩阵表示样本点的分布情况,样本数据的均值和协方差矩阵如表 1 所列。

表 1 说明性数据信息

数据点	均值	协方差矩阵	数据点	均值	协方差矩阵
正类 1	[9,11]	$\begin{bmatrix}0.17 & 0.00\\0.00 & 6.00\end{bmatrix}$	负类 1	[−6,−2]	$\begin{bmatrix}0.72 & 1.71\\1.71 & 5.44\end{bmatrix}$
正类 2	[14,3]	$\begin{bmatrix}0.17 & 0.00\\0.00 & 6.00\end{bmatrix}$	负类 2	[2,8]	$\begin{bmatrix}6.00 & 0.00\\0.00 & 0.17\end{bmatrix}$
正类 3	[9,−6]	$\begin{bmatrix}0.17 & 0.00\\0.00 & 6.00\end{bmatrix}$	负类 3	[−2,−3]	$\begin{bmatrix}6.00 & 10.00\\0.00 & 0.17\end{bmatrix}$

4.1.2 实验设计与结果

首先仅基于样本均值做一次标准 TWSVM 得到分类超平面(如图 3(a)中的实线所示),再根据 3.3 节提出的流程求解 TWSVM-U 模型,从而得到修正后的分类超平面(如图 3(a)中的虚线所示)。测试数据是由训练样本分布产生的随机点。SVM,TWSVM 和 TWSVM-U 3 个模型的预测准确率随测试样本数量变化的波动情况如图 3(b)所示,其中三角形线,六角星形线和菱形线依次表示 SVM,TWSVM 和 TWSVM-U 3 个模型。



(a) (b)

图 3 说明性数据实验结果

4.2 UCI 数据集实验

4.2.1 数据描述

实验选取的 6 组 UCI 数据集信息如表 2 所列,对于具有多个类别的数据集,实验时只选取前两类。在验证 SVM,TWSVM 和 TWSVM-U 模型时,将数据集中的数据点视为其对应的多维正态分布的均值点。为了体现两类数据的分散性差异,正类数据和负类数据的协方差矩阵分别取为其对应类别中的总体样本协方差矩阵的不同倍数,如表 2 中的正类倍数和负类倍数所示。

表 2 UCI 数据集信息

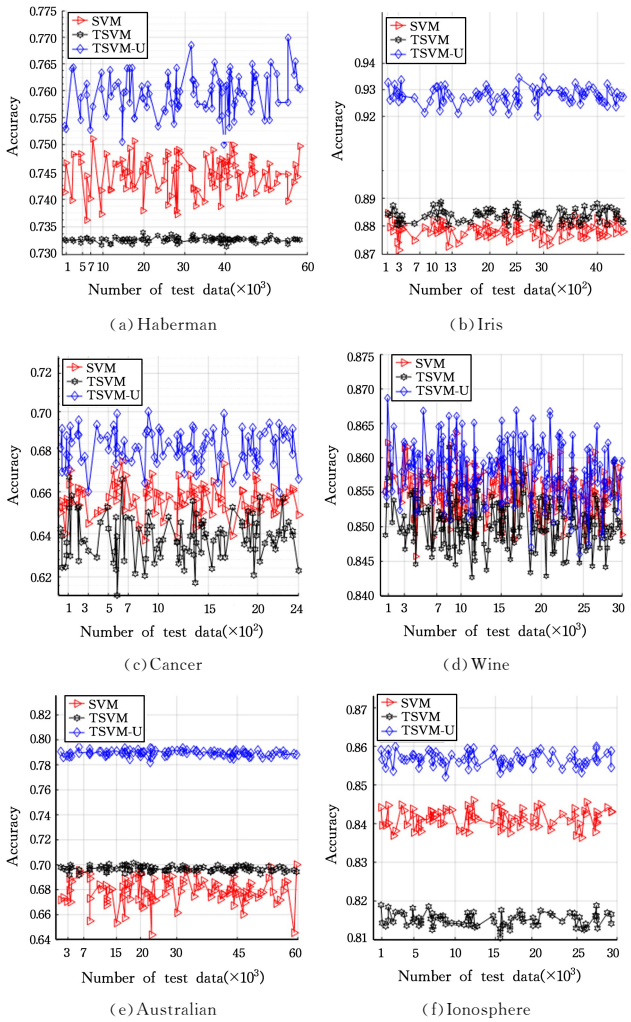
数据集	维度	样本数	正类数	负类数	正类倍数	负类倍数
Haberman	3	306	225	81	0.1	5
Iris	4	100	50	50	1	30
Cancer	9	116	52	64	0.1	10.5
Wine	13	130	59	71	0.05	10
Australian	14	690	383	307	0.05	20
Ionosphere	34	351	225	126	1	10

