巡天光谱分类前的预处理——流量标准化

李乡儒[1,2], 刘中田[1,2], 胡占义[1*], 吴福朝[1], 赵永恒[2]

1. 中国科学院自动化所模式识别实验室, 北京 100080
2. 中国科学院国家天文台, 北京 100012

摘 要 由于噪声、畸变、观测环境和观测设备、以及流量未定标等因素的影响, 对天体光谱进行自动识别之前, 需要对它进行相应的标准化/预处理。文章研究了对巡天光谱自动分类前的预处理——光谱流量的标准化问题。通过分析光谱流量的干扰因素及其特点, 提出了流量数量级变化的基本模型, 并给出了相应的流量标准化方法。通过对正常星系和类星体的分类实验, 表明文章所给基本模型的正确性, 以及所给流量标准化方法良好的性能。并且从理论上分析、比较、解释了上述方法在性能上的差异。特别需要指出的是, 研究表明文献中通常采用的流量标准化方法的效果较差。该研究结果对于大型光谱巡天所产生的海量数据的其它自动处理研究(例如, 红移测量, 天体表面有效温度, 和化学丰度估计等)也具有重要的指导意义。

关键词 光谱分类; 流量模型; 流量标准化; 预处理; 类星体; 主成分分析
中图分类号: TN911.7 文献标识码: A 文章编号: 1000-0593(2007)07-1448-04

引 言

随着天文观测技术的进步、数据获取能力的提高和大型光谱巡天计划的相继实施(如: SDSS[1]计划, 2dF计划和LAMOST项目等), 光谱数据的自动处理越来越受到重视和关注, 例如, Connoliy[2]和Galaz[3]分别采用主分量分析对红移值已知的星系光谱进行自动分类, 覃冬梅[4]采用主成分分析和最近邻方法研究了恒星的光谱型自动分类; 最近, 刘中田[5]、李乡儒[6]等研究了未知红移天体的自动分类, 许馨[7]等研究了天体红移的自动测量, 张健楠[8-10]等研究了恒星大气基本物理参量(表面有效温度、表面重力加速度、化学丰度)的自动估计。

由于噪声和畸变的影响, 对天体光谱进行自动识别之前, 需要对它进行相应的预处理, 以抑制与自动处理或分析任务无关的信息, 例如, 去噪、校正、流量标准化, 连续谱归一化、剔除离群点、去掉天光线, 以及特征提取等。噪声和畸变可能有多种原因: 光学透镜的残次、光电传感器的非线性、不当的焦距、以及大气的扰动(不均衡、不稳定)等等。通过特征提取和特征选择能够去除或减少与当前的分类任务无关的冗余信息, 这不仅能减少问题的复杂性, 提高处理效率, 还能极大地简化分类器的设计[11]。预处理的目的是改善光谱数据, 抑制不需要的变形或增强某些对于后续自动处理重要的光谱特征。因此, 预处理在很多情况下是非常必要的。

本文研究了巡天光谱进行自动分类前的预处理——光谱流量标准化问题。由于天气状况、月光亮度、星等、积分时间等因素的影响, 连续谱会发生畸变, 导致不同观测的光谱流量数量级会有差异, 这增加了自动分类的难度、复杂性和准确性。目前, 在巡天光谱自动分类研究中, 常用的标准化方法是流量归一化[12](在1.2节将做具体介绍)。但是本文研究表明, 该标准化方法的效果较差。

下一节, 我们通过分析光谱数据的特点提出了流量标准化方法的基本思想, 并给出了五种不同的流量标准化方法; 在第二节, 介绍了本文中使用的实验方法和实验数据, 并通过实验研究了光谱流量标准化的必要性和不同标准化方法的性能。最后, 我们结合实验结果从理论上分析了所给标准化方法的特点、抗噪能力方面的差异。

1 基本思想与方法

1.1 天体光谱的特点分析

天体光谱是天体辐射的一种描述。绝大多数天体都有辐射, 其表现形式有连续辐射和谱线辐射之分。连续辐射指的

收稿日期: 2005-12-06, 修订日期: 2006-03-16
基金项目: 国家"863"项目计划(2003AA133060)和国家自然科学基金项目(60202013)资助
作者简介: 李乡儒, 1972年生, 中国科学院自动化研究所国家模式识别实验室在读博士研究生
*通讯联系人 e-mail: huzy@nlpr.ia.ac.cn

