基于小波特征的星系光谱分类

刘 蓉[1], 段福庆[2], 刘三阳[1], 吴福朝[2]
(1. 西安电子科技大学数学系, 陕西西安 710071; 2. 中科院自动化所模式识别国家重点实验室, 北京 100080)

摘 要: 提出了一种新的星系光谱分类方法. 首先, 对原始光谱进行四级小波分解, 选择主要包含谱线信息的第四级小波系数作为光谱的小波特征; 然后, 利用主分量分析对光谱的小波特征进行特征压缩, 得到光谱的识别特征; 最后, 利用 Fisher 线性判别分析实现分类. 该方法能够在红移值未知的情况下, 对流量未定标的星系光谱进行识别. 通过实验与其他几种分类方法进行了比较. 实验结果表明, 本文方法具有较强的鲁棒性, 在流量未定标情况下的识别效果优于其他几种分类方法.

关键词: 光谱分类; 小波特征; 主分量分析; Fisher 线性判别分析
中图分类号: TP391 文献标识码: A 文章编号: 0372-2112 (2005) 11-2059-04

Spectral Classification of Galaxy Based on Wavelet Feature

LIU Rong[1], DUAN Fu-qing[2], LIU San-yang[1], WU Fu-chao[2]
(1. *Department of Mathematics, Xidian University, Xi'an, Shaanxi 710071, China;* 2. *National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. Beijing 100080, China*)

Abstract: A technique for classification of galaxy spectra is proposed. At first, a four-level wavelet decomposition of the original spectrum is performed, and the wavelet coefficient at the fourth level, which mainly includes the information of spectral lines, is chosen as the wavelet feature of the spectrum. Secondly, principal components analysis is used to compress the wavelet feature and to get the recognition feature of the spectrum. Finally, Fisher linear discriminant analysis is employed for classification. This approach can recognize the galaxy spectrum whose flux is uncalibrated and redshift is unknown. Comparisons with several other classification techniques were made by experiments. Experiment results show the proposed method is robust and superior to other methods under the condition that flux is uncalibrated.

Key words: spectral classification; wavelet feature; principal components analysis; Fisher linear discriminant analysis

1 引言

我国正在建造的 LAMOST[1] (大天区面积多目标光纤光谱望远镜) 在建成后的数年内将能获得 10^7 左右的星系光谱. 面对这些海量数据, 传统的以人工为主的光谱处理方法显然不能满足需求, 研究光谱的自动处理方法迫在眉睫.

天体光谱主要由连续谱、谱线和噪声组成. 连续谱对应光谱中缓变的低频成份, 谱线分为吸收线和发射线. 图 1 左侧为两条观测光谱, 其中, 横轴为波长, 纵轴为相对流量, 较粗的曲线为拟合出的连续谱, 吸收线是位于连续谱下方的凸出区域, 发射线是位于连续谱上方的凸出区域, 图中已标出部分谱线. 连续谱和谱线都是光谱的重要特征, 但 LAMOST 不对光谱进行流量定标处理, 因而产生的光谱的连续谱不反映真实的强度分布信息, 只有谱线才是可靠的特征. 根据光谱上是否存在较强的发射谱线, 星系分为正常星系 *NG* 和活动星系 *AG*, 前者的光谱基本上没有强发射谱线. 按形态来划分, *NG* 分为四种 (*E*0、*S*0、*Sa*、*Sb*), *AG* 分为七种 (*Sc*, *Sb*1-*Sb*6). 在 LAMOST 光谱处理中, 光谱分类是红移测量的前奏, 分类的正确与否直接影响到红移测量的可信度. 本文的目的是实现正常星系和活动星系的光谱识别. 由于红移的影响, 在固定的观测波段, 相同类型的光谱也会有不同的表现. 因此红移值未知会给星系光谱的识别带来很大难度. 国际上发表的众多涉及星系的光谱分类方法[2~4]都是在红移值已知的前提下先将光谱移回静止波长再进行分类. 文献[5]结合主分量分析 PCA 和最优判别面实现星系的光谱分类, 其中 PCA 是针对原始光谱进行的, 选取的所有样本都是经过流量定标的光谱, 连续谱的特征在其中起着较大的作用. 将连续谱归一化 (用原始光谱除以连续谱以去除连续谱的影响) 后, 该方法不再有效. 本文提出了一种基于光谱的小波特征的星系光谱分类方法. 由于所选择的小波特征主要包含谱线的信息, 因此适用于流量未定标的星系光谱分类.

