第26卷，第7期
2006年7月

光 谱 学 与 光 谱 分 析
Spectroscopy and Spectral Analysis

Vol. 26, No. 7, pp1368-1372
July, 2006

基于小波降噪与支持向量机的恒星光谱识别研究

邢 飞，郭 平*

北京师范大学信息科学与技术学院，北京 100875

摘 要 提出了一种对恒星光谱识别的新方法。根据恒星光谱数据的特性，我们以支持向量机为核心技术构建光谱识别器。由于恒星光谱数据通常含有较高的噪声，如果直接进行分类，识别率往往较低。因此作者首先采用小波分析的方法对原始光谱数据进行降噪预处理，提取光谱的特征，然后馈送到支持向量机完成对光谱数据的最终识别。利用实际光谱数据(Jacoby，1984)对所提出的技术进行检测，实验结果表明使用这种小波分析结合支持向量机的技术的识别效果要优于使用支持向量机结合主分量分析降维技术的识别方法。另外，作者还比较了支持向量机与传统甄别分析的分类性能，对实际及合成光谱进行实验的结果显示了支持向量机的识别正确率不但优于常见的5种甄别分析方法的识别率，而且有较强的泛化能力。

主题词 恒星光谱识别；支持向量机；小波降噪；主分量分析；甄别分析

中图分类号：TP391.4 **文献标识码**：A **文章编号**：1000-0593(2006)07-1368-05

引 言

恒星光谱分类是天体光谱自动识别中的重要组成部分，是恒星和星系天文学中的一个重要研究方向，它的出现大大地促进了恒星天文学的发展。由于光谱数据数量十分巨大，手工进行光谱分类是不现实的。因此，为了达到对光谱数据的自动处理，需要研究快速自动的恒星光谱分类技术。

目前，在模式识别研究领域中人们已经研究出了很多分类方法[1]，其中甄别分析是一种常见的监督性学习分类技术。近年来出现了多种基于甄别分析的分类方法，它们均被应用于恒星光谱分类中[2]。二次甄别分析(quadratic discriminant analysis，QDA)是一种被广泛使用的甄别分析方法，它在训练样本数量充足的情况下有着较好得分类效果。但在实际的恒星光谱分类问题中，由于缺乏相关的专家知识，往往很难获得足够的恒星光谱训练样本数据。另一方面，由于每条光谱数据的维数相当高，估量协方差矩阵(estimated covariance matrix)会出现奇异的情况，使得分类出现病态(Ill-posed)问题，导致严重的错分情况。人们常用降维和重整化的方法来解决这种高维小训练样本问题，线性甄别分析(linear discriminant analysis，LDA)可以被看作一种重整化分类器，它用公共协方差矩阵代替类协方差矩阵，当总训练样本数目大于数据维数时，用此方法可以很好地解决估量协方差矩阵奇异的问题。但是当总训练样本数目小于数据维数时，LDA中用到的公共矩阵也会变为奇异矩阵。在这种情况下，重整化甄别分析(regularized discriminant analysis，RDA)通过对类协方差矩阵和公共协方差矩阵的整合可以达到很好的分类效果。而留一法协方差矩阵估计方法(leave-one-out covariance matrix estimate，LOOC)和库勒巴克-雷伯勒信息量度重整化方法(Kullback-Leibler information measure，KLIM)通过分别加入对角矩阵和带系数的单位矩阵也可很好地解决此类问题。RDA，LOOC和KLIM这三种方法中所用到的参数均是由留一交叉检验(Leave-one-out cross validation)方法确定的。

支持向量机(support vector machines，简称SVM)是Vapnik等[3,4]提出的一类新型机器学习方法。由于其出色的学习性能，该技术已成为机器学习界的研究热点，并在很多领域都得到了成功的应用[5]。由于支持向量机在高维小训练样本情况下有着很好泛化能力，因此这个特性可以很好的应用于恒星光谱分类中。同时考虑到原始光谱数据含有较高的噪声，而小波变换是近年来发展起来的一种很好的信号分析手段，它具有良好的时频局域化特性，能通过伸缩和平移对信号进行多分辨率分析，能聚焦到对象的任意细节，因此我们首先采用小波变换的方法对原始光谱数据进行降噪，然后将降噪后的光谱数据作为支持向量机的输入完成恒星光谱识别。

