利用带无标签数据的双支持向量机对恒星光谱分类

刘忠宝[1,2]，雷宇飞[1]，宋文爱[2]，张 静[2]，王 杰[3]，屠良平[4]

1. 泉州信息工程学院软件学院，福建 泉州 362000
2. 中北大学软件学院，山西 太原 030051
3. 中国科学院新疆天文台，新疆 乌鲁木齐 830011
4. 辽宁科技大学理学院，辽宁 鞍山 114051

摘 要 恒星光谱分类是天文技术与方法领域一直关注的热点问题之一。随着观测设备持续运行和不断改进，人类获得的光谱数量与日俱增。这些海量光谱为人工处理带来了极大挑战。鉴于此，研究人员开始关注数据挖掘算法，并尝试对这些光谱进行数据挖掘。近年来，神经网络、自组织映射、关联规则等数据挖掘方法广泛应用于恒星光谱分类。在这些方法中，支持向量机(SVM)以其强大的学习能力和高效的分类性能而备受推崇。SVM的基本思想是试图在两类样本之间找到一个最优分类面将两类分开。SVM在求解时，通过将其最优化问题转化为具有(QP)形式的凸问题，进而得到全局最优解。尽管该方法在实际应用中表现优良，但为了进一步提高其分类能力，有的学者提出双支持向量机(TSVM)。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。TSVM的计算效率较之传统SVM提高近4倍，因此，自TSVM提出后便受到研究人员的持续关注，并出现若干改进算法。在恒星光谱分类中，一般分类算法都是根据历史观测光谱来建立分类模型，其中最关键的是对光谱进行人工标注，这项工作极为繁琐，且容易犯错。如何利用已标记的光谱以及部分无标签的光谱来建立分类模型显得尤为重要。因此，提出带无标签数据的双支持向量机(TSVMUD)用以实现对恒星光谱智能分类的目的。该方法首先将光谱分为训练数据集和测试数据集两部分；然后，在训练集上进行学习，得到分类依据；最后利用分类依据对测试集上的光谱进行验证。继承了双支持向量机的优势，更重要的是，在训练集上学习分类模型过程中，不仅考虑有标记的训练样本，也考虑部分未标记的样本。一方面提高了学习效率，另一方面得到更优的分类模型。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TSVM以及K近邻(KNN)等传统分类方法相比，带无标签数据的双支持向量机TSVMUD具有更优的分类能力。然而，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。该工作将借鉴海量数据随机采样思想，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；智能分类；双支持向量机；无标签数据

中图分类号：TP391 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2019)03-0948-05

引 言

恒星光谱分类是天文技术和方法领域研究的热点问题之一。随着观测设备持续运行和不断改进，人类获得的光谱数量与日俱增。这些海量光谱为人工处理带来了巨大挑战。鉴于此，人们开始考虑利用自动化技术，特别是数据挖掘算法来处理这些光谱。近年来，不断涌现出一些卓有成效的研究成果。Bazarghan等提出基于人工神经网络的自组织映射(self-organzing Map，SOM)算法，该算法可以直接对光谱分类，而无需进行预先训练[1]。Navaro等提出的人工神经网络系统通过在温度、光度敏感的光谱中选择线强指数集进行训练，实现对低信噪比光谱的分类[2]。Bolton等利用高质量的光谱源实现光谱的分类[3]。Hernandez等利用稀疏表示和词典学习方法实现光谱分类[4]。Gray等建立了一个MKCLASS专家系统来对MK光谱进行分类[5]。Fuentes等将多层感知

收稿日期：2018-02-11，**修订日期**：2018-06-19

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项目(2016)，中北大学杰出青年基金支持计划项目(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