2 光谱特征提取

在流量未定标的情况下, 光谱的信息主要体现在谱线上, 由于红移和可观测区间的影响, 一条光谱上只有很少的几根谱线. 对于不同的光谱, 谱线的类型、位置和强度都不尽相同,

收稿日期: 2004-10-26; 修回日期: 2005-07-29
基金项目: 国家 863 计划(No. 2003AA133060)

因此可以说光谱的信息是一种局部信息. 小波变换最大的特点是能够对信号进行多尺度局域分析, 因而它经常被用于光谱处理领域[6~8]. 本文选取光谱的小波特征的原因是: 首先, 连续谱对应光谱中缓变的低频成份, 谱线和噪声对应光谱中的高频成份, 光谱的小波系数主要反映了谱线和噪声的信息, 因此分类器受连续谱的影响较小; 其次, 信号和噪声在小波变换的各尺度上具有不同的传播特性: 随着尺度的增大, 小波系数中信号和噪声对应的模极大值分别增大和减小, 连续若干次小波变换之后, 与噪声对应的模极大值已基本去除或幅值很小. 当输入的识别特征中主要包含谱线信息时, 分类器对噪声的敏感性就会大大降低.

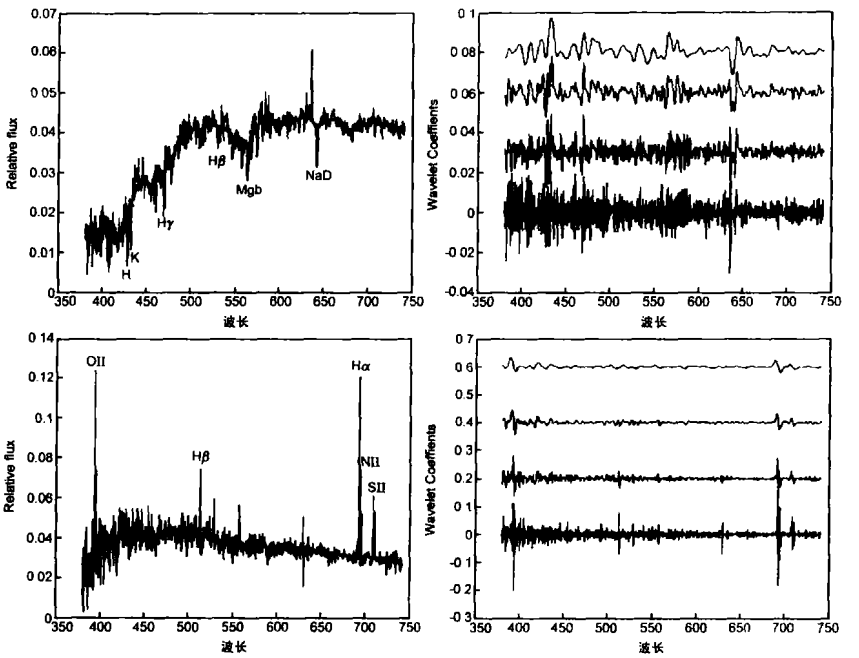

图 1 实测光谱(左)及其四级小波系数(右, 从下到上尺度分别为 1、2、3、4)