收稿日期：2005-03-16，**修订日期**：2005-06-26

基金项目：国家自然科学基金(60275002)和教育部留学回国人员科研启动基金资助

作者简介：邢 飞，1981年生，北京师范大学信息科学与技术学院硕士研究生 * 通讯联系人

1　基本原理

1.1　小波变换原理

小波变换可同时在时域和频域上分析信号的局部特性。平方可积函数 $f(t)\in L^2(R)$的连续小波变换定义为

$$WT_f(a,b)=\frac{1}{\sqrt{a}}\int_{-\infty}^{\infty}f(t)\Psi\cdot\left(\frac{t-b}{a}\right)\mathrm{d}t=\langle f(t),\Psi_{a,b}(t)\rangle \tag{1}$$

其中，小波变换的核函数 $\Psi_{a,b}=\frac{1}{\sqrt{a}}\Psi\left(\frac{t-b}{a}\right)$是母小波 $\Psi(t)$的时间平移 b 和尺度伸缩 a 的结果。

应用小波变换降噪主要是利用小波变换能较好地表示信号的局部结构特征，以及信号局部结构特征下所表现的奇异性不同于噪声所表现的奇异性。在恒星光谱分析中我们直接利用 Mallat 算法将信号分解为高频和低频信息，设定一个截断尺度，使频率高于此尺度下的小波空间向量全部置为零，然后进行信号重构即可达到去噪的作用。在小波系数的取舍问题上，采用 Donoho 提出的通用阈值算法[6]。通用阈值 T 由下式定义：

$$T=\sqrt{2\log_2 S} \tag{2}$$

其中对于小波包变换计算，$S=N\ \ln N$；对于离散小波变换计算，$S=N$。Donoho 还提出了选择系数的软硬阈值法。在硬阈值法中，所有变换系数的绝对值与通用阈值 T 作比较。如果一个系数值小于通用阈值 T，则等于 0，

$$\text{硬阈值}\ \alpha_j=\begin{cases}0 & |\alpha_j|<T\\ \alpha_j & |\alpha_j|\geqslant T\end{cases} \tag{3}$$

上式中，α_j 代表小波域上的系数。软阈值法的系数由下式确定：

$$\text{软阈值}\ \alpha_j=\begin{cases}0 & |\alpha_j|<T\\ |\alpha_j-T| & |\alpha_j|\geqslant T\end{cases} \tag{4}$$

1.2　支持向量机原理

SVM 是以结构化风险最小化(SRM)代替常用的经验风险最小化(ERM)作为优化准则，其基本思想是对于非线性可分样本，将其输入向量经非线性变换映射到另一个高维空间 Z 中，在变换后的空间中寻找一个最优的分界面(超平面)，使其推广能力最好。以两类模式的分类为例说明其基本原理。

设线性可分的样本集 n 有个样本(x_i,y_i)，其中 $i=1,2,\cdots,n,\boldsymbol{x}\in R^N$，$y\in\{-1,1\}$是类别标号。在高维空间中，将两类样本无错分开的分类超平面满足

$$\boldsymbol{w}\cdot\boldsymbol{x}+b=0,\quad \boldsymbol{w}\in R^N,\ b\in R \tag{5}$$

