基于知识的红移测量和谱线证认方法

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 西安 710071)
[②](中国科学院自动化所模式识别国家重点实验室 北京 100080)

摘 要：该文给出了一种基于知识的天体光谱的红移测量和谱线证认方法。首先，利用特征谱线的相关知识对红移候选和特征谱线候选进行了定义，并根据定义交叉确认红移候选和特征谱线候选；然后，利用Parzen窗法对所得到的红移候选集进行密度估计；最后，确定密度最大的红移候选，将落入其Parzen窗内的所有红移候选值进行平均得到红移，与这些红移候选值相对应的特征谱线候选即为特征谱线。与现有的基于谱线匹配的方法相比，该方法对谱线提取效果的依赖程度较低。实验结果表明：该方法的鲁棒性较好，正确率较其它基于谱线匹配的方法有较大提高。

关键词：光谱分析，红移测量，谱线证认，密度估计
中图分类号：TP391.4 文献标识码：A 文章编号：1009-5896(2006)01-0076-04

Red Shift Determination and Spectral Line Identification Based on Knowledge

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

Abstract A novel method for redshift determination and spectral line identification of celestial spectra is presented, which is based on the knowledge of feature spectral lines. Firstly, definition of redshift candidate and feature spectral line candidate is given, and the candidates are cross-validated according to the definition; Secondly, the density is estimated at every redshift candidate by using the Parzen window technique; Finally, the average of redshift candidates in Parzen window of the redshift candidate with maximum density is the redshift, and the feature spectral line candidates corresponding to those redshift candidates are feature spectral lines. Compared with other methods of the same kind, this method has a lower dependence on the quality of spectral line extraction. Experiments show that this method is robust and the correct rate is encouraging.

Key words Spectral analysis; Redshift determination; Spectral line identification; Density estimate

1 引言

大规模星系光谱巡天计划(如 SDSS, 2dF, LAMOST 等)获取的天体光谱数以亿计，因而自动的光谱处理方法对天文学界有着重要的意义。天体光谱主要由连续谱、谱线和噪声组成。连续谱是不聚集在任何特定波长处的连续辐射产生的光谱。谱线分为吸收谱线和发射谱线，分别是由于天体中的原子、分子在发生能级跃迁时吸收或辐射特定波长处的能量所体现出的特征， 谱线线心对应的波长称之为谱线的特征波长。图1上端为两条观测光谱，其中，横轴为波长，纵轴为流量，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，图中已标出特征谱线。由于天体背离地球运动，因而我们观测到的谱线的波长要比其静止波长大，即谱线向红端移动，这就是所谓的红移现象。红移值是河外天体最重要的物理参量。设谱线的静止波长为 λ'，观测波长为 λ，z 是红移值，则

$$\lambda = (1+z)\lambda' \tag{1}$$

2004-12-15 收到，2005-03-14改回
国家863计划(2003AA133060)和国家九五重大科学工程LAMOST项目资助课题

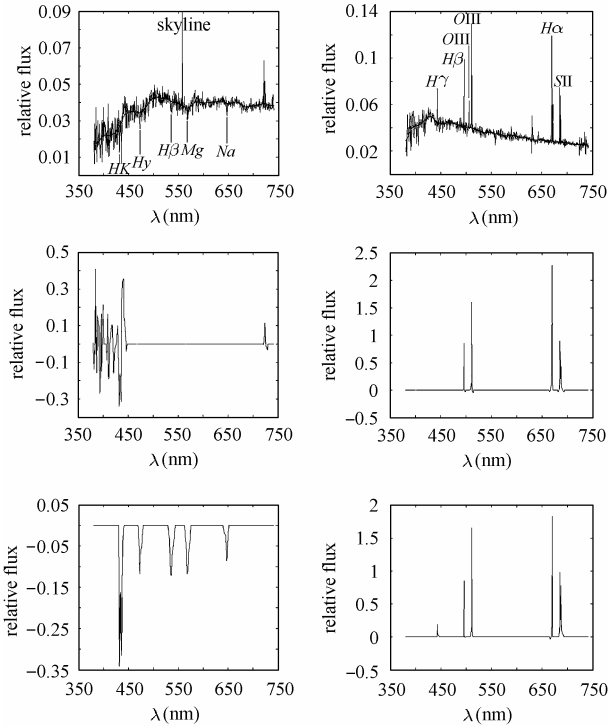

图1　谱线提取及证认 (上，原始光谱和连续谱；中，谱线提取结果；下，谱线证认结果)