4.2.2 实验设计与结果

本文采用 10 折交叉验证评估模型,即将数据随机分为 10 等份,每次实验选取 9 份数据为训练样本,并计算剩余 1 份数据的分布,测试样本为此分布产生的随机点。最终的测试准确率为 10 次实验准确率的均值。

表 3 UCI 数据集 3000 测试样本的准确率

数据集	SVM	TWSVM	TWSVM-U
Haberman	73.1±0.02	73.4±0.01	**77.0±0.04**
Iris	88.4±0.03	87.5±0.03	**91.7±0.04**
Cancer	65.6±0.11	65.4±0.10	**66.0±0.14**
Wine	86.7±0.06	85.8±0.06	**87.3±0.04**
Australian	68.0±0.10	69.4±0.03	**79.2±0.05**
Ionosphere	84.0±0.03	81.2±0.04	**85.7±0.02**



(a) Haberman (b) Iris

(c) Cancer (d) Wine

(e) Australian (f) Ionosphere

图 4 UCI 数据集上的实验结果

SVM,TWSVM 和 TWSVM-U 3 个模型在 3000 随机测试样本上的准确率如表 3 所列。图 4 给出了 3 个模型在 6 个数据集上的准确率随测试样本数目的波动情况,其中三角形线、六角星形线和菱形线分别表示 SVM,TWSVM 和 TWSVM-U 3 个模型的准确率。

4.3 实验结果分析

说明性数据实验的结果(见图 3)表明,TWSVM-U 模型对标准 TWSVM 模型求解得到的分类超平面进行了修正,且修正程度与数据在超平面法方向的分散程度正相关,这与我们的动机相吻合,且分类准确率随着测试样本的增多趋于稳定。UCI 数据集上的实验结果(见图 4)表明,TWSVM-U 模型分类准确率随测试样本增加保持一定的稳定性,且分类准确率普遍提升,表 3 表明在随机产生测试样本点时 TWSVM-U 相对于另外两种模型仍具有一定的优势。

结束语 标准 TWSVM 模型未考虑数据的分布特征,因此在处理带有随机性的数据集时存在明显不足,本文通过考虑数据的分布特征对标准 TWSVM 模型中的距离进行加权,得到了 TWSVM-U 模型,一系列的实验表明,TWSVM-U 在处理带分布的数据集时能明显提高分类准确率。并且当数据

的随机性很弱时，TWSVM-U 模型会退化为 TWSVM 模型，因此 TWSVM-U 模型可以看作是一个更广义的分类模型。

接下来，进一步将模型推广到非线性分类并修正模型，使其对随机性很弱的数据集也能展现出优势。

参 考 文 献

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

[2] 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.

[3] JAYADEVA, 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.

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

[5] PENG X J, XU D. Twin Mahalanobis distance-based support vector machines for pattern recognition [J]. Information Sciences, 2012, 200: 22-37.

[6] CHEN S G, WU X J. A new fuzzy twin support vector machine for pattern classification [J]. International Journal of Machine Learning and Cybernetics, 2018, 9: 1553-1564.

[7] KHEMCHANDANI R, GOYAL K, CHANDRA S. TWSVR: Regression via Twin Support Vector Machine[J]. Neural Networks, 2016, 74: 14-21.

[8] WANG Z, SHAO Y H, BAI L, et al. Twin support vector machine for clustering[J]. IEEE Trans Neural Netw Learn Syst, 2015, 26(10): 2583-2588.

[9] CURRAN J M, BUCKLETON J S. An investigation into the performance of methods for adjusting for sampling uncertainty in DNA likelihood ratio calculations[J]. Forensic Science International Genetics, 2011, 5(5): 512-516.

[10] MA X, DJOUADI S M, CHARALAMBOUS C D. Optimal Filtering Over Uncertain Wireless Communication Channels [J]. IEEE Signal Processing Letters, 2011, 18(6): 359-362.

[11] POWELL W B, BOUZAIENE-AYARI B, BERGER J, et al. The Effect of Robust Decisions on the Cost of Uncertainty in Military Airlift Operations[J]. Acm Transactions on Modeling & Computer Simulation, 2011, 22(1): 1-19.

[12] BI J, ZHANG T. Support Vector Classification with Input Data Uncertainty[J]. Proc. of Neural Inf. proc. systems, 2004, 17: 161-168.