通过对向量系数 $\boldsymbol{w}$ 进行归一化，可以使所有样本满足 $|g(\boldsymbol{x}_i)|\geqslant 1$，这样分类间隔就等于 $2/\|\boldsymbol{w}\|$，因此使分类间隔最大实际上就是使 $\|\boldsymbol{w}\|$ 最小；考虑到线性不可分的情况，引入了软边缘最优超平面的概念，即引入非负变量 ξ_i，最优分类面问题可表示成如下的约束优化问题，即在式(6)的约束下

$$y_i(\boldsymbol{x}_i\boldsymbol{w}+b)\geqslant 1-\xi_i,\quad i=1,2,\cdots,n \tag{6}$$

求函数

$$\boldsymbol{\Phi}(\boldsymbol{w},\xi)=\frac{1}{2}\|\boldsymbol{w}\|^2+C\left(\sum_{i=1}^{n}\xi_i\right) \tag{7}$$

的最小值。为此，可利用 Lagrange 函数把原问题转化为较简单的 Wolfe 对偶问题：在约束条件

$$\sum_{i=1}^{n}y_i\alpha_i=0,\quad 0\leqslant\alpha_i\leqslant C,\ i=1,2,\cdots,n \tag{8}$$

之下对 α_i 求解下列函数的最大值：

$$Q(\alpha)=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j(\boldsymbol{x}_i,\boldsymbol{x}_j) \tag{9}$$

求解上述问题后得到的最优分类函数是

$$f(\boldsymbol{x})=\mathrm{sgn}\{(\boldsymbol{w}\cdot\boldsymbol{x}_i)+b\}=\mathrm{sgn}\left\{\sum_{i=1}^{n}\alpha_i y_i(\boldsymbol{x}_i,\boldsymbol{x})+b\right\} \tag{10}$$

通过上述讨论可以看出，最优分类识别函数(10)只包含待测样本与支持向量之间的内积。因此对于非线性可分的特征空间，考虑通过一个非线性映射 T 将特征 $\boldsymbol{x}$ 映射到高维线性特征空间 F 中。高维线性空间中的内积可以定义为

$$K(\boldsymbol{x}_i,\boldsymbol{x}_j)=T(\boldsymbol{x}_i)\cdot T(\boldsymbol{x}_j) \tag{11}$$

$K(\boldsymbol{x}_i,\boldsymbol{x}_j)$称为核函数，它的选择需要满足 Mercer 条件。采用不同的内积核函数将导致不同的支持向量机算法，目前采用较多的 3 类核函数包括多项式内积函数，径向基函数和 S 型内积函数[3,4]。

支持向量机方法的优点在于没有必要知道映射 T 的具体形式，而只需定义高维空间中的内积运算 $K(\boldsymbol{x}_i,\boldsymbol{x}_j)$即可，这样即使变换后空间维数增加很多，计算的复杂度也没有太大的变化。

2　光谱分析实验

本实验中所用到的恒星光谱数据来自天文数据中心(astronomical data center，ADC)。实验中采用了 Jacoby(1984)中的 161 条光谱数据[7]。按照温度由高到低，一共有 7 种主要的恒星光谱数据，它们分别是 O-He Ⅱ谱线、B-He Ⅰ谱线、A-H 谱线、F-Ca Ⅱ谱线、G-strong 金属线、K-bands 扩散线和 M-very 红外线。其中 O，B 型星温度最高，可达二三万度，所以电离谱线比较强，中性原子谱线比较弱。而 K，M 型星温度较低，只有三四千度，所以电离谱线比较弱，中性原子谱线比较强，并出现分子带谱线。图 1 显示了这 161 条光谱前三维主分量投影。

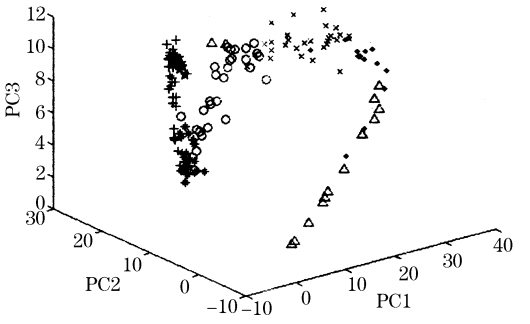

Fig. 1　Distribution of stellar spectra in first three principal component space

在分类前，我们首先对原始恒星光谱进行归一化处理，将每条光谱的流量强度调整到 0～1 之间，这样就可以用同样的标准比较光谱间的差别。从图 2 中我们可以看到归一化后的七类恒星谱线的分布。

