利用融合数据分布特征的模糊双支持向量机对恒星光谱分类

刘忠宝[1,2]，秦振涛[1]，罗学刚[1]，周方晓[1]，张 靖[1]

1. 攀枝花学院数学与计算机学院，四川 攀枝花 617000
2. 中北大学软件学院，山西 太原 030051

摘 要 恒星光谱分类是天文学研究的一个热点问题。随着观测光谱数量的急剧增加，传统的人工分类无法满足实际需求，急需利用自动化技术，特别是数据挖掘算法来对恒星光谱进行自动分类。关联规则、神经网络、自组织网络等数据挖掘算法已广泛应用于恒星光谱分类。其中，支持向量机(SVM)分类能力突出，被广泛应用于恒星光谱分类。该方法试图在两类样本之间找到一个最优分类面将两类分开。该方法具有较高的时间复杂度，计算效率有限。双支持向量机(TWSVM)的出现有效地解决了SVM面临的效率问题。该方法通过构造两个非平行的分类面将两类分开，每一类靠近某个分类面，而远离另一个分类面。TWSVM的计算效率较之传统SVM提高近4倍，因此，自TWSVM提出后便受到研究人员的持续关注。但上述方法在分类决策时，一方面没有考虑数据的分布特征，另一方面较易受噪声点和奇异点的影响，分类效率难以显著提升。鉴于此，在双支持向量机的基础上，提出融合数据分布特征的模糊双支持向量机(TWSVM-SDP)。该方法引入线性判别分析(LDA)的类间离散度和类内离散度，用以表征光谱数据的分布性状；引入模糊隶属度函数用以降低噪声点和奇异点对分类结果的影响。在SDSS DR8恒星光谱数据集上的比较实验表明，与支持向量机SVM、双支持向量机TWSVM等传统分类方法相比，融合数据分布特征的模糊双支持向量机TWSVM-SDP具有更优的分类能力。该方法亦存在一定的局限性，其中一大难题是其无法处理海量光谱数据。接下来将利用大数据处理技术，来对所提方法在大数据环境下的适应性展开进一步研究。

关键词 恒星光谱；分类；数据分布特征；模糊隶属度；双支持向量机

中图分类号：TP391 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2019)04-1307-05

引 言

作为一种典型的智能分类模型，支持向量机(support vector machine，SVM)具有优良的分类能力，已被广泛应用于恒星光谱分类。近年来，与支持向量机相关的研究成果不断涌现，较为典型的成果有：张怀福等提出分别利用小波包分析和支持向量机对天体光谱进行特征提取和智能分类方法[1]；Peng等利用支持向量机从大型巡天项目中搜寻类星体[2]；刘忠宝等在支持向量机中引入流形判别分析，用以提升支持向量机的分类能力[3]；Shi等利用支持向量机从SDSS DR9中对发射线星系进行分类[4]；Liu在支持向量机中引入将线性判别分析(linear discriminant analysis，LDA)中的类间离散度和类内离散度，用以表征数据的分布性状，确保支持向量机在分类决策时将数据的分布性状考虑在内[5]。

尽管支持向量机在实际应用中表现优良，但其时间复杂度过高，无法处理较大规模数据的分类问题。双支持向量机(twin support vector machine，TWSVM)[6]的提出有效地解决了SVM面临的上述问题。TWSVM的计算效率较之传统SVM提高近4倍。然而，该方法亦面临一些挑战：(1)分类决策时只关注各类数据之间的绝对间隔，并未考虑光谱数据的分布特征；(2)易受到噪声点和奇异点的影响。鉴于此，提出融合数据分布特征的模糊双支持向量机(fuzzy twin support vector machine with spectral distribution properties，TWSVM-SDP)。在该方法中，引入LDA中的类间离散度和类内离散度用以表征恒星光谱数据的分布特征；引入模糊隶属度函数用以降低噪声点和奇异点对分类结果的影响。SDSS DR8恒星光谱数据集的比较实验表明，与SVM，TWSVM等传统分类方法相比，TWSVM-SDP具有更优的分类能力。