对于河外天体，红移测量和谱线证认是天体光谱分析的首要任务。

从式(1)可以看出：红移测量和谱线证认是密切相关的，两者中任何一方的解决都会使另一个问题迎刃而解。已有的自动方法可以分为两类：一类是基于交叉相关的方法，另一类是基于谱线匹配的方法。第一类中最经典的方法是Tonry和Davis的交叉相关法[1]，其核心是用实测光谱和一系列静止模板光谱做交叉相关，在所有交叉相关函数中寻找最大峰，这个峰的位置和宽度分别决定了红移值和误差。当目标光谱和模板光谱比较相似时，这种方法能给出精确的结果。但实际上观测光谱和模板光谱之间往往有一定的差别，尤其是发射线光谱。Glazebrook[2]等人提出的PCAZ方法将这种方法进行了推广。PCAZ是用PCA构造一组正交模板，然后用正交模板的线性组合与观测光谱做交叉相关，方法中认为正交模板的线性组合可以补偿目标光谱与模板光谱之间的不匹配。由于可处理波长段的限制，这类方法只适合小红移的光谱。第二类中较典型的有密度估计法[3]，Hough变换法[4]。这类方法的核心思想是根据式(1)寻找谱线提取结果中与静止模板的特征谱线耦合最多的那些谱线。这类方法严重地依赖谱线提取的效果，当谱线提取的结果中特征谱线较少而伪特征谱线较多时，通常很难得到正确的结果。为了降低谱线提取的影响，本文给出了一种基于知识的红移测量和谱线证认方法，利用静止模板中特征谱线的相关知识交叉确认红移候选和特征谱线候选，能够将谱线提取过程中丢失的一些特征谱线找出来，并且可以降低伪特征谱线的影响。

本文组织结构如下：第 2 节简要地介绍了谱线的自动提取过程；第 3 节是红移和谱线的证认，是本文的重点；第 4 节给出实验结果；最后是本文的结论。

2　谱线提取

由于谱线提取对本文方法的影响较小，因此可以采用现有的任一种谱线提取方法[5, 6]。为保持文章的完整性，本节将简要地介绍本文的谱线提取过程。

2.1 光谱的预处理

预处理包括光谱去噪和连续谱归一化。

天体光谱噪声主要表现为两种形式：一种是在固定波长处的宇宙线，为强脉冲噪声；另一种是随机白噪声。本文先在可能存在宇宙线的固定波长附近进行中值滤波，去除宇宙线噪声，然后采用小波软阈值法[7]去除随机白噪声。

连续谱归一化就是用光谱除以连续谱，其目的是去除连续谱的影响。归一化后的光谱称之为谱线光谱。本文采用小波变换的方法[5]拟合连续谱。为便于后续的谱线搜索，我们将谱线光谱减一，这样发射线的线心强度为正，而吸收线的线心强度为负。

2.2　谱线特征提取

(1)　阈值处理　计算整个谱线光谱的均方根RMS，设置阈值为 2RMS 。对谱线光谱上强度绝对值小于阈值的点，将其强度置为零。

(2) 搜索谱线线心。在阈值处理后的谱线光谱上搜索局部极大值点，我们将其视为谱线的线心。设 $f(\lambda_{i-1})$，$f(\lambda_i)$，$f(\lambda_{i+1})$ 为相邻三个点处的强度，如满足 $|f(\lambda_{i-1})|<|f(\lambda_i)|$ 且 $|f(\lambda_i)|>|f(\lambda_{i+1})|$ ，则称 $f(\lambda_i)$ 为局部极大值。