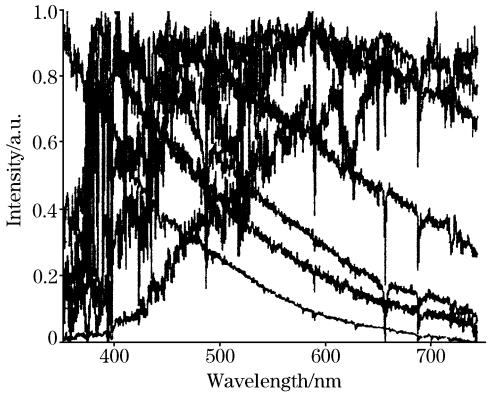

Fig. 2 Seven main types of stellar spectra

在一般的监督学习方法中，包括两个数据集，一个用于构造分类器，称为训练样本集；另一个用于检验分类器的性能，称为测试样本集。我们采用 bootstrap 技术进行实验，每类恒星光谱中随机选取 10 条作为训练样本，剩下的作为检验样本用于计算分类正确率(correct classification rate，CCR)。实验重复 25 次计算 CCR 的均值和方差。

首先，我们将归一化后的光谱数据直接作为 SVM 的输入进行分类，从表 1 可以看到，分类正确率可达到 81.66%，标准差为 3.75。

Table 1 Mean and standard deviation of the classification accuracy of SVM，PCA＋SVM and wavelet＋SVM

方法	CCR/%	标准差
原始数据	81.66	3.75
PCA	81.30	2.90
wavelet	93.26	3.08

考虑到光谱数据通常含有较高的噪声，为提高分类正确率，我们采用小波变换方法对原始光谱数据进行降噪，降噪后的光谱数据作为输入送入 SVM 进行分类。从图 3 中可以看出小波去噪的效果，实验结果显示了分类正确率高达 93.26%，标准差为 3.08。通过假设检验(t 检验)可以看出，在显著性水平 $\alpha=0.05$ 的情况下，利用小波降噪结合 SVM 的方法对恒星光谱数据进行分类的正确率要明显优于直接用 SVM 进行分类的正确率(p-value$=1.97\times10^{-8}$)。

主分量分析(principal component analysis，PCA)[8]是一种用较少数量的特征对样本进行描述以达到降低特征空间维数的方法，它广泛应用于信号处理、统计学和神经网络计算，是数据降维和去噪的有力工具。作为比较，我们首先用 PCA 对原始数据进行预处理，从图 4 中可以看出，前 10 PCs 的重建错误率只有 0.43%，所以我们将原始数据降至 10 维，然后送入 SVM 完成最终分类。从图 3 中可以看到经 PCA 降维重建后的谱线，实验结果显示了此种方法的分类正确率只有 81.30%，通过假设检验可以看出，在显著性水平 $\alpha=0.05$ 的情况下，用 PCA＋SVM 进行分类的分类正确率和直接用 SVM 分类相比较并未有显著提高(p-value$=0.71$)。同时也可以看出 wavelet＋SVM 的方法比 PCA＋SVM 的方法有更高的识别率(p-value$=2.01\times10^{-10}$)。导致应用 PCA 进行降维预处理分类正确率未有显著提高的原因是由 SVM 的基本思想决定的，SVM 是将非线性可分样本经非线性变换映射到高维可分的空间，也就是说 SVM 在高维样本的处理上有很好的性能。