收稿日期：2018-03-17，**修订日期**：2018-08-05

基金项目：国家自然科学基金项目(U1731128，11803080)，山西省自然科学基金项目(201601D011042)，山西省高等学校创新人才支持计划(2016)，中北大学杰出青年基金支持计划(2017)资助

作者简介：刘忠宝，1981年生，中北大学软件学院教授 e-mail：liuzb@nuc.edu.cn

1 背景知识

给定 N 个样本集为 $X=\{(x_1\ y_1),\cdots,(x_N,\ y_N)\}$，其中 $x_i\in R^m$，$y_i\in\{-1,1\}$为类别标签。当 $1\leqslant i\leqslant N_1$ 时，$y_i=1$；当 $1\leqslant i\leqslant N_2$ 时，$y_i=-1$，且 $N=N_1+N_2$。

1.1 支持向量机

支持向量机通过构造一个分类超平面将两类隔开。设分类超平面为 $w^{\mathrm{T}}x+b=0$，分类间隔为 $2/\|w\|$，SVM 的最优化问题可描述为

$$\min_{w,\,b,\,\xi_i}\ \frac{1}{2}w^{\mathrm{T}}w+C\sum_{i=1}^{N}\xi_i$$

$$\text{s. t.}\quad y_i(w^{\mathrm{T}}x_i+b)\geqslant 1-\xi_i,\ \xi_i\geqslant 0\quad i=1,\cdots,N$$

其中，C 为惩罚因子，ξ_i 为松弛因子。

由 Lagrangian 定理将原问题转化为如下对偶问题

$$\max_{\alpha}\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{1}-\frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{\alpha}$$

$$\text{s. t.}\quad \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{Y}=\mathbf{0},\ \boldsymbol{\alpha}\geqslant 0$$

其中 $\boldsymbol{\alpha}=[\alpha_1,\cdots,\alpha_N]^{\mathrm{T}}$，$\mathbf{1}=[1,\cdots,1]^{\mathrm{T}}$，$\boldsymbol{Q}=[y_iy_jx_i^{\mathrm{T}}x_j]$，$\boldsymbol{Y}=[y_1,\cdots,y_N]^{\mathrm{T}}$，$\mathbf{0}=[0,\cdots,0]^{\mathrm{T}}$。

1.2 双支持向量机

双支持向量机 TWSVM 试图找到两个非平行的分类面将两类分开。设矩阵 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 分别表示属于 1 类和−1 类的数据集，设两类超平面方程分别为 $w_+^{\mathrm{T}}x+b_+=0$ 和 $w_-^{\mathrm{T}}x+b_-=0$，则 TWSVM 的最优化问题可表示为以下形式：

(TWSVM1)

$$\min_{w_+,\,b_+,\,\xi}\ \frac{1}{2}(Aw_++e_1b_+)^{\mathrm{T}}(Aw_++e_1b_+)+c_1e_2^{\mathrm{T}}\xi$$

$$\text{s. t.}\quad -(Bw_++e_2b_+)^{\mathrm{T}}+\xi\geqslant e_2\quad \xi\geqslant 0$$

(TWSVM2)

$$\min_{w_-,\,b_-,\,\xi}\ \frac{1}{2}(Bw_-+e_2b_-)^{\mathrm{T}}(Bw_-+e_2b_-)+c_2e_1^{\mathrm{T}}\xi$$

$$\text{s. t.}\quad (Aw_-+e_1b_-)^{\mathrm{T}}+\xi\geqslant e_1\quad \xi\geqslant 0$$

其中，c_1 和 c_2 为惩罚因子；e_1 和 e_2 为全由 1 组成的列向量，ξ 为松弛因子。

由 Lagrangian 定理将原问题转化为如下对偶形式：

(TWSVM1)

