

基于谱线特征匹配的恒星光谱自动识别方法

刘中田¹, 邱宽民¹, 赵瑞珍²

1. 北京交通大学电子信息工程学院, 北京 100044

2. 北京交通大学计算机与信息技术学院, 北京 100044

摘要 我国正在实施的大型巡天项目(LAMOST 项目), 急需恒星光谱的自动识别系统。文章给出了一种基于谱线特征匹配的恒星光谱自动识别方法。该方法由以下主要步骤组成: (1) 利用小波变换的方法对观测光谱进行谱线特征提取; (2) 将提取出的特征和恒星谱线的特征模板进行相关匹配; (3) 根据相关匹配结果进行恒星光谱识别。通过对 Sloan Digital Sky Survey (SDSS), Data Release Four (DR4) 中的大量真实光谱数据实验表明, 该方法具有对噪声鲁棒等特点, 正确识别率高达 96.7%。该方法可对相对定标的巡天光谱进行自动识别, 符合 LAMOST 数据的要求, 可为天文学家进行恒星和银河系的结构等研究提供帮助。

关键词 恒星光谱识别; 谱线特征匹配; 小波变换; 特征提取

中图分类号: TN911.7 文献标识码: A 文章编号: 1000-0593(2008)06-1435-04

引言

高质量 CCD 探测器的应用, 扩大了人们的观测能力。我国正在建设的 LAMOST 项目, 预计每个观测夜能获得数万条光谱数据。对于如此大量的巡天观测数据人工无法完成光谱识别的任务, 自动的光谱识别方法将成为必然的选择。LAMOST 巡天得到的光谱复杂多样, 有各种类型的光谱, 研究恒星光谱的自动识别方法具有重要意义。作者已经进行了光谱中分子吸收带较强的晚型恒星识别研究^[1,2], 本文结合以前的工作, 目的是识别光谱中谱线较强的较早型恒星, 主要包括 O, B, A, F, G 和 K 型及其子次型恒星。综合这两方面的工作, 就可以构造出恒星光谱的自动识别系统。

目前, 恒星的识别方法主要是从恒星光谱型和光谱次型进行分析的。Kurtz 应用交叉相关认证和主分量分析方法(PCA), 将 MK 分类方法应用到低分辨率光谱(1.4 nm)。Gulati 等应用神经网络方法做恒星光谱次型的分类工作^[3]。覃冬梅等应用主分量分析方法结合最近邻分类方法, 对恒星光谱型和光谱次型进行自动识别^[4]。现有的方法主要是在已知天体为恒星的前提下, 利用连续谱的信息进行识别的。对于 LAMOST 来说, 观测光谱的连续谱流量仅相对定标, 以前方法很难应用于 LAMOST 项目中。

本文提出了一种基于谱线特征匹配的恒星光谱自动识别方法。该方法利用小波变换的方法提取谱线特征, 然后进行

谱线特征匹配, 以克服低信噪比和光谱流量仅相对定标对恒星光谱识别造成的不利影响。为了能够有效描述恒星光谱的谱线特征, 本文构造了一组包含多种恒星子型的谱线特征模板。本文方法的优点在于: (1) 特征提取方法抗噪声能力强, 简单易行; (2) 没有利用恒星光谱的连续谱信息, 适用于连续谱仅相对定标的巡天项目。大量真实光谱数据实验表明, 本文方法十分有效, 正确识别率高达 96.7%。

1 恒星谱线的特征提取

光谱信噪比较低时谱线很难准确提取, 本文利用小波变换的方法提取谱线特征^[5,6], 以避开较难的低信噪比下的谱线提取问题。在谱线特征提取过程中, 本文首先对观测光谱进行 3 层小波变换, 对第 1, 2, 3 层半高频小波系数分别进行分层反变换并且归一化^[7]得到结果如图 1 所示。

恒星谱线以窄线为主, 但是也有一些恒星(如 B 型白矮星)的谱线相对较宽。如图 1(a), 对于谱线相对较窄的恒星光谱, 提取出的第 1, 2, 3 层谱线小波特征都较理想。但是, 对于谱线较宽的恒星光谱(如图 1(b), 谱线较宽的发射线恒星光谱), 由于受噪声影响较大, 较宽的谱线上叠加了噪声, 第一层谱线小波特征易受噪声影响(如图 1(b)右上)。所以, 为了保证谱线特征能够充分提取, 以及克服噪声的不利影响, 我们利用如图 2 方法提取恒星谱线特征。