神经网络和 PCA 方法用于恒星光谱的次型分类，该方法对于矮星和巨星光度型的分类可信度达到 95%以上。他们还将 PCA 方法应用于恒星光谱降维，这为恒星光谱降维方法的研究提供了重要参考[6]。有研究提出一种新的基于支持向量机(support vector machine，SVM)的非活动天体与活动天体的自动分类方法。姜斌等针对 LAMOST 光谱的特点，首先利用拉普拉斯特征映射(Laplacian eigenmap，LE)对光谱进行特征提取，然后利用神经网络进行分类[7]。Bu 等利用等距特征映射(isometric feature map，ISOMAP)和支持向量机[8]以及局部线性嵌入(locally linear embedding，LLE)算法[9]来对恒星光谱进行分类。Cai 等利用加权频繁模式树发现恒星光谱的关联规则[10]。刘忠宝等针对大规模光谱数据分类问题，提出非线性集成分类方法[11]，该方法首先将大规模光谱数据分为若干子集，在每个子集上运行传统分类器并得到分类结果，最后将各子集的结果进行集成，得到最终分类结果。此外，Liu 还提出“LPP+SVM”分类策略，即首先利用保局投影(locality preserving projections，LPP)算法进行高维光谱降维处理，然后利用支持向量机进行分类[12]。

上述分类方法均属于有监督的学习方法，其工作流程一般分为两个阶段：一个是训练，另一个是预测。在训练阶段，上述方法要求事先给出带类别标签的训练样本。然而，光谱的类别信息往往依赖于人工标注，这项工作极为繁琐，且容易犯错。此外，通过对历史观测光谱学习得到的分类模型，对于新获取的光谱(还未进行人工标注)未必有效，如果重新训练分类模型，时间代价又过于庞大，因此，如何在分类模型中融入无标记样本进行学习值得深入研究。在众多分类模型中，双支持向量机(twin support vector machine，TSVM)的计算效率较之传统 SVM 提高近 4 倍，因此，自 TSVM 提出后便受到研究人员的持续关注。鉴于此，本文在 TSVM 的基础上，提出带无标签数据的双支持向量机(twin support vector machine with unlabeled data，TSVMUD)用以对恒星光谱智能分类。通过在 SDSS DR8 恒星光谱数据集上与 SVM，TSVM，KNN(K nearest neighhor)等分类方法的比较实验来验证所提方法的有效性。

1　双支持向量机

双支持向量机(TSVM)试图找到两个分类面将两类分开。设将两类样本分别存放于矩阵 $\boldsymbol{A}$ 和 $\boldsymbol{B}$。分别定义如下两个分类面[见式(1)]

$$w_+^{\mathrm{T}}x+b_+=0 \text{ 和 } w^{\mathrm{T}}x+b_-=0 \tag{1}$$

在 TSVM 中，每一类都接近于相对应的分类面，而远离另一分类面，并且两个分类面之间应有一定距离。基于上述分析，可得 TSVM 的最优化问题[见式(2)和式(3)]

$$\min_{w_+,\,b_+,\,\xi_+} \frac{1}{2}\|Aw_++e_+\,b_+\|^2+C_1e^T\,\xi_+$$
$$\text{s. t.}\quad -(Bw_++e_-\,b_+)\geqslant e_--\xi_+,$$
$$\xi_+\geqslant 0 \tag{2}$$

$$\min_{w_-,\,b_-,\,\xi_-} \frac{1}{2}\|Bw_-+e_-\,b_-\|^2+C_2e_+^T\,\xi_-$$
$$\text{s. t.}\quad (Aw_-+e_+\,b_-)\geqslant e_+-\xi_-,$$
$$\xi_-\geqslant 0 \tag{3}$$

其中，C_1 和 C_2 为惩罚因子；$\xi_{(\pm)}=[\xi_1^{(\pm)},\ \xi_2^{(\pm)},\ \xi_3^{(\pm)},\ \cdots,\ \xi_l^{(\pm)}]$为松弛因子，其保证算法具有一定的容错性；$e_+$ 和 e_- 均表示全 1 列向量。