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

$$\text{s. t.}\quad 0\leqslant\alpha\leqslant c_1$$

其中 $H=[A\quad e_1]$，$G=[B\quad e_2]$。

(TWSVM2)

$$\max_{\gamma}e_1^{\mathrm{T}}\gamma-\frac{1}{2}\gamma^{\mathrm{T}}P(Q^{\mathrm{T}}Q)P^{\mathrm{T}}\gamma$$

$$\text{s. t.}\quad 0\leqslant\gamma\leqslant c_2$$

其中 $P=[A\quad e_1]$，$Q=[B\quad e_2]$。

1.3 LDA

LDA 是一种经典的特征提取方法，该方法提取的特征具有很好的可分性，即同类之间的距离尽可能近，异类之间的距离尽可能远。其优化问题可描述为

$$J(w)=\max_{w}\frac{w^{\mathrm{T}}S_Bw}{w^{\mathrm{T}}S_Ww}$$

其中 $S_B=\sum_{i=1}^{c}\frac{N_i}{N}(\bar{x}_i-\bar{x})(\bar{x}_i-\bar{x})^{\mathrm{T}}$，$S_W=\sum_{i=1}^{c}\sum_{j=1}^{N_i}\frac{1}{N}(x_{ij}-\bar{x}_i)(x_{ij}-\bar{x}_i)^{\mathrm{T}}$，$c$ 表示类别数，N_i 表示第 i 类的规模，$\bar{x}_i$ 和 $\bar{x}$ 分别表示第 i 类均值和所有光谱数据均值。

2 融合数据分布特征的模糊双支持向量机

融合数据分布特征的模糊双支持向量机 TWSVM-SDP 在双支持向量机 TWSVM 的基础上，引入 LDA 中的类间离散度 S_W 和类内离散度 S_B 用以表征光谱数据的分布特征，引入模糊隶属度函数 s 用以降低噪声点和奇异点对分类结果的影响。设两类超平面方程分别为 $w_+^{\mathrm{T}}x+b_+=0$ 和 $w_-^{\mathrm{T}}x+b_-=0$，TWSVM-SDP 的最优化问题可描述为：

(TWSVM-SDP1)

$$\min_{w_+,\,b_+,\,\xi}\ \frac{1}{2}(Aw_++e_1b_+)^{\mathrm{T}}(Aw_++e_1b_+)+\frac{1}{2}\beta_1w_+^{\mathrm{T}}(S_W-S_B)w_++c_1s_1e_2^{\mathrm{T}}\xi\tag{1}$$

$$\text{s. t.}\quad -(Bw_++e_2b_+)^{\mathrm{T}}+\xi\geqslant e_2\quad \xi\geqslant 0\tag{2}$$

(TWSVM-SDP2)

$$\min_{w_-,\,b_-,\,\xi}\ \frac{1}{2}(Bw_-+e_2b_-)^{\mathrm{T}}(Bw_-+e_2b_-)+\frac{1}{2}\beta_2w_-^{\mathrm{T}}(S_W-S_B)w_-+c_2s_2e_1^{\mathrm{T}}\xi\tag{3}$$

$$\text{s. t.}\quad (Aw_-+e_1b_-)^{\mathrm{T}}+\xi\geqslant e_1\quad \xi\geqslant 0\tag{4}$$

其中 c_1 和 c_2 为惩罚因子；β_1 和 β_2 为平衡参数；e_1 和 e_2 为全由 1 组成的列向量。

令 TWSVM-SDP1 的 Lagrangian 函数为

$$L(w_+,\ b_+,\ \xi,\ \alpha,\ \beta)=\frac{1}{2}(Aw_++e_1b_+)^{\mathrm{T}}(Aw_++e_1b_+)+\frac{1}{2}\beta_1w_+^{\mathrm{T}}(S_W-S_B)w_++c_1s_1e_2^{\mathrm{T}}\xi-\alpha^{\mathrm{T}}[-(Bw_++e_2b_+)^{\mathrm{T}}+\xi-e_2]-\beta^{\mathrm{T}}\xi\tag{5}$$

其中 Lagrangian 乘子 $\alpha\geqslant 0$，$\beta\geqslant 0$。

L 分别对 w_+，b_+，ξ 求导并令导数等于 0，可得式(6)—式(8)

$$\frac{\partial L}{\partial w_+}=0\Leftrightarrow A^{\mathrm{T}}(Aw_++e_1b_+)+\beta_1(S_W-S_B)w_++B^{\mathrm{T}}\alpha=0\tag{6}$$

$$\frac{\partial L}{\partial b_+}=0\Leftrightarrow e_1^{\mathrm{T}}(Aw_++e_1b_+)+e_2^{\mathrm{T}}\alpha=0\tag{7}$$

$$\frac{\partial L}{\partial\xi}=0\Leftrightarrow c_1s_1e_2-\alpha-\beta=0\quad\Leftrightarrow 0\leqslant\alpha\leqslant c_1s_1\tag{8}$$