(3) 搜索谱线边界。在阈值处理前的谱线光谱上，以谱线线心为中心向两端搜索谱线的边界。我们用 $LA=\{L_i=(\lambda_i,t_i,\ \mathrm{width}_i),\ i=1,\cdots,N\}$ 来表示提取出的谱线信息。其中，L_i 表示第i个谱线，λ_i 表示谱线的特征波长，t_i 表示谱线的线型(1表示发射型谱线，-1表示吸收型)，width_i 表示谱线宽度。图1中间部分为谱线提取过程得到的谱线。

3　红移和谱线的证认

3.1 特征谱线的知识获取

每一类天体光谱一般都会有其特定的特征谱线，这些谱线的静止特征波长是已知的。谱线证认的目的就是要对目标光谱进行特征谱线的识别。为了获取特征谱线的相关知识，我们首先对静止模板光谱进行预处理，然后在其谱线光谱上搜索该类天体光谱的特征谱线。本文用 $\mathrm{LT}=\{L_i'=(\lambda_i',t_i',$

$\text{width}_i'),\ i=1,\cdots,M\}$ 来表示静止模板光谱中特征谱线的谱线信息。

3.2 红移候选和特征谱线候选

假设 $L=(\lambda,t,\text{width})$ 为目标光谱中的一个谱线，$L'=(\lambda',t',\text{width}')$ 为静止模板光谱的一个特征谱线，计算

$$zc=\frac{t\lambda}{t'\lambda'}-1,\quad rc=\frac{\text{width}}{\text{width}'} \tag{2}$$

定义1 称 zc 为红移候选、$L=(\lambda,t,\text{width})$ 为特征谱线候选，如果满足下述条件：

(1) $zc\ge 0$；

(2) $rc>C(1+zc),\ C\in(0,1)$。

红移候选和特征谱线候选必须满足第1个条件的原因是：$zc<0$ 的情况有两种：一种是从目标光谱中提取出的谱线 L 的线型和静止模板中特征谱线 L' 的线型不一致；另一种是两者的谱线线型相同但特征波长明显不匹配(红移不为负值)。红移候选和特征谱线候选必须满足第2个条件的原因是：如果 zc 是真正的红移值，说明谱线 L 和特征谱线 L' 是同一种特征谱线，则理论上应该有 $rc=1+zc$。但是，由于受噪声和光谱分辨率的影响，这个关系是很难成立的。本文选 $C=0.8$。

根据定义1，我们给出确定红移候选和特征谱线候选的交叉确认算法如下。

步骤1 利用提取出的目标光谱的谱线集 LA 和静止模板中的特征谱线集 LT，根据式(2)得到红移候选集 $Z=\{z_i,\ i=1,2,\cdots\}$ 和特征谱线候选集 $\text{LC}=\{L_i=(\lambda_i,t_i,\ \text{width}_i),\ i=1,2,\cdots\}$。

步骤2 对每一个红移候选 $z_k\in Z$ 和静止模板中的每一个特征谱线 $L_l'\in LT$，根据式(1)得到波长 $\lambda=(1+z_k)\lambda_l'$，如谱线候选集 LC 中已经存在该波长处的谱线，则不用做任何处理；否则，在目标光谱的谱线光谱上波长为 λ 附近搜索谱线，记得到的谱线为 $L_0=(\lambda_0,t_0,\text{width}_0)$，按式(2)计算 $z=\frac{t_0\lambda_0}{t_l'\lambda_l'}-1,\quad r=\frac{\text{width}_0}{\text{width}_l'}$，如果 z 满足红移候选条件，则将谱线 L_0 加入特征谱线候选集 LC，将 z 加入红移候选集 Z。

3.3 利用密度估计确定红移和特征谱线

很显然，在特征谱线候选集中，只有特征谱线才会与静止模板中对应的同一种特征谱线存在红移耦合关系式(1)。因此在红移候选集中，红移值附近的点一般来说是最密集的。图1中左侧光谱的红移候选值分布如图2所示，其中，横轴为序号，纵轴为红移候选值，线段所示即为红移附近的点。因此，只要找到红移候选集 Z 中密度最大的点并对其附近的点进行平均即可得到较为准确的红移值。