收稿日期: 2006-11-18, 修订日期: 2007-02-15

基金项目: 国家“863”高技术研究发展计划(2003AA133060)和北京交通大学科技基金项目(2006RC028)资助

作者简介: 刘中田, 1979 年生, 北京交通大学电子信息工程学院讲师, e-mail: liuzht@bjtu.edu.cn

©1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

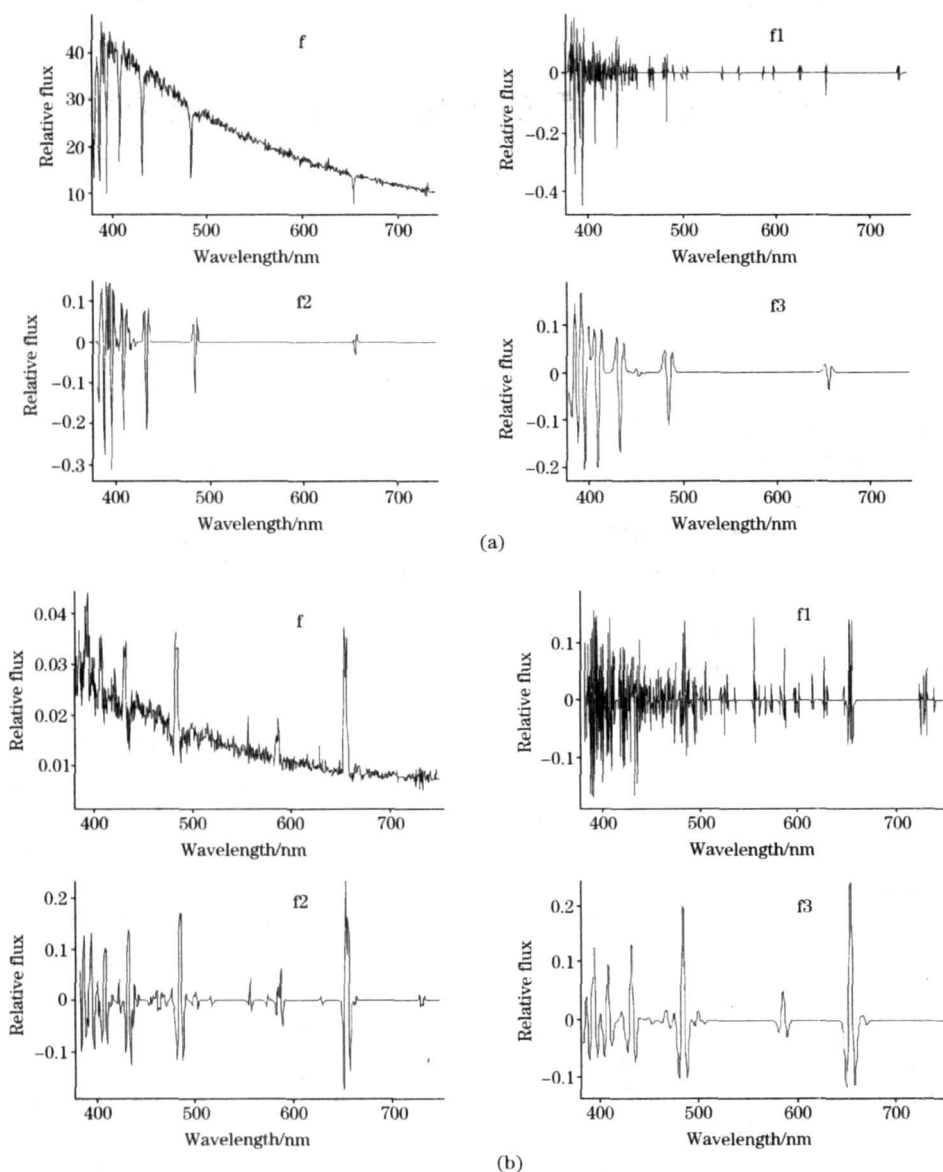


Fig 1 The results of extracting the features of spectral lines

(a): A-type star; (b): Emission line star

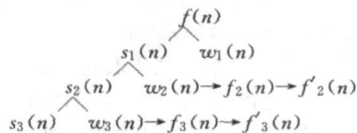


Fig. 2 Extracting the features of spectral lines

其中, $f(n)$ 是原始光谱信号, $s_i(n)$ 和 $w_i(n)$ 分别是第 i 层小波变换后的半低频和半高频小波系数, $f_i(n)$ 是对 $w_i(n)$, $i = 2, 3$ 分别进行小波反变换后的结果, $f'_i(n)$, $i = 2, 3$ 是求得的第 2, 3 层小波特征。则恒星谱线特征 $f'(i)$, $i = 1, 2, \dots, 2n$ 计算如下:

$$f'(i) = \begin{cases} f'_2(i) & i = 1, 2, \dots, n \\ f'_3(i-n) & i = n+1, n+2, \dots, 2n \end{cases} \quad (1)$$

2 谱线特征模板构造及相关匹配值计算

恒星光谱数据选取 Jacoby^[8] 的恒星数据集 161 条和 SDSS DR1 数据释放的恒星模板数据 20 条。利用主分量分析方法进行降维, 方差贡献率设定为 0.95, 最后得到 2 维 PCA 降维数据, 并可用二维图形表示^[4]。本文从 O 型到 K 型恒星数据中选取 24 条标准恒星数据用于构造恒星谱线特征模板, 主要是由 O5, B5, A5, F5, G5 和 K5 型恒星数据, 以及各类边界处的恒星数据构成。