是不聚集在任何特定波长处的、低阶而连续的辐射[12,13]。天体中的原子、分子等在不同能级之间跃迁就会吸收或发射谱线，不同的原子和分子有其特定的谱线。天体的辐射特性可以用不同波长处强度(即，流量)的分布来描述。天光背景、宇宙射线等因素会对光谱有一定的影响。同时，我们观测到的天体辐射也会受到探测器热噪声、读出噪声以及一些未知源等的影响。因此，天体光谱主要是由连续光谱和叠加在其上的吸收线、发射谱线和噪声组成的。

由于天体亮度或距离地球远近的不同，我们观测到的同一类天体的辐射能量在数量级上可能也会相差很大。例如，假设一条光谱的理想观测是 $\boldsymbol{x}_I=(x_1, x_2, \cdots, x_n)^{\mathrm{T}}$，但是由于上述因素的影响，我们观测到的数据可用下面的模型近似

$$\alpha\boldsymbol{x}_I = (\alpha x_1, \alpha x_2, \cdots, \alpha x_n)^{\mathrm{T}}$$

其中 $\alpha>0$。也就是说，同一类天体光谱的流量数量级(即，单位或量纲)会因它本身的亮度或距离等的差异而成为一个可变因素。所以，这种情况下光谱流量数量级是一个不稳定的因素，它增加了自动分类的难度、复杂性和分类的误差。需要通过相应的预处理以消除或抑制它对自动分类的不利影响。本文通过实验研究了该模型的合理性和有效性。

对于任意 i，$j\leqslant n$，因为 $\alpha x_i/\alpha x_j = x_i/x_j$，所以这时谱线之间的相对强弱没有受到影响。也就是说，这种情况下，光谱中谱线强度的相对比值仍具有可比性。同时，在天文上认为，表征天体的是光谱在不同波长处流量的相对比例。也就是说，两条光谱只要各条谱线之间的相对比例一样，则它们代表同一个天体的光谱；不同的线强比代表了不同类型的天体。因此，光谱的各个分量同时放缩一个因子，不应该影响自动处理的结果。

1.2　流量标准化方法

假设 $\boldsymbol{x}$ 是一条光谱，记为：$\boldsymbol{x}=(x_1, x_2, \cdots, x_n)^{\mathrm{T}}$，它是 n 维空间中的一个向量；$\sigma(x)$ 是由 $\boldsymbol{x}$ 定义的一个标量。由前面对光谱特点的分析知道，如果 $\sigma(x)$ 与 $\boldsymbol{x}$ 的各个分量同数量级，则以

$$\boldsymbol{y} = \boldsymbol{x}/\sigma(x) \tag{1}$$

代替原始的光谱 $\boldsymbol{x}$ 不仅能消除数量级的影响，而且不改变自动处理的结果。下面我们通过定义不同的标准化因子 $\sigma(x)$ 给出几种具体的标准化方法：

单位化(unit)

$$\sigma(x) = \sqrt{\sum_{i=1}^{n} x_i^2} \tag{2}$$

这是目前光谱自动处理中常用的一种流量标准化因子[12]。

均值(mean)

$$\sigma(x) = \sum_{i=1}^{n} x_i/n \tag{3}$$

如果对光谱 $\boldsymbol{x}$ 的各个分量从小到大排序 $\mathrm{sort}(\boldsymbol{x})=(x_{(1)}, x_{(2)}, \cdots, x_{(n)})$，则可以定义如下标准化因子：

中值(median)

$$\sigma(x) = \mathrm{median}(x_1, x_2, \cdots, x_n) = \begin{cases} x_{([n+1])/2}, & n\text{ 为奇数} \\ (x_{(n/2)} + x_{(n/2+1)})/2, & n\text{ 为偶数} \end{cases} \tag{4}$$

最大值(max)

$$\sigma(x) = \max(x_1, x_2, \cdots, x_n) = x_{(n)} \tag{5}$$

最小值(min)

$$\sigma(x) = \min(x_1, x_2, \cdots, x_n) = x_{(1)} \tag{6}$$

为了下文书写方便，我们把与(2)~(6)式对应的流量标准化方法分别记为 S_{unit}，S_{mean}，S_{median}，$S_{\max}$，$S_{\min}$。其中，S_{unit} 是目前光谱自动处理中常用的标准化方法。