本文采用Parzen窗法[8]进行密度估计。定义窗函数 $\phi(u)=\begin{cases}1, & |u|\le\varepsilon\\ 0, & \text{其它}\end{cases}$，则密度估计函数为 $\hat f(z)=\frac{1}{P}\sum_{j=1}^{P}\phi(z-z_j)$，其中 $z_j\in Z$，P 为红移候选的个数，ε 为设定的红移误差上界，本文选 ε 为 0.001。令 $\hat f(z_k)=\max\{\hat f(z_j),\ j=1,2,\ldots,P\}$，$RC=\{z_{k,i}\mid z_{k,i}\in Z,\ \phi(z_k-z_{k,i})=1,\ i=1,2,\cdots K\}$ 为红移候选集 Z 中所有落入 z_k 的 Parzen 窗内的红移候选值，则红移值为 $z=\sum_{i=1}^{K}z_{k,i}\Big/K$，证认出的特征谱线为 $\{(L_{k,i},L_{k,i}')\mid \lambda_{k,i}=(1+z_{k,i})\lambda_{k,i}',\ z_{k,i}\in RC,\ i=1,2,\cdots,K\}$，其中 $L_{k,i}\in LC,\ L_{k,i}'\in LT$。图1下端为证认出的特征谱线。

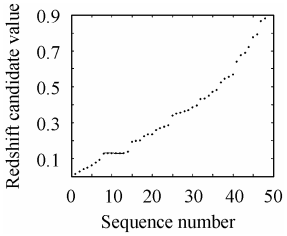

图2 红移候选分布

3.4 讨论

在确定特征谱线候选和红移候选时，我们仅采用了谱线的特征波长、线型和宽度信息作为判据，而没有对谱线的强度信息进行约束，这是由于发射谱线的强度在不同光谱间的变化较大。图1右侧光谱的谱线证认结果中丢失了 506nm 处的特征谱线 OIII，这说明观测光谱中的特征谱线也有可能不满足谱线候选定义中的谱线宽度约束。对于由此丢失的特征谱线，我们可以在红移确定以后，在光谱上直接搜索谱线，无需对其进行谱线宽度约束。

4 实验与分析

在本节中我们给出了两组实验，第1组为模拟光谱实验，第2组是 SDSS 实测光谱的实验。在第2组实验中，将本文方法与密度估计法和 Hough 变换法进行了比较。

4.1 模板光谱

实验中所采用的正常星系的模板光谱是通过PCA将Kinney&Calzetti[9]的 4 个正常星系(NG)的静止模板合成的一个模板，如图3(a)；采用的活动星系(AG)的模板光谱是通过PCA将Kinney&Calzetti的 7 个活动星系的静止模板合成的一个模板，如图3(b)。

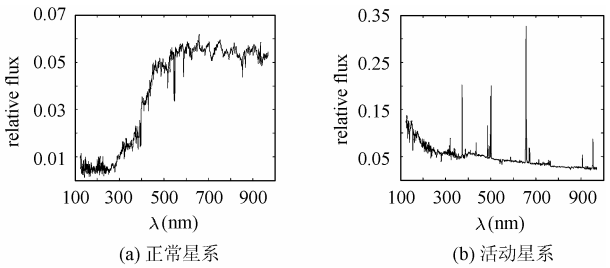

图3 模板光谱

4.2 模拟实验

按红移区间 $z \le 0.5$,步长为 0.01 分别对 Kinney&Calzetti 的 11 个静止模板光谱进行红移模拟，截取波长段为 380nm—742nm 的部分,得到两组模拟光谱。第 1 组包括 204 个正常星系光谱，第 2 组包括 357 个活动星系光谱。我们对这些模拟光谱加了不同均方差 σ 的高斯白噪声，在每个信噪比下($SNR = 1/\sigma$)进行了 50 次实验，对实验结果取平均，得到正确率随信噪比的变化如图 4 所示。从图中可以看出：当信噪比大于 10 时，两组数据的正确率都达到了 90%以上，并且正确率随着信噪比的提高而逐渐提高，这说明该方法的鲁棒性较好；在信噪比低于 12 时，活动星系的正确率大于正常星系，这是因为吸收谱线的强度一般较发射谱线弱，更容易受到噪声的污染，因而在低信噪比时吸收谱线的匹配效果较发射谱线差；当信噪比大于 12 时。正常星系的正确率大于活动星系，这是因为吸收谱线在不同光谱中的表现较发射谱线更为稳定。