由式(6)—式(7)可得

$$w_+=\beta_1^{-1}(S_W-S_B)^{-1}(A-B)^{\mathrm{T}}\alpha\tag{9}$$

$$b_+=-e_2^{\mathrm{T}}\alpha-e_1^{\mathrm{T}}A\beta_1^{-1}(S_W-S_B)(A-B)^{\mathrm{T}}\alpha\tag{10}$$

将式(9)—式(10)代入式(5)，可得

$$\max_{\alpha}\frac{1}{2}\alpha^{\mathrm{T}}\alpha+\frac{1}{2}\beta_1^{-1}\alpha^{\mathrm{T}}(A-B)[(S_W-S_B)^{-1}]^{\mathrm{T}}\cdot(S_W-S_B)(S_W-S_B)^{-1}(A-B)^{\mathrm{T}}\alpha\tag{11}$$

令 $G=A-B$，$H=(S_W-S_B)^{-1}$，式(11)转化为

$$\max_{\alpha}\alpha^{\mathrm{T}}\alpha+\beta_1^{-1}\alpha^{\mathrm{T}}GH^{\mathrm{T}}H^{-1}HG^{\mathrm{T}}\alpha\tag{12}$$

$$\text{s.t.}\quad 0 \leqslant \alpha \leqslant c_1 s_1 \tag{13}$$

同理可得 TWSVM-SDP2 的对偶形式

$$\max_{\gamma} \gamma^{\mathrm{T}} \gamma + \beta_2^{-1} \gamma^{\mathrm{T}} Q H^T H^{-1} H Q^{\mathrm{T}} \gamma \tag{14}$$

$$\text{s.t.}\quad 0 \leqslant \gamma \leqslant c_2 s_2 \tag{15}$$

其中 $Q=B-A$，$H=(S_W-S_B)^{-1}$。

TWSVM-SDP 的决策函数为

$$f(x) = \arg \min_{k=+,-} | w_k^{\mathrm{T}} x + b_k | \tag{16}$$

2.1　算法描述

TWSVM-SDP 的算法流程如下：

输入：训练数据集 X_Train

输出：测试数据集 X_Test 中样本的类属

步骤 1：将目标光谱分为训练数据集和测试数据集；

步骤 2：利用 Lagrangian 乘子法将 TWSVM-SDP 最优化问题转化为如式(12)—式(15)所示的对偶形式；

步骤 3：在训练数据集 X_Train 上运行的 TWSVM-SDP 算法，得到分类依据；

步骤 4：计算如式(16)所示的决策函数；

步骤 5：利用步骤 4 得到的决策函数对测试数据集中的任一样本 $x \in X$_Test 判定类属，从而得到 TWSVM-SDP 算法的分类精度。

3　实验分析

实验采用美国斯隆巡天发布的 SDSS DR8 恒星光谱数据作为实验数据集。实验对象是 K 型、F 型、G 型恒星光谱，其中 K 型光谱包括 K1，K3，K5 和 K7 四类次型，其信噪比(signal noise ratio，SNR)区间为(50，60)；F 型光谱包括 F2，F5，F9 三类次型，其中 F2 次型光谱信噪比区间为(50，60)，F5 次型光谱信噪比区间为(65，70)，F9 次型信光谱噪比区间为(75，80)；G 型光谱包括 G0，G2，G5 三类次型，其中 G0 次型光谱信噪比区间为(50，60)，G2 次型光谱信噪比区间为(55，60)，G5 次型信噪光谱信噪比区间为(50，70)，实验数据集如表 1(a)—(c)所示。

表 1(a)　K 型恒星光谱规模

Table 1(a)　The total number of K stars

Stellar subclass type	K1	K3	K5	K7
SNRs	(50，60)	(50，60)	(50，60)	(50，60)
Number	3 302	3 176	3 048	1 132