2.1 光谱的小波变换

我们首先从多分辨分析引入小波变换. 令 $\varphi(t)$ 为 $L^2(R)$ 上的函数, 称满足下列条件的子空间 $\{V_m, m\in Z\}$ 为 $L^2(R)$ 的一个多分辨分析, 如果

①$V_m \subset V_{m-1}$, $m\in Z$; ②$\bigcap_{m\in Z} V_m = \{0\}$; ③$\bigcup_{m\in Z} V_m = L^2(R)$; ④$x(t)\in V_m \Leftrightarrow x(2t)\in V_{m-1}$; ⑤$\varphi(t)\in V_0$, 且 $\{\varphi(t-n)\}_{n\in Z}$ 是 V_0 的标准正交基. 称 $\varphi(t)$ 为多分辨分析 $\{V_m, m\in Z\}$ 的尺度函数. 由④⑤可知, $\{\varphi_{m,n}(t)=2^{-m/2}\varphi(2^{-m}t-n)\}_{n\in Z}$ 构成 V_m 的标准正交基. 令 $\{W_m, m\in Z\}$ 为 $L^2(R)$ 上的子空间序列, 且 $W_m \oplus V_m = V_{m-1}$, 令 $\psi(t)$ 为对应尺度函数 $\varphi(t)$ 的小波函数, $\psi(t)$ 与 $\varphi(t)$ 正交, 则 $\{\psi_{m,n}(t)=2^{-m/2}\psi(2^{-m}t-n)\}_{n\in Z}$ 构成 W_m 的标准正交基, $L^2(R)$ 空间的正交分解为 $L^2(R)=\sum_{m=-\infty}^{\infty}\oplus W_m$.

定义 $L^2(R)$ 空间的内积 $\langle f(t), g(t)\rangle = \int_{-\infty}^{\infty} f(t)g^*(t)\,dt$. 令 $c_{j,n}=\langle f(t), \varphi_{j,n}(t)\rangle$ 和 $d_{j,n}=\langle f(t), \psi_{j,n}(t)\rangle$, $c_{j,n}$ 和 $d_{j,n}$ 分别被称为 $f(t)$ 在尺度为 j 时的尺度系数和小波系数, 它们分别是信号在尺度为 j 时的低频分量和高频分量, 对应了信号 $f(t)$ 在空间 V_j 和 W_j 中的投影. 通常在实际应用时是将 $\varphi(t)$ 和 $\psi(t)$ 与滤波器对应起来, 把 $\varphi(t)$ 看作一个低通滤波器 $H(\omega)$, 把 $\psi(t)$ 看作一个高通滤波器 $G(\omega)$. 令 $h(n)$ 和 $g(n)$ 分别为相应滤波器的冲激响应. 本文采用卷积型的 Mallat 快速算法[9]. 设 $D=S^d_{2^0}f[n]\,(n\in Z)$ 为原始信号 f 的离散采样序列, $W^d_{2^j}f[n]\,(n\in Z)$ 为 D 在尺度 j 上的小波系数, $S^d_{2^j}f[n]\,(n\in Z)$ 为 D 在尺度 j 上的尺度系数, $\{S^d_{2^J}f, (W^d_{2^j}f)_{1\leqslant j\leqslant J}\}$ 构成了信号 D 的二进小波分解. Mallat 快速分解算法如下:

$$S^d_{2^{j+}}f = S^d_{2^j}f * h_j,\quad W^d_{2^{j+}}f = S^d_{2^j}f * g_j,\quad j=0\sim J-1 \qquad (1)$$

其中, h_j 和 g_j 分别表示 h 和 g 中每相邻的两系数间插入 2^j-1 个零点构成的新滤波器的冲激响应, * 表示两个向量的卷积运算.

本文采用 Spline2 小波[10]对光谱进行四级小波分解, 得到光谱的四级小波系数. 图 1 所示为两条实测光谱及其四个尺度上的小波系数(从下到上尺度分别为 1、2、3、4). 可以看出, 第四级小波系数(也即尺度为 4)的能量主要分布在谱线上, 噪声的成份已经很小, 因此我们选择光谱的第四级小波系数为初始小波特征.