一个新的样本点 x 的类属判定取决于如下的决策函数[式(4)]

$$\text{class}\quad i=\arg\min_{k=+,\,-}\frac{|\,x^{\mathrm{T}}w_{(k)}+b_{(k)}\,|}{\|\,w_{(k)}\,\|} \tag{4}$$

其中 i=｛−1，+1｝。

2　带无标签数据的双支持向量机

2.1　最优化问题

假设给定训练数据集 $\widetilde{T}=T\cup U=\{(x_1,\ y_1),\ (x_2,\ y_2),\ \cdots,\ (x_l,\ y_l)\}\cup\{x_1^*,\ x_2^*,\ \cdots,\ x_u^*\}$，其中 T 为有标签数据集，U 为无标签数据集。$x_i\in R^n$，$y_i=\{+1,\ -1\}$，其中 $i=1,\ 2,\ \cdots,\ l$；$x_m^*\in R^n$，其中 $m=1,\ 2,\ \cdots,\ u$。l 和 m 分别表示有标签数据和无标签数据规模。

带无标签数据的双支持向量机 TSVMUD 在双支持向量机 TSVM 的基础上，引入无标签数据，因此，在建立优化问题时，可以将目标函数分为两部分：一部分针对有标签数据分类，另一部分针对无标签数据分类。TSVMUD 的最优化问题表示为式(5)和式(6)

$$\min_{w_+,\,b_+,\,\xi_+,\,\psi}\frac{1}{2}\|Aw_++e_+\,b_+\|^2+C_1e^T\,\xi_++De_u^T\psi$$
$$\text{s. t.}\quad -(Bw_++e_-\,b_+)\geqslant e_--\xi_+,$$
$$(Uw_++e_ub_+)+\psi\geqslant(\varepsilon-1)e_u,$$
$$\xi_+\geqslant 0,\ \psi\geqslant 0 \tag{5}$$

$$\min_{w_-,\,b_-,\,\xi_-,\,\psi^*}\frac{1}{2}\|Bw_-+e_-\,b_-\|^2+C_2e_+^T\,\xi_-+De_u^T\psi^*$$
$$\text{s. t.}\quad (Aw_-+e_+\,b_-)\geqslant e_+-\xi_-,$$
$$-(\boldsymbol{U}w_-+e_ub_-)+\psi^*\geqslant(\varepsilon-1)e_u,$$
$$\xi_-\geqslant 0,\ \psi^*\geqslant 0 \tag{6}$$

其中矩阵 $\boldsymbol{U}$ 用于存放无标签数据；C_1 和 C_2 为针对有标签数据的惩罚因子，D 为针对无标签数据的惩罚因子；$\xi_{(\pm)}=[\xi_1^{(\pm)},\ \xi_2^{(\pm)},\ \xi_3^{(\pm)},\ \cdots,\ \xi_l^{(\pm)}]$为针对有标签数据的松弛因子，$\psi^{(*)}=[\psi_1^{(*)},\ \psi_2^{(*)},\ \psi_3^{(*)},\ \cdots,\ \psi_u^{(*)}]$为针对无标签数据的松弛因子；$e_+$，$e_-$ 以及 e_u 均表示全 1 列向量。

根据拉普拉斯乘子法，引入拉普拉斯算子 α_+ 和 β_+，可得(5)式的对偶形式[式(7)]

$$\min_{\alpha_+,\,\beta_+}\frac{1}{2}(G^{\mathrm{T}}\alpha_+-J^{\mathrm{T}}\beta_+)^{\mathrm{T}}(H^{\mathrm{T}}H)^{-1}(G^{\mathrm{T}}\alpha_+-J^{\mathrm{T}}\beta_+)$$
$$-e_-^{\mathrm{T}}\,\alpha_+-(\varepsilon-1)e_u^{\mathrm{T}}\beta_+$$
$$\text{s. t.}\quad 0\leqslant\alpha_+\leqslant C_1e_-,$$
$$0\leqslant\beta_+\leqslant De_u \tag{7}$$