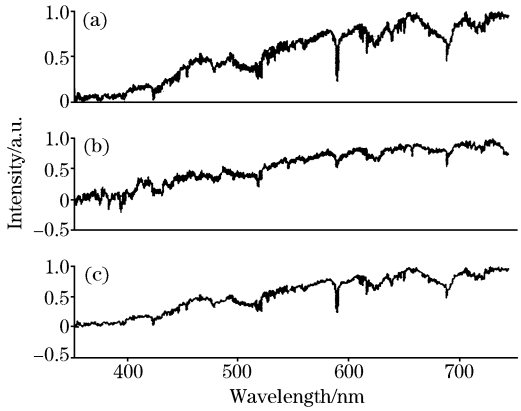

Fig. 3 De-noising result

(a)：original spectrum；
(b)：with PCA (10 PCs reconstructed spectrum)；
(c)：with wavelet de-noising

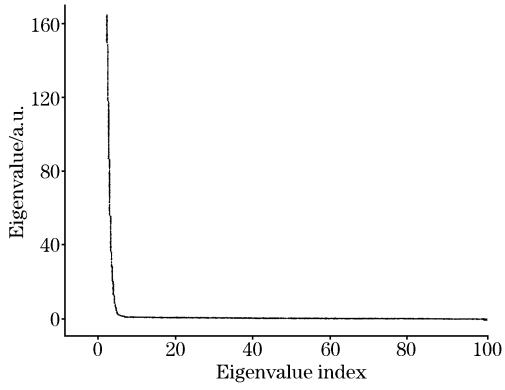

Fig. 4 Eigenvalue in decreasing order

对于 1.2 节中提到的 3 种 SVM 核函数，我们在实验中对其性能进行了比较。表 2 显示了实验结果，可以看到，以径向基函数和 S 型内积函数为核函数的 SVM 对恒星光谱的识别性能要明显好于多项式内积函数。

Table 2 Classification accuracy comparison of the three kernels of SVM

核函数	CCR/%	标准差
多项式内积函数	88.62	3.37
径向基函数	93.26	3.08
S 型内积函数	93.39▯▯3.29	

甄别分析是一种传统的监督学习分类技术。在接下来的实验中我们比较上述我们提到的五种甄别分析方法与支持向量机的分类效果。所用到的光谱数据是来自天文数据中心的进化合成光谱库，数据包括了 457 条光谱，可被分为三大类。每条谱线含有 1 221 个波长点，波长范围在 9.1 nm 到 1.6×10^5 nm 之间。由于传统的甄别分析对数据的维数比较敏感，实验中我们将原始恒星光谱数据利用主分量分析法分别降至 100 维，40 维，10 维和 6 维，然后采用 bootstrap 技术，从每类光谱中随机选取 50 个样本作为进行实验，将每类的 50 个样本随机分成两部分，15 个作为训练样本，剩下的 35 个作为测试样本用于计算分类正确率。实验重复 25 次计算 CCR 的均值和方差。表 3 显示了分类效果，括号中是标准差。

Table 3　Classification accuracy comparison of the conventional discriminant analysis and the SVM

方法	$d=100$	$d=40$	$d=10$	$d=6$
QDA	N/A	N/A	94.08(1.80)	96.21(0.84)
LDA	N/A	70.51(8.21)	93.87(1.33)	94.88(0.47)
RDA	92.67(3.14)	91.45(5.22)	95.11(0.63)	96.02(0.52)
LOOC	77.24(4.65)	83.48(3.25)	87.52(1.62)	88.55(1.10)
KLIM	94.29(5.77)	95.64(3.92)	96.51(0.49)	96.88(0.51)
SVM	97.52(1.59)	97.90(1.83)	97.14(0.78)	97.62(1.65)