设这些数据为 $S_1, S_2, \dots, S_{24}$, 点数为 n 。利用恒星谱线的特征提取方法进行计算, 分别得到 $S'_1, S'_2, \dots, S'_{24}$, 点数为 $2n$ 。将这些特征按行排列, 即可构成一组恒星谱线的特征模板 $M_{24 \times 2n}$ 。

如果计算出观测光谱 $f(n)$ 的谱线特征 $f'(i)$, $i = 1, 2$

..., $2n$, 则匹配值 r_j 由下式计算:

$$r_j = \sum_{i=1}^{2n} f'(i) \times M(j, i) \quad j = 1, 2, \dots, 24 \quad (2)$$

由于有 24 个恒星谱线特征模板, 所以得到了 24 个模板相关匹配值。分别求得这 24 个匹配值中的最大值 η_m 和最小值 η_n , 按照如下规则判别恒星: 如果 $|\eta_m| > |\eta_n|$ 且 $|\eta_m| > \sigma$, 则判别该天体为恒星。其中, σ 是待定的阈值。这里, 如果 $|\eta_m| < |\eta_n|$, 则说明 $\eta_n < 0$, 观测光谱 $f(n)$ 中的相应恒星谱线位置具有发射线特征, 在此基础上, 如果对 $|\eta_n|$ 设置一个阈值且辅助相应恒星谱线的检测, 就能够有效地识别发射线恒星。因此, 本方法可以进一步发展成一个相对独立的系统, 该系统能够有效处理特殊光谱恒星。

3 恒星光谱的自动识别

恒星光谱识别器(单阈值)结构如图 3 所示。

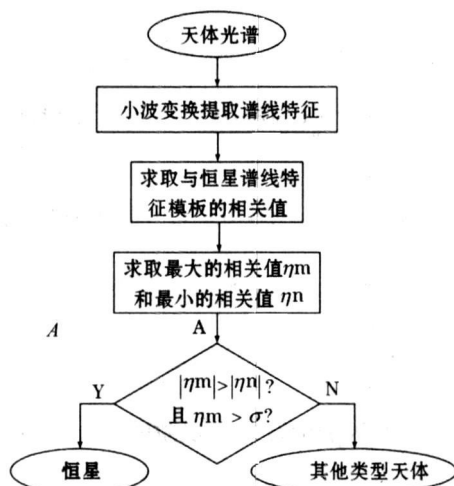


Fig 3 The process of recognizing the stars

具体算法由下述几个步骤构成: (1) 首先利用特征提取方法构建一组恒星谱线的特征模板; (2) 对观测光谱进行特征提取, 然后用该特征和恒星谱线的特征模板做相关匹配计算; (3) 求得最大的匹配值 η_m 和最小的匹配值 η_n ; (4) 设定阈值 σ , 如果 $|\eta_m| > |\eta_n|$ 且 $|\eta_m| > \sigma$, 则判别该天体为恒星。