表 1(b)　F 型恒星光谱规模

Table 1(b)　The total number of F stars

Stellar subclass type	F2	F5	F9
SNRs	(50，60)	(65，70)	(75，80)
Number	1 416	1 671	1 535

通过与 SVM，TWSVM 等分类方法的比较来验证所提方法 TWSVM-SDP 的有效性。上述分类方法的性能与所选的参数有关。本文选用 10 折交叉验证法获取实验参数，而参数的选择采用网格搜索法。在 SVM 和 TWSVM 中，惩罚因子在网格{0.01，0.05，0.1，0.5，1，5，10}中搜索。实验选取基于距离的模糊隶属度函数。分别选取实验对象的 30%，40%，50%，60%和 70%作为训练数据集，而剩余样本作为测试数据集。实验结果如表 2(a)—(c)所示，其中括号前的值表示样本规模，括号中的值表示所占比例。

表 1(c)　G 型恒星光谱规模

Table 1(c)　The total number of G stars

Stellar subclass type	G0	G2	G5
SNRs	(50，60)	(55，60)	(50，70)
Number	1 188	1 463	295

表 2(a)　K 型恒星数据集上的比较实验结果

Table 2(a)　The comparative experimental results on the K-type dataset

Training size	Test size	SVM	TWSVM	TWSVM-SDP
30%(3 197)	70%(7 461)	0.668 7	0.709 7	0.829 9
40%(4 263)	60%(6 395)	0.710 2	0.729 6	0.850 4
50%(5 329)	50%(5 329)	0.810 7	0.820 2	0.870 5
60%(6 395)	40%(4 263)	0.867 7	0.880 9	0.940 4
70%(7 461)	30%(3 197)	0.891 1	0.910 9	0.960 0
Average classification accuracy		0.789 7	0.810 3	0.890 2

表 2(b)　F 型恒星数据集上的比较实验结果

Table 2(b)　The comparative experimental results on the F-type dataset

Training size	Test size	SVM	TWSVM	TWSVM-SDP
30%(1 387)	70%(3 275)	0.600 3	0.650 4	0.750 0
40%(1 865)	60%(2 797)	0.660 0	0.700 8	0.780 5
50%(2 331)	50%(2 331)	0.710 0	0.760 6	0.821 5
60%(2 797)	40%(1 865)	0.800 0	0.841 3	0.900 8
70%(3 275)	30%(1 387)	0.785 1	0.839 9	0.891 8
Average classification accuracy		0.711 1	0.758 6	0.828 9

表 2(c)　G 型恒星数据集上的比较实验结果

Table 2(c)　The comparative experimental results on the G-type dataset

Training size	Test size	SVM	TWSVM	TWSVM-SDP
30%(884)	70%(2 062)	0.618 3	0.641 6	0.730 8
40%(1 178)	60%(1 768)	0.681 6	0.701 4	0.810 0
50%(1 473)	50%(1 473)	0.740 0	0.780 7	0.874 4
60%(1 768)	40%(1 178)	0.790 3	0.831 1	0.920 9
70%(2 062)	30%(884)	0.848 4	0.869 9	0.957 0
Average classification accuracy		0.735 7	0.764 9	0.858 6

由表 2(a)—(c)可以看出：随着训练样本规模的增大，

SVM，TWSVM 和 TWSVM-SDP 三种分类方法的分类精度呈上升趋势(训练样本为 70%的 F 型恒星光谱的情况除外)。在不同规模的训练样本情况下，TWSVM-SDP 较之 SVM 和 TWSVM 具有更优的分类能力。从平均分类性能看，与 SVM 和 TWSVM 相比，TWSVM-SDP 的分类表现更优。究其原因，TWSVM-SDP 一方面继承了 TWSVM 计算效率较高的优势，另一方面其在分类决策时考虑了光谱数据的分布性状，通过引入模糊隶属度函数来降低噪声点和奇异点对分类结果的影响。因此，与 SVM 和 TWSVM 相比，TWSVM-SDP 在恒星光谱分类中表现更优。