同理可得(6)式的对偶形式[式(8)]

$$\min_{\alpha_-,\,\beta_-}\frac{1}{2}(H^{\mathrm{T}}\alpha_--J^{\mathrm{T}}\beta_-)^{\mathrm{T}}(G^{\mathrm{T}}G)^{-1}(H^{\mathrm{T}}\alpha_--J^{\mathrm{T}}\beta_-)$$
$$-e_+^{\mathrm{T}}\,\alpha_--(\varepsilon-1)e_u^{\mathrm{T}}\beta_-$$
$$\text{s. t.}\quad 0\leqslant\alpha_-\leqslant C_2e_+,$$

$$0 \leqslant \beta_- \leqslant De_u \tag{8}$$

其中 $H=[A, e_+]$, $G=[B, e_-]$, $J=[U, e_u]$。

一个新的样本点 x 的类属判定取决于如式(4)所示的决策函数，如式(9)和式(10)

$$\begin{bmatrix} w_+ \\ b_+ \end{bmatrix} = -(H^{\mathrm{T}}H)^{-1}(G^{\mathrm{T}}\alpha_+ - J^{\mathrm{T}}\beta_+) \tag{9}$$

$$\begin{bmatrix} w_- \\ b_- \end{bmatrix} = -(G^{\mathrm{T}}G)^{-1}(H^{\mathrm{T}}\alpha_- - J^{\mathrm{T}}\beta_-) \tag{10}$$

2.2 算法描述

TSVMUD 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。训练数据集中包含一定比例的有标签数据和无标签数据。

步骤 2：利用拉格朗日乘子法将 TSVMUD 最优化问题转化为如式(7)和式(8)所示的对偶形式；

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

步骤 4：计算如式(4)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 TSVMUD 算法的分类精度。

3 实验分析

实验采用美国斯隆巡天发布是 SDSS DR8 的恒星光谱数据作为实验数据集。实验对象是 K 型光谱中信噪比在 50～60 之间的 3 302 条 K1 次型光谱，3 176 条 K3 次型光谱，3 048条 K5 次型光谱以及 1 132 条 K7 次型光谱。其中随机选取 80%的光谱作为有标签样本，其余的 20%样本去掉其类别标签，作为无标签样本。实验的软硬件环境包括：3GHz Pentium4 CPU，4G RAM，Windows 7，MATLAB 7.0。

通过与 SVM，TSVM 和 KNN 等传统分类方法的比较来验证所提方法 TSVMUD 的有效性。上述分类方法的性能与所选的参数有关。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 TSVM 中，惩罚因子在网格{0.01，0.05，0.1，0.5，1，5，10}中搜索；在 K 邻近算法(K nearest neighbor，KNN)中，参数 K 在网格{1，5，10，15，20，25，30}中搜索。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。上述数据集中有标签样本和无标签样本的比例为 4∶1。由于 SVM 和 TSVM 是经典的有监督学习方法，即上述方法无法对无标签数据进行训练。因此，为了表示方便，需要事先对无标签数据进行随机分类处理。KNN 方法对无标签数据进行 K 近邻计算，即首先找到与无标签数据最近的 K 个有标签的近邻，然后根据“少数服从多数”的原则，确定无标签数据的类属。实验结果如表 1—表 4 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。

由表 1—表 4 可以看出，随着训练样本规模的增大，SVM，TSVM，KNN 和 TSVMUD 等分类精度呈上升趋势。在 K1，K3，K5 和 K7 次型数据集上，与 SVM，TSVM 和 KNN 相比，TSVMUD 的分类精度均最优。从平均精度角度看，TSVMUD 的平均分类精度远高于其他三种方法。产生上述实验结果的原因是：由于 SVM，TSVM 和 KNN 属于监督学习方法，其无法对无标签数据进行学习。本实验为了比较方便，故对上述三种方法进行了预处理，这种预处理具有一定的随机性和不确定性，并对实验结果有一定影响。而本文所提的 TSVMUD 方法擅长处理混有标签和无标签数据的