本文结合基于最近邻规则的恒星分类器研究恒星的识别问题, 将单阈值恒星识别器的流程图做一定程度的改变, 如图 4 所示。在图 4 中, 流程图入口 B 点前的框图和图 3 中 A 点前的部分相同, 这里进行了省略。 η_j 是恒星最近邻分类器得出的最近距离值, β 是一个阈值。 σ_1 和 σ_2 是对 η_m 设置的两个阈值, 通过实验发现, 设定两个阈值比只设置一个阈值的实验正确率高。

4 实验及结论

4.1 实验数据选择

实验数据选择 SDSS DR4 中的恒星数据 10 000 条, 类星体数据 10 000 条和星系数据 10 000 条, 包括正常星系和活动

星系)。随机均分为两个数据集: 数据集 1 和数据集 2。

4.2 单阈值 σ 的确定

利用数据集 1 进行计算, 不同阈值 σ 对应的正确率曲线图见图 5。

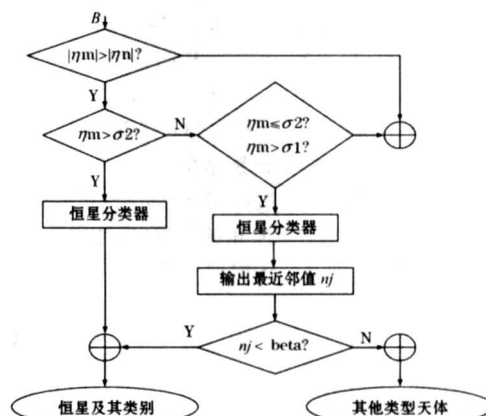


Fig. 4 The process of stellar recognition and classification

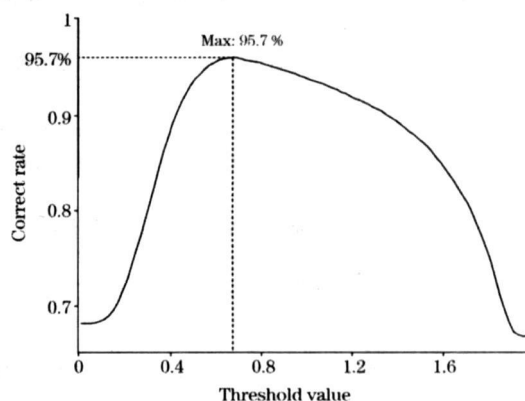


Fig. 5 The curve of the correct rate with respect to the threshold σ

如果设定一个阈值 σ , 则当 $\sigma = 0.66$ 时, 实验正确率达到最大值 95.7%。利用数据集 2 进行测试, 正确率为 95.6%。

4.3 双阈值 σ_1 和 σ_2 的确定

选取数据集 1 进行实验, 正确率曲线如图 6 所示。

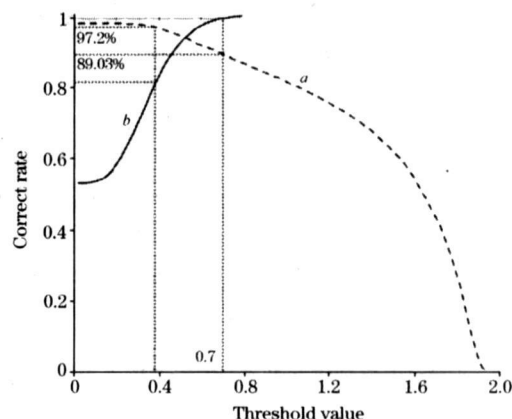


Fig 6 The curve of the correct rate

a: Other; b: Star

由图 6 可知, 曲线 a 表示不同阈值 σ 下恒星光谱单独统计的正确识别率; 曲线 b 表示其他类型光谱在不同阈值 σ 下的实验正确率。当 $\sigma_2 = 0.7$ 时, 其他类型光谱的错误识别率小于 1%。可见, 当选择 $\sigma_2 = 0.7$ 时, 识别出的天体基本上就是恒星, 所以我们选择阈值 $\sigma_2 = 0.7$ 。当 $\sigma_1 = 0.4$ 时, 恒星的正确识别率能达到 97.2%, 此时, 其他类型光谱的错误识别率较高。我们可以选择 $\sigma_1 = 0.4$, 在大于 σ_1 且小于 σ_2 的阈值区间, 结合恒星分类器进一步提高恒星的识别率。