2.2 PCA 特征提取

PCA 是一种常用的特征提取方法. 对于高维数据, PCA 计算一组捕获了数据中最大方差的正交方向, 使得数据能以最小方差重构. 这些矢量称之为主分量矢量, 而样本在主分量矢量上的投影称之为样本的主分量. 利用主分量能够以较少的维数来描述数据, 同时又最大地保留了数据的结构. 从图 1 可以看出, 光谱的小波特征中存在较大的冗余信息, 因此我们通过 PCA 对其进行特征压缩.

为了使训练样本中包含尽可能多的谱线特征, 我们采用了 Kinney&Calzetti[11] 的 11 个静止光谱模板, 其中四个 NGs 和七个 AGs 模板, 对这些模板进行红移模拟, 红移区间取为 $0\leqslant z\leqslant 1.2$, 步长为 0.01, 截取波长为 380~742nm 的部分, 得到 1331 个光谱样本, 其中 484 个是 NGs, 847 个是 AGs. 对这些光谱按 0.5nm 的分辨率重新采样, 这样每个光谱的维数为 725 维. 由于模板中噪声较小, 我们对这 1331 个模拟光谱加入均方差为 0.1 的高斯白噪声. 文献[11]指出: 天体光谱的噪声理论上可以认为是泊松噪声, 可用高斯白噪声近似, 光谱的信噪比等于噪声标准差的倒数. 将加入噪声后的 1331 个光谱作为训练样本进行主分量分析, 方差贡献率取 95%.

设 $\{X_i, 1\leqslant i\leqslant N\}$ 为训练样本的小波特征集合, N 为训练样本个数, 这里 N 为 1331. 假设每个光谱上有 M 个采样点, 也即小波特征向量为 M 维. 主分量分析的具体步骤为: 首先,

将每个训练样本的小波特征向量 X_i 单位化, 由这些单位向量构成矩阵 $X_{N\times M}$; 然后, 求协方差矩阵 X^TX 的特征值和特征向量, 将特征值从大到小排序, 选取占特征值总和 95% 的前 K 个特征值对应的单位特征向量组成 PCA 特征空间变换矩阵 $H_{M\times K}$. 本文得到的 K 为 81. 将一条光谱的小波特征变换到 PCA 特征空间即可得到其主分量特征.

3　分类器设计

本文利用 Fisher 线性判别法来设计分类器. Fisher 线性判别法是要找到一个投影方向, 使两类样本在这个方向上的投影具有最佳的可分性. 这个投影方向称为 Fisher 矢量. 求解 Fisher 矢量的具体步骤参见文献[12].

我们根据上节训练样本的主分量特征来求取一个 Fisher 矢量, 并将这些训练样本的主分量特征投影到 Fisher 矢量上. 利用这些投影点确定一个合适的分界阈值 x_0, 对于待测的星系光谱, 只要将其识别特征投影到 Fisher 矢量上, 将其投影点与 x_0 相比较即可做出分类决策. 为方便分析, 我们将训练样本的主分量特征投影到二维平面上显示出来, 如图 2 所示, 横轴为 Fisher 矢量 $\boldsymbol{P}$, 纵轴为与其正交的一个矢量 $\boldsymbol{G}$, 图中加号表示 NG, 点表示 AG. 可以看出, 训练样本的识别特征在 Fisher 矢量上的投影能够较好地分开. 本文采用了一个线性支撑矢量机来确定阈值 x_0, 得到 $x_0 = -0.06$.