表 1 K1 次型光谱上的实验结果

Table 1 The experimental results on the K1 subclass spectra

训练数据规模	测试样本规模	SVM	TSVM	KNN	TSVMUD
991(30%)	2 311(70%)	0.321 1	0.411 9	0.563 4	0.660 8
1 321(40%)	1 981(70%)	0.402 8	0.491 2	0.648 1	0.731 4
1 651(50%)	1 651(50%)	0.503 3	0.530 0	0.702 0	0.791 0
1 981(60%)	1 321(40%)	0.582 1	0.582 1	0.751 7	0.840 3
2 311(70%)	991(30%)	0.650 9	0.689 2	0.797 1	0.891 0
平均精度		0.492 0	0.540 1	0.692 5	0.782 9

表 2 K3 次型光谱上的实验结果

Table 2 The experimental results on the K3 subclass spectra

训练数据规模	测试样本规模	SVM	TSVM	KNN	TSVMUD
953(30%)	2 223(70%)	0.380 1	0.421 1	0.521 4	0.610 0
1 270(40%)	1 906(70%)	0.450 7	0.501 6	0.581 3	0.691 0
1 588(50%)	1 588(50%)	0.530 2	0.570 5	0.641 1	0.740 6
1 906(60%)	1 270(40%)	0.629 9	0.684 3	0.711 0	0.830 7
2 223(70%)	953(30%)	0.698 8	0.711 4	0.771 2	0.878 3
平均精度		0.537 9	0.577 8	0.645 2	0.750 1

表 3 K5 次型光谱上的实验结果

Table 3 The experimental results on the K5 subclass spectra

训练数据规模	测试样本规模	SVM	TSVM	KNN	TSVMUD
914(30%)	2 134(70%)	0.402 5	0.472 4	0.560 0	0.671 5
1 219(40%)	1 829(60%)	0.481 1	0.519 9	0.620 6	0.741 4
1 524(50%)	1 524(50%)	0.551 2	0.601 0	0.690 9	0.802 5
1 829(60%)	1 219(40%)	0.648 1	0.760 5	0.784 2	0.851 5
2 134(70%)	914(30%)	0.678 3	0.760 4	0.811 8	0.902 6
平均精度		0.552 4	0.622 8	0.693 5	0.793 9

表 4 K7 次型光谱上的实验结果

Table 4 The experimental results on the K7 subclass spectra

训练数据规模	测试样本规模	SVM	TSVM	KNN	TSVMUD
340(30%)	792(70%)	0.352 3	0.411 6	0.521 5	0.603 5
453(40%)	679(60%)	0.421 2	0.490 4	0.593 5	0.695 1
566(50%)	566(50%)	0.545 9	0.561 8	0.682 0	0.775 6
679(60%)	453(40%)	0.604 9	0.653 4	0.752 8	0.827 8
792(70%)	340(30%)	0.650 0	0.694 1	0.791 2	0.894 1
平均精度		0.514 9	0.562 3	0.668 2	0.759 2

分类问题，因此，在不同规模的训练样本上，TSVMUD 均具有最优的分类性能。

4　结　论

针对已有恒星光谱分类方法面临的无法处理无标签光谱的不足，提出带无标签数据的双支持向量机 TSVMUD。该方法在双支持向量机 TSVM 的基础上，引入无标签数据以实现对恒星光谱智能分类的目的。该方法在训练集上学习分类模型时，不仅考虑有标记的训练样本，也考虑部分未标记的样本。这样，一方面提高了学习效率，另一方面得到更优的分类模型。在 SDSS DR8 恒星光谱数据集上与 SVM，TSVM 和 KNN 等传统分类方法相比，所提方法 TSVMUD 具有更优的分类精度。然而，该方法亦存在无法处理海量光谱数据的不足。进一步将借鉴海量数据随机采样思想，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。