4 结 论

针对 SVM 面临的计算效率低的问题，研究人员提出 TWSVM，该方法的计算效率较之传统 SVM 提高近 4 倍。然而，上述方法在分类决策时并未考虑数据的分布特征，且易受噪声点和奇异点的影响，因而分类效率难以显著提升。鉴于此，提出融合数据分布特征的模糊双支持向量机 TWSVM-SDP。该方法在双支持向量机 TWSVM 的基础上，通过引入 LDA 中的类间离散度 S_W 和类内离散度 S_B 用以表征光谱数据的分布特征，引入模糊隶属度函数 s 用以降低噪声点和奇异点对分类结果的影响。SDSS DR8 恒星光谱数据集上的比较实验表明，与传统的 SVM 和 TWSVM 相比，TWSVM-SDP 具有更优的分类能力。然而，TWSVM-SDP 无法有效处理大规模光谱分类问题，接下来将利用大数据处理技术对所提方法在大数据环境下的适应性展开进一步研究。

References

[1] ZHANG Huai-fu, ZHAO Rui-zhen, LUO A-li(张怀福，赵瑞珍，罗阿理). Journal of Beijing Jiaotong University(北京交通大学学报), 2008, 32(2): 30.

[2] Peng N B, Zhang Y X, Zhao Y H, et al. Monthly Notices of the Royal Astronomical Society, 2012, 425(4): 2599.

[3] LIU Zhong-bao, WANG Zhao-ba, ZHAO Wen-juan(刘忠宝，王召巴，赵文娟). Spetroscopy and Spectral Analysis(光谱学与光谱分析), 2014, 34(1): 263.

[4] Shi F, Liu Y Y, Sun G L, et al. Monthly Notices of the Royal Astronomical Society, 2015, 453(1): 122.

[5] Liu Z B. Journal of Astrophysics and Astronomy, 2016, 37(2): 9.

[6] Jayadeva R K, Khemchandani R, Chandra S. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(5): 905.

Stellar Spectra Classification by Support Vector Machine with Spectral Distribution Properties

LIU Zhong-bao[1,2], QIN Zhen-tao[1], LUO Xue-gang[1], ZHOU Fang-xiao[1], ZHANG Jing[1]
1. School of Mathematics and Computer Science, Panzhihua University, Panzhihua 617000, China
2. School of Software, North University of China, Taiyuan 030051, China

Abstract Stellar spectra classification is one of hot spots in astronomy. With hundreds and thousands of spectra obtained by researchers, it is a big challenge to process them manually. It's urgent to apply the automatic technologies, especially the data mining algorithms, to classify the stellar spectra. Neural networks, self organization mapping, association rules and other data mining algorithms have been utilized to classify the stellar spectra in recent years. In these methods, Support Vector Machine (SVM), as a typical classification method, is widely used in the stellar spectra classification due to its good learning capability and excellent classification performance. The basic idea of standard SVM is to find an optimal separating hyper-plane between the positive and negative samples. Its time complexity is so high that its classification efficiencies can't be greatly improved. Twin Support Vector Machine (TWSVM) is proposed to deal with the above problem. It aims at generating two non-parallel hyper-planes such that each plane is close to one class and as far as possible from the other one. The learning speed of TWSVM is approximately four times faster than the classical SVM. The limitation of TWSVM is that it doesn't take spectral distribution properties into consideration, and its efficiencies are prone to be influenced by noise and singular points. In view of this, Fuzzy Twin Support Vector Machine with Spectral Distribution Properties (TWSVM-SDP) is proposed, in which between-class scatter and within-class scatter in Linear Discriminant Analysis (LDA) is introduced to describe the spectral distribution properties and the fuzzy membership function is introduced to decrease the influences of noise and singular points. Comparative experiments on SDSS DR8 stellar spectra datasets verity TWSVM-SDP performs better than SVM and TWSVM. However, some limitations exist in TWSVM-SDP, for example, how to deal with the mass spectra is quite difficult to solve. We will research the adaptability of our proposed method in the big data environment based on big data technologies.

Keywords Stellar spectra; Classification; Spectral distribution properties; Fuzzy membership function; Twin support vector machine

(Received Mar. 17, 2018; accepted Aug. 5, 2018)