2　实　验

在光谱的自动分类研究中，会面临到各种各样的分类问题，例如，恒星、星系、晚型星等；而且，对具体的分类问题合适的分类器也多种多样。但是，光谱流量预处理是一种基础性的问题，研究中我们充分考虑了所选用数据和方法的代表性。

2.1　数据

因为Galaxy和Qso一般红移较大、距离比较远，受到的噪声影响较多，这增加了自动分类的难度，所以，我们针对这两种天体光谱的分类问题通过实验研究了光谱的流量标准化问题。对这两类天体，我们从SDSS发布的DR2数据库中各随机选择4 000条数据。波长范围截取3 800~9 000，步长为0.5 nm。

2.2　k 近邻方法

在本研究中，我们选用 k 近邻方法[14]。对于一个两分类问题，假设未知类别样本 x 的 k 个近邻中有 k_i 个样本来自第 i 类，$i=1, 2$，且 $k_{i_0}=\max_{i=1,2} k_i$。如果使用 k-近邻方法进行分类，则样本 x 被决策为来自第 i_0 类。因为这种算法不涉及太多的数学计算，易于实现，而且在训练样本足够多的情况下，它具有良好的性能，这已经在理论上得到了证明[14]，所以，它不仅在大量的模式识别问题中得到了成功的应用，例如文献[15]，手写数字识别，卫星图像和EKG分类等；而且，已经成为评价其它分类方法的比较标准。所以，该方法适于本文的研究工作。

2.3　主成分分析

因为光谱数据的维数很大，在我们的实验中是1070维，所以需要通过特征提取和特征选择来进行特征约简，这样不仅能提高处理效率，还能够在一定程度上抑制噪声影响。主成分分析(PCA)是一种无监督的非参数方法，主要目的是寻找最小均方意义下最能够代表原始数据的投影方向。该方法已经在信号处理上得到成功应用，关于它的详细介绍请参考文献[16]。

2.4　实验结果

为了保证实验在统计意义上的代表性，每个实验都重复十次，每次都是从2.1节所述的实验数据库中随机选择一半作为训练集，剩余的为测试集。我们在4维PCA空间中分别测试了不同标准化方式下分类性能(正确率)及其稳定性(本文指对于不同的训练数据和测试数据，分类正确率的变化情况)，结果如下：

1　有效性，可参见表1。

Table 1 Mean of the classification results under the above standardization methods in 4 dimensional PCA space (%)

标准化方法	qso 正确率	galaxy 正确率	整体正确率
不标准化	93.81	90.55	92.18
S_{unit}	94.98	89.87	92.43
S_{mean}	96.45	92.43	94.44
S_{min}	92.78	88.52	90.65
S_{median}	95.77	91.70	93.73
S_{max}	98.46	91.04	94.75

2 稳定性, 可参见表 2。

Table 2 Standard deviation of the classification results under the above standardization methods in 4 dimensional PCA space

标准化方法	qso 正确率均方差	galaxy 正确率均方差	整体正确率均方差
不标准化	0.147	0.123	0.114
S_{unit}	0.141	0.117	0.109
S_{mean}	0.045	0.041	0.026
S_{min}	0.064	0.067	0.025
S_{median}	0.091	0.147	0.113
S_{max}	0.035	0.041	0.013

3 分析与结论

(1) 如表 1 中的实验结果所示, 流量标准化后(S_{unit}, S_{mean}, S_{median}, S_{max})分类性能得到了明显提高, 这充分说明了 1.1 节所述因素的影响和流量标准化的必要性。但是, 使用 S_{min} 方法流量标准化后分类效果反而更差, 对此, 后面将会做进一步的分析。

(2) 由(2)式和(3)式对标准化因子的定义知道, 在目前常用的流量标准化方法方法 S_{unit} 中, 各个波长处的噪声干扰逐渐积累起来; 而 S_{mean} 流量标准化中噪声影响能够互相抵消。所以, 正如表 1 和表 2 中的实验结果所示, S_{unit} 标准化方法受噪声干扰严重, 它在有效性和稳定性方面都要明显差于 S_{mean} 方法。

(3) 如果假定不同波长处的噪声干扰独立同分布(或近似独立同分布), 根据中心极限定理[17], 当光谱向量的维数趋向于无穷大时, 均值标准化因子中的噪声干扰依概率 1 收敛于 0; 也就是说, 只要光谱向量的维数足够高, 噪声在标准化方法 S_{mean} 中的影响可以忽略不计。而光谱一般是上千维空间中的向量(我们实验中的是 1070 维), 所以正如表 1 和表 2 中的实验结果所示, 标准化方法 S_{mean} 在有效性和稳定性方面都非常好。