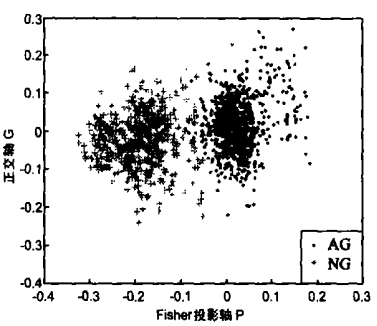

图 2　训练样本的识别特征在 Fisher矢量上的投影示意, 横轴为 Fisher矢量

下面为本文分类算法概述:

step1　根据公式(1)对待分类光谱进行小波分解, 选取第四级小波系数作为小波特征 F.

step2　将小波特征向量单位化后投影到 PCA 特征空间提取其识别特征 $F' = FH / \| F \|$.

step3　将 F' 投影到 Fisher 矢量 $\boldsymbol{P}$ 上, 得到投影点 $x = P^TF'$.

step4　如 $x \geqslant x_0$, 则该光谱属于活动星系光谱, 否则为正常星系光谱.

4　实验与分析

实验中采用了三组测试数据: 测试集 1 为模拟数据, 我们采用 Kinney&Calzetti 的 11 个静止模板进行红移模拟, 红移区间取为 $z \leqslant 1.2$, 步长为 0.001, 得到 13211 个光谱样本, 其中 4804 个是 NGs, 8407 个是 AGs; 测试集 2 为来自美国 SDSS 巡天中 0271 天区至 0275 天区中 1574 条正常星系的观测光谱; 测试集 3 为这五个天区中的 373 条活动星系的观测光谱.

支撑矢量机[13](SVM)是一种经常被研究的光谱分类方法[8,11], 我们基于不同的特征对 Fisher 方法和 SVM 进行了一系列比较. 实验中的训练数据均为 2.2 节提到的 1331 个模拟光谱. 实验结果如表 1 所示, 表中同时也给出了各种方法的适用场合(A: 流量未定标, B: 流量定标). 其中, SVM 的程序采用 SVM[light]程序[13], 核函数是高斯核. 实验中拟合连续谱的方法[7]为: 对原始光谱作四级小波分解, 将小波系数全部置零后, 再进行小波重构来逼近原始光谱的连续谱.

表 1　几种方法的比较

试验	特征提取	分类方法	测试集 1	测试集 2	测试集 3	适用情况
1	小波特征+ PCA	Fisher	98.6%	97.97%	100%	A/B
2	小波特征+ PCA	SVM (906 支撑)	67%	0.7%	99.38%	A/B
3	原始光谱+ PCA	Fisher	97.2%	99.6%	100%	B
4	原始光谱+ PCA	SVM (59 支撑)	99.8%	92.64%	96.5%	B
5	连续谱归一化 + PCA	SVM (997 支撑)	64.94%	0	100%	A/B
6	连续谱归一化	Fisher	90.3%	67.3%	43.7%	A/B

从表 1 中可以看出: ①使用同样的光谱特征, Fisher 方法明显优于 SVM 方法. ②基于连续谱归一化后的光谱特征的识别率(第五、六组)远远低于基于原始光谱特征的识别率(第三、四组), 这充分说明, 连续谱的特征在基于原始光谱特征的识别中起着很大的作用, 因此, 基于原始光谱特征的识别方法不适于流量未定标的情况. ③本文基于光谱小波特征的方法(第一组)对模拟数据和实测数据都有较高的识别率, 尽管其识别结果稍次于第三组的方法, 但本文方法同时适用于流量定标和不定标两种情况. ④同样是基于谱线特征的识别, 第一组基于光谱的小波特征优于第六组基于连续谱归一化后的光谱特征的识别, 这是因为本文选取的小波特征主要是谱线在尺度为 4 时的高频特征, 而第六组方法利用的是原始的谱线特征. 在第六组中未进行 PCA 特征提取是因为训练样本的 PCA 特征在 Fisher 矢量上的投影是不可分的.

为检验本文方法的鲁棒性, 我们对测试集 1 加了不同均方差的高斯白噪声, 在每个信噪比下进行了 100 次实验, 取其平均值, 得到识别率随信噪比的变化如图 3 所示, 可以看出, 该方法具有较好的鲁棒性, 在信噪比较低的情况下也能得到较好的识别率.