References

[1] Bazarghan M. Astrophysics and Space Science, 2012, 337(1): 93.
[2] Navarro S G, Corradi R L M, Mampaso A. Astronomy and Astrophysics, 2012, 538(1): 143.
[3] Bolton A S, Schlegel D J, Aubourg E, et al. The Astronomical Journal, 2012, 144(5): 507.
[4] Hernandez R D, Barreto H P, Robles L A, et al. Experimental Astronomy, 2014, 38(1): 193.
[5] Gray R O, Corbally C J. The Astronomical Journal, 2014, 147(4): 80.
[6] Fuentes O, Gulati R K. Proceedings of the 7th Texas-Mexico Conference on Astrophysics: Flows, Blows and Glows, 2001. 209.
[7] JIANG Bin, LI Zi-xuan, QU Mei-xia, et al(姜　斌，李紫宣，曲美霞，等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2016, 36(7): 2275.
[8] Bu Y D, Chen F Q, Pan J C. New Astronomy, 2014, 28: 35.
[9] Bu Y D, Pan J C, Jiang B, et al. Publications of the Astronomical Society of Japan, 2013, 65(4): 173.
[10] Cai J H, Zhao X J, Sun S W, et al. Research in Astronomy and Astrophysics, 2013, 13(3): 334.
[11] Liu Z B, Song L P, Zhao W J. Monthly Notices of the Royal Astronomical Society, 2016, 455(4): 4289.
[12] Liu Z B. Journal of Astrophysics and Astronomy, 2016, 37(2): 1.

Stellar Spectra Classification by Support Vector Machine with Unlabeled Data

LIU Zhong-bao[1,2], LEI Yu-fei[1], SONG Wen-ai[2], ZHANG Jing[2], WANG Jie[3], TU Liang-ping[4]

1. School of Software, Quanzhou University of Information Engineering, Quanzhou　362000, China
2. School of Software, North University of China, Taiyuan　030051, China
3. Xinjiang Astronomical Observatory, Chinese Academy of Sciences, Urumqi　830011, China
4. School of Science, University of Science and Technology Liaoning, Anshan　114051, China

Abstract　Stellar spectra classification is one of hot spots in astronomical techniques and methods. With continuous operation and improvement of observation apparatus, hundreds and thousands of spectra were obtained by researchers, which presented challenges to process them manually. In view of this, data mining algorithms have attracted more attentions, and have been utilized to deal with the 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 algorithms, Support Vector Machine (SVM) is much more popular 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. SVM as a convex programming problem has a unique optimal solution, which can be posed as a quadratic programming (QP) problem. In order to further improve the classification efficiency, Twin Support Vector Machine (TSVM) has been proposed. 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 TSVM is approximately four times faster than that of the classical SVM. TSVM receives many attentions since it shows low computational complexity, and many variants of TSVM have been proposed in literatures. During the process of stellar spectra classification, the classification model is built based on the observation data. The key step is to manually label the spectra, which is time-consuming and painstaking. Therefore, how to construct the spectra classification model based on the labeled and unlabeled spectra is a problem de-

serving study. In order to effectively classify the stellar spectra, Twin Support Vector Machine with Unlabeled Data (TSVMUD) is proposed in this paper. In TSVMUD, the stellar spectra are firstly divided into two parts, one is for training, and the other is for test. Then, the proposed method TSVMUD is utilized on the training data and the classification model is obtained. At last, the spectra in the test dataset are verified by the classification model. TSVMUD not only preserve the advantage of low computational complexity, but also improve the classification efficiency by taking both the labeled and unlabeled data into consideration. The comparative experiments on the SDSS datasets verify that TSVMUD performs better than the traditional classifiers, such as SVM, TSVM, KNN (K Nearest Neighbor). However, some limitations exist in TSVMUD, for example, how to deal with the mass spectra is quite difficult to solve. Inspired by random sampling, we will research the adaptability of our proposed method in the big data environment based on big data technologies.

Keywords Stellar spectra; Intelligent classification; Twin support vector machine; Unlabeled data

(Received Feb. 11, 2018; accepted Jun. 19, 2018)