从实验结果中我们可以看出，当 PCA 投影维数为 6 维和 10 维时，六种甄别分析方法均有较稳定的分类性能。而当 PCA 投影维数逐渐升高时，QDA 和 LDA 分类器先后出现了不适定问题，导致了分类效果极不稳定，而 3 种重整化分类方法通过调整估量协方差矩阵避免了矩阵奇异的问题。在 3 种重整化分类器中，RDA 分类器随维数增加，分类性能略有下降；LOOC 分类器虽未出现不适定问题，但分类性能下降得较快；而 KLIM 始终保持着较高的识别率。由于重整化分类器都是采用交叉检验技术进行重整化参数估计，所以在计算时间上要远高于一般分类器，而对于大型的恒星光谱分类系统，往往有数万甚至数十万光谱需要处理，因此分类器的计算效率是不可忽视的。对于支持向量机，实验结果显示 PCA 投影维数对其识别率影响非常小，它的识别率要高于文中提到的 5 种甄别分析方法，并且有着更稳定的分类性能(更低的标准差)。在计算效率方面，支持向量机要略逊色于 QDA 和 LDA，但优于 3 种重整化分类器 RDA，LOOC 和 KLIM。

3　结果与讨论

本文提出了一种恒星光谱识别的新技术，该技术结合了小波变换和支持向量机技术。首先利用小波变换的方法对原始光谱进行降噪，将那些与光谱特征无关的噪声去掉，然后将降噪后的光谱送入支持向量机完成最终分类。对实际光谱的识别结果表明，这种恒星光谱识别器有很好的分类效果，我们从统计学的角度用假设检验(t-检验)证明了它的分类效果明显优于直接用 SVM 进行分类的效果，也优于用 PCA 结合 SVM 的分类效果。通过比较支持向量机和甄别分析对恒星光谱的识别性能我们可以看到，支持向量机要优于甄别分析，而且不受光谱数据维数的限制，在计算效率上也要优于重整化分类器。尽管这种恒星光谱识别模型还需要进一步的评估，但本文实验结果已经充分表明了其准确稳定的分类性能，相信这种分类器将会被广泛地应用在光谱识别中。

参　考　文　献

[1] GUO Ping, QIN Dong-mei, HU Zhan-yi(郭　平，覃冬梅，胡占义). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2003, 23(4): 811.

[2] WANG Xi, XING Fei, GUO Ping. Proceedings of SPIE, 2003, 5286: 758.

[3] Vapnik V. The Nature of Statistical Learning Theory. New York: Springer-Verlag, 1995.

[4] Cortes C, Vapnik V. Machine Learning, 1995, 20: 273.

[5] QIN Dong-mei, HU Zhan-yi, ZHAO Yong-heng(覃冬梅，胡占义，赵永恒). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2004, 24(4): 507.

[6] Donoho D L. IEEE Transaction on Information Theory, 1995, 41: 613.

[7] Jacoby G H, Hunter D A, Christian C A. J. Suppl. Ser., 1984, 56: 278.

[8] Jolliffe I T, Principal Component Analysis, New York: Springer-Verlag, 1986.

Stellar Spectral Recognition Based on Wavelet De-Noising and SVM

XING Fei, GUO Ping*
School of Information Science and Technology, Beijing Normal University, Beijing 100875, China

Abstract The present paper describes a new technique for stellar spectral recognition. Considering the characteristics of stellar spectral data, support vector machine (SVM) was adopted to build a recognition system as kernel. Because stellar spectral data sets are usually extremely noisy, the correct classification rate of direct applying SVM is low. Consequently, wavelet de-noising method was proposed to reduce noise first and extract the main characteristics of stellar spectra. Then SVM was used for the recognition. Based on the real-world stellar spectra contributed by Jacoby et al. (1984), it has proven that there will be a better performance using this composite classifier which combines wavelet and SVM than using SVM with principle component analysis data dimension reduction technique. From the experiment of comparison of discriminant analysis and SVM based on stellar spectra for evolutionary synthesis, we can see that the correct classification rate of SVM is higher than that of discriminant analysis methods, and a well generalization ability is achieved.

Keywords Stellar spectral classification; Suppor vector machine; Wavelet de-noising; Principal component analysis; Discriminant analysis

(Received Mar. 16, 2005; accepted Jun. 26, 2005)

* Corresponding author