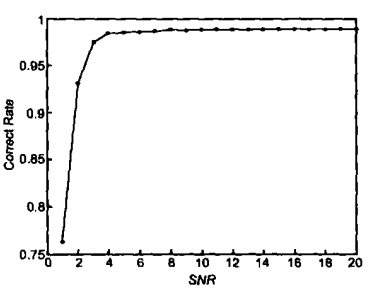

图 3　识别率随信噪比的变化

5　结论

光谱的自动处理对包含我国的 LAMOST 在内的一些大型光谱巡天计划有着非常重要的意义. 本文基于光谱的小波特征实现星系光谱的识别, 解决了流量未定标和红移值未知情况下的正常星系和活动星系光谱的分类. 通过实验对几种不同的光谱特征和分类方法进行了比较, 实验结果表明, 本文基于小波特征的分类方法具有较强的鲁棒性, 在流量未定标情

况下的识别效果优于其他几种分类方法.

参考文献:

[1] 中国科学院. LAMOST 项目计划建议书[R]. 1995.

[2] A J Connolly, A S Szalay, M A Bershady, et al. Spectral classification of galaxies: an orthogonal approach[J]. Astronomical Journal, 1995, 110(3), 1071-1082.

[3] Gaspar Galaz, Valerie de Lapparent. The ESO-sculptor survey: spectral classification of galaxies with[J]. Astronomy and Astrophysics, 1998, 332(2): 459-478.

[4] D Zaritsky, A I. Zabludoff, A W Jeffrey. Spectral Classification of galaxies along the Hubble sequence[J]. Astronomical Journal, 1995, 110(7): 1602-1614.

[5] D Qin, Z Hu, Y Zhao. New automated classification technique of galaxy spectra with Z< 1.2 based on PCA-ODP[A]. J L Starck, F D Murtagh, eds. Proceedings of SPIE(vol. 4847)[C]. Bellinghan: SPIE, 2002. 362-370.

[6] 罗阿理, 赵永恒. 使用小波技术自动搜寻天体谱线[J]. 天体物理学报. 2000, 20(4): 427-436.

[7] 赵瑞珍, 胡占义, 赵永恒. 谱线自动提取的小波变换零交叉点方法[J]. 光谱学与光谱分析. 2005, 25(1): 153-156.

[8] P Guo, F Xing, Y G Jiang. Stellar data classification using SVM with wavelet transformation[A]. Proceedings of the IEEE Conference on Systems, Man and Cybernetics[C]. CD-ROM, 2004. 5894-5899.

[9] Mallat S, Zhong S. Characterization of signals from multiscale edges[J]. IEEE Trans PAMI, 1992, 14(7): 710-732.

[10] K R Castleman. Digital Image Processing[M]. USA, Prentice Hall, 1996.

[11] 覃冬梅. 天体光谱信号的自动识别研究[D]. 北京: 中科院自动化所博士论文, 2003.

[12] 边肇祺, 张学工, 等. 模式识别[M]. 北京: 清华大学出版社, 2000.

[13] T Joachims. Making large-scale SVM learning practical[A]. Advances in Kernel Methods - Support Vector Learning[C]. B. Sch? lkopf and C. Burges and A. Smola (Eds.), Cambridge, MA, MIT-Press, 1999.

作者简介:

刘 蓉 女, 1972 年生于陕西西安, 分别于 1996 年和 1999 年获西北大学学士学位和硕士学位, 现为西安电子科技大学数学系讲师及在职博士生, 主要研究方向为应用统计、模式识别等.
E-mail: rliu@mail.xidian.edu.cn.

段福庆 男, 1973 年生于陕西渭南, 分别于 1995 年和 1998 年获西北大学学士学位和硕士学位, 现为中科院自动化所模式识别国家重点实验室博士研究生, 主要研究方向为模式识别、信号处理、计算机视觉等.