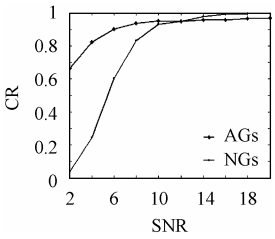

图 4　正确率-信噪比曲线

4.3 实测光谱实验

采用美国 SDSS 巡天观测数据中 0271-0275 天区的 1795 个星系光谱，包括 1574 个正常星系和 221 个活动星系光谱，这些光谱的信噪比平均为 13，红移值已经给出。我们分别采用本文方法和文献[3]中密度估计方法计算红移，其中的谱线提取方法均为本文第 2 节所介绍的方法，计算结果如表 1 所示。可以看出，本文方法的正确率较另外两种方法有较大的提高。这是因为密度估计法和 Hough 变换法仅利用了经过谱线提取过程得到的谱线，其结果直接受谱线提取效果的影响；而本文方法在红移候选和特征谱线候选的交叉确认过程中能够寻找谱线提取过程中遗失的特征谱线，这就增加了确定红移的证据，同时也降低了谱线提取效果对红移测量的影响。

表 1　实验结果对比

方 法	正常星系 error	活动星系 error	平均正确率
本文方法	108	6	93.65%
密度估计法	191	13	88.64%
Hough 变换法	196	12	88.41%

5　结束语

光谱的自动分析技术在大规模星系光谱巡天中有着非常重要的意义。本文给出的基于知识的红移测量和谱线证认方法是一种基于谱线匹配的方法，与已有的谱线匹配方法不同的是：本文方法利用特征谱线的相关知识对红移候选和特征谱线候选进行交叉确认，能够寻找到在谱线提取过程中遗失的特征谱线，从而降低了谱线提取效果对红移测量的影响。当目标光谱中存在大多数特征谱线不满足特征谱线候选定义中的谱线宽度约束时，本文方法得到的红移精度较差，这是本文方法的一个缺点。然而，这种情况在实际的观测光谱中仅是少数，实验结果证实了这一点。与现有的红移测量方法一样，本文的目标光谱类别是已知的，因而错误的光谱分类结果将导致错误的红移测量和谱线证认。

参 考 文 献

[1] Tonry J and Davis M. A survey of galaxy redshifts. I. data reduction techniques. *The Astronomical Journal*, 1979, 84(10): 1511 – 1525.

[2] Glazebrook K, Offer A R and Deeley K. Automatic redshift determination by use of principal component analysis. I. Fundamentals. *The Astrophysical journal*, 1998, 492(1): 98 – 109.

[3] 段福庆, 吴福朝, 罗阿理等. 用于红移测量的基于密度估计的模板匹配法. 光谱学与光谱分析,2005, 25(11):1895 – 1898.

[4] 周虹,黄凌云,罗曼丽. 一种基于 Hough 变换和神经网络的分层类星体识别方法. 电子科学学刊,2000, 22(7)：529 – 535.

[5] 赵瑞珍. 天体光谱的噪声去除与特征提取.中国科学院自动化研究所博士后出站报告, 2003.

[6] 罗阿理,赵永恒. 使用小波技术自动搜索谱线. 天体物理学报.2000, 20(4):427 – 436.

[7] Donoho D L. De-noising by soft-thresholding. *IEEE Trans.on IT*, 1995.5, 41(3): 613 – 627.

[8] 边肇祺，张学工等. 模式识别. 北京清华大学出版社， 2000, 65:70

[9] Kinney A L, Calzetti D, Bohlin R C, et al. Template ultraviolet to near-infrared spectra of star-forming galaxies and their application to k-corrections. *Astrophysical Journal*, 1996, 467(8): 38 – 60

刘　蓉：　女，1972 年生，讲师，研究方向为应用统计，模式识别等.

段福庆：　男，1973 年生，博士生，研究方向为模式识别，信号处理等.

刘三阳：　男，1959 年生，教授，研究方向为优化理论，应用统计等.

吴福朝：　男，1957 年生，研究员，研究方向为计算机视觉，模式识别等.