4.4 阈值 n_j 的确定和恒星识别结果

实验中, 我们还需要对最近邻距离 n_j 设定一个阈值。实验发现, 当 $n_j = 0.02$ 时, 实验正确率最高, 达到 96.8%。用数据集 2 进行测试, 恒星总体识别正确率为 96.7%。

4.5 结论

通过实验我们发现, 对于谱线较强的恒星光谱, 利用谱线特征匹配的方法进行识别, 实验正确率较高。该方法具有以下特点, (1) 没有利用恒星光谱的连续谱信息, 比较适用于连续谱仅相对定标的巡天项目中; (2) 由于识别方法比较直观, 可以综合前人提出的恒星细分类方法, 构造性能较强的恒星识别与分类器, 提高光谱识别的正确率。另外, 本文的谱线特征提取方法兼顾了恒星发射线的特点, 使得相关匹配值具有恒星发射线的鉴别功能, 如果设置一个阈值且辅助相应恒星发射线的检测, 就能够有效地识别发射线恒星。综上所述, 本文方法不仅识别正确率较高, 而且具有较好的拓展性能。

参 考 文 献

- [1] LIU Zhong-tian, ZHAO Rui-zhen, ZHAO Yong-heng, et al(刘中田, 赵瑞珍, 赵永恒, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2005, 25(7): 1158.
- [2] LIU Zhong-tian, LI Xiang-ru, WU Fu-chao, et al(刘中田, 李乡儒, 吴福朝, 等). Acta Electronica Sinica(电子学报), 2007, 35(1): 67.
- [3] Gulati R, Gupta R A, Gothoskar P, et al. Astrophysical Journal, 1994, 426: 340.
- [4] 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), 2003, 23(1): 182.
- [5] Mallat S. IEEE Trans. Pattern Anal. Machine Intell., 1989, 11(7): 674.
- [6] LIU Zhong-tian, WU Fu-chao, LUO A-li, et al(刘中田, 吴福朝, 罗阿里, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2006, 26(2): 372.
- [7] LIU Zhong-tian, LI Xiang-ru, WU Fu-chao, et al(刘中田, 李乡儒, 吴福朝, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2006, 26(9): 1738.
- [8] Jacoby G H, Hunter D A, Christian C A. Astrophysical Journal, Supplement, 1984, 56: 257.

A Method for Automatic Recognition of Stellar Spectra Based on Feature Matching of Spectral Lines

LIU Zhong-tian¹, QIU Kuan-min¹, ZHAO Rui-zhen²

1. School of Electronics and Information Engineering, Beijing Jiaotong University, Beijing 100044, China

2. School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China

Abstract The LAMOST project, the world's largest sky survey project being implemented in China, urgently needs an automatic stars recognition system. The present paper presents a method for automatic recognition of stellar spectra based on feature matching of spectral lines. This method consists of three main steps: First, the features of spectral lines in the observed spectra are extracted using the wavelet transform. Then, the correlations between the extracted features and the feature templates of the stellar spectral lines are computed. Finally, based on the results of the former step, the stellar spectra can be recognized. The extensive experiments with real observed spectra from the SDSS DR4 show that the method can robustly recognize stellar spectra, and the correct rate of this method is as high as 96.7%. This method is designed to automatically recognize stellar spectra with relative flux and low signal-to-noise ratio, which is applicable to the LAMOST data and helps in the structure study of stars and galaxy etc.

Keywords Recognition of stellar spectra; Feature matching of spectral lines; Wavelet transform; Feature extraction

(Received Nov. 18, 2006; accepted Feb. 15, 2007)