(4) 在(4)~(6)式中, 如果对 $\sigma(x)$ 作如下分解 $\sigma(x)=\sigma_p(x)+\varepsilon=\sigma_p(x)(1+\varepsilon/\sigma_p(x))$, 其中 $\sigma_p(x)$ 表示去除噪声影响后理想条件下的标准化因子, ε 代表噪声干扰项。因为 $1+\varepsilon/\sigma_p(x)\to 1(\sigma_p(x)\to+\infty)$, 也就是说, 对于给定的噪声, $\sigma_p(x)$ 越大噪声相对影响越小, 所以标准化方法 S_{min}, S_{median}, S_{max} 的分类效果依次提高(见表 1)。

参 考 文 献

[1] Kent S M. Astrophysics and Space Science, 1994, 217(1-2): 27.

[2] Connolly A J, Szalay A S. Astron. J., 1995, 110(3): 1071.

[3] Galaz G, Lapparent V. Astronomy and Astrophysics, 1998, 332(2): 459.

[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] 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.

[6] LI Xiang-ru, WU Fu-chao, HU Zhan-yi, et al(李乡儒, 吴福朝, 胡占义, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2005, 25(11): 1889.

[7] XU Xin, LUO A-li, WU Fu-chao, et al(许 馨, 罗阿理, 吴福朝, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2005, 25(6): 996.

[8] ZHANG Jian-nan, WU Fu-chao, LUO A-li, et al(张健楠, 吴福朝, 罗阿理, 等). Astronomical Research and Technology——Publications of National Astronomical Observatories of China(天文研究与技术——国家天文台台刊), 2004, 1(4): 249.

[9] ZHANG Jian-nan, WU Fu-chao, LUO A-li, et al(张健楠, 吴福朝, 罗阿理, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2005, 25(12): 2088.

[10] ZHANG Jian-nan, WU Fu-chao, LUO A-li, et al(张健楠, 吴福朝, 罗阿理, 等). Acta Astronomica Sinica(天文学报), 2005, 46(4): 406.

[11] Theodoridis S, Koutroumbas K. Pattern Recognition. United States: Academic Press, 1999.

[12] XU Xin, WU Fu-cha, HU Zhan-yi, et al(许 馨, 吴福朝, 胡占义, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2006, 26(1): 182.

[13] HE Xiang-tao(何香涛). Observational Cosmology(观测宇宙学). Beijing: Science Press(北京: 科学出版社), 2002. 4.

[14] BIAN Zhao-qi, ZHANG Xue-gong(边肇祺, 张学工). Pattern Recognition (模式识别). Beijing: Tsinghua University Press(北京: 清华大学出版社), 2000.
[15] Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. Springer-Verlag, 2001.
[16] Duda R O, Hart P E, Stork D G. Pattern Classification, Second Edition. Beijing: China Machine Press, 2003.
[17] WEI Zong-shu(魏宗舒). Probability and Statistics(概率论与数理统计教程). Beijing: Higher Education Press(北京: 高等教育出版社), 1983. 10.

Celestial Spectrum Flux Standardization for Classification

LI Xiang-ru[1, 2], LIU Zhong-tian[1, 2], HU Zhan-yi[1 *], WU Fu-chao[1], ZHAO Yong-heng[2]
1. National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100080, China
2. National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100012, China

Abstract Celestial spectra should be preprocessed before automated classification to eliminate the disturbance of noise, observation environment, and flux aberrance. In the present work, the authors studied the spectrum flux standardization problem. By analyzing the disturbing factors and their characteristics, the authors put forward a theoretical model for spectra flux, and correspondingly give several flux standardizing methods. The rationality/correctness of the model, and the satisfactory performance of the proposed methods have been obtained by the experiments over normal galaxies (NGs) and quasi-stellar object (Qso). Furthermore, the authors theoretically analyze, compare and evaluate them. In particular, this work indicated that the conventional method is worse than the proposed one. And the investigation is also particularly significant for other automatic spectrum processing study, e.g. redshift determination, effective temperature, metallic estimation, etc.

Keywords Celestial spectrum classification; Flux model; Flux standization; Preprocessing; Qso; Principal component analysis (PCA)

(Received Dec. 6, 2005; accepted Mar. 16, 2006)

* Corresponding author