[13] WENZEL F, GALY-FAJOU T, DEUTSCH M, et al. Bayesian Nonlinear Support Vector Machines for Big Data[C]// Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Cham, Springer, 2017: 307-322.

[14] DEISENROTH M P, FOX D, RASMUSSEN C E. Gaussian Processes for Data-Efficient Learning in Robotics and Control [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(2): 408-423.

[15] LANCKRIET G R G, GHAOUI L E, BHATTACHARYYA C, et al. A robust minimax approach to classification[J]. Journal of Machine Learning Research, 2003, 3(3): 555-582.

[16] TZELEPIS C, MEZARIS V, PATRAS I. Linear Maximum Margin Classifier for Learning from Uncertain Data [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(12): 2948-2962.

[17] Han R J, Cao Q L. Fuzzy chance constrained least squares twin support vector machine for uncertain classification[J]. Journal of Intelligent & Fuzzy Systems, 2017, 33(5): 3041-3049.

[18] JAYADEVA, KHEMCHANDANI R, CHANDRA S. Twin Support Vector Machines: Models, Extensions and Applications [M]. Cham, Springer, 2017: 43-53.

[19] 李航. 统计学习方法[M]. 北京：清华大学出版社，2012：225-228.

[20] DUA D, EFI K T. UCI Machine Learning Repository [EB/OL]. http://archive.ics.uci.edu/ml.

（上接第 406 页）

[2] CONG G, JENSEN C S, WU D. Efficient retrieval of the top-k most relevant spatial web objects [C] // Proceedings of the VLDB Endowment. 2009: 337-348.

[3] AHN J, JO B, JUNG S. Multiple Domain-Based Spatial Keyword Query Processing Method Using Collaboration of Multiple IR-Trees[C]// Proceedings of the 7th International Conference on Emerging Databases. Springer, Singapore, 2018: 183-192.

[4] ZHENG K, SU H, ZHENG B, et al. Interactive top-k spatial keyword queries[C]// 2015 IEEE 31st International Conference on Data Engineering (ICDE). IEEE, 2015: 423-434.

[5] ZHANG D, CHAN C Y, TAN K L. Processing spatial keyword query as a top-k aggregation query[C]// Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval. ACM, 2014: 355-364.

[6] LIU X, CHEN L, WAN C. LINQ: A framework for location-aware indexing and query processing[J]. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(5): 1288-1300.

[7] GUTTMAN A. R-trees: a dynamic index structure for spatial searching[J]. ACM SIGMOD Record, 2016, 14(2): 47-57.

[8] ZANG X, HAO P, GAO X, et al. QDR-Tree: An Efcient Index Scheme for Complex Spatial Keyword Query[J]. arXiv preprint arXiv: 1804. 10726, 2018.

[9] JUNG H R, YONG S K, CHUNG Y D. QR-tree: An efficient and scalable method for evaluation of continuous range queries [J]. Information Sciences, 2014, 274(8): 156-176.

[10] NAIR S H, SINHA A, VACHHANI L. Hilbert's space-filling curve for regions with holes[C]// 2017 IEEE 56th Annual Conference on Decision and Control (CDC). IEEE, 2017: 313-319.

[11] CHEN Y Y, SUEL T, MARKOWETZ A. Efficient query processing in geographic web search engines[C]// Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data. ACM, 2006: 277-288.

[12] KHODAEI A, SHAHABI C, LI C. Hybrid indexing and seamless ranking of spatial and textual features of web documents [C]// International Conference on Database and Expert Systems Applications. Springer-Verlag, 2010: 450-466.

[13] SANTOKI K. Indexing and Searching on a Hadoop Distributed File System[EB/OL]. http://www.drdobbs.com/parallel/indexing-and_searching-on-a-hadoop-distr/226300241?pgno=3.

[14] CHRISTOFORAKI M, HE J, DIMOPOULOS C, et al. Text vs. space: efficient geo-search query processing[C]// Proceedings of the 20th ACM International Conference on Information and Knowledge Management. ACM, 2011: 423-432.

[15] GÖBEL R, HENRICH A, NIEMANN R, et al. A hybrid index structure for geo-textual searches[C]// Proceedings of the 18th ACM conference on Information and Knowledge Management. ACM, 2009: 1625-1628.

[16] WU D, MAN L Y, JENSEN C S, et al. Efficient continuously moving top-k spatial keyword query processing[C]// IEEE, International Conference on Data Engineering. IEEE, 2011: 541-552.