第31卷,第10期
2011年10月

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

Vol. 31, No. 10, pp2856-2860
October, 2011

一种新的光谱特征提取方法

李乡儒[1], 冯春明[2], 王永俊[1], 卢 瑜[1]

1. 华南师范大学数学科学学院, 广东 广州 510631
2. 聊城大学东昌学院数学系, 山东 聊城 252000

摘 要 研究了天体光谱的特征提取问题, 这是光谱自动处理中的一个关键环节。通过特征提取, 不仅能够约简数据、减少冗余, 而且亦能抑制噪声干扰, 对识别系统的精度和效率均有重要影响。提出了一种基于空间转换和分解的特征分析模型(STP), 基于此, 可实现对常用光谱特征提取方法的分析, 例如, 无监督的主成分分析(PCA), 小波变换(Wavelet), 有监督的支持向量机(SVM), 相关向量机(RVM)和线性判别分析方法(LDA)等。在STP模型中, 关注的核心要素是特征提取中对数据成分的分解、重组, 以及噪声的抑制和冗余的消除。亦在STP框架的基础上, 给出了一种逻辑和实现均较为简单的特征提取方法: 基于曲线拟合与下采样的光谱特征提取(EFCD)。研究的一个重要发现是, 在一些分类问题中文献中设计巧妙的特征提取方法并不一定是决定性的: 即使采用通常的信号下采样方法提取特征, 亦能获得良好的光谱识别性能, 而重要的仅仅是需要将特征数量保持在一定的水平以上即可。研究中, 选用的测试数据是SDSS中的Galaxy和QSO两类河外天体实测光谱, 他们一般具有较大的红移, 在天体光谱识别中具有较强的代表性。

关键词 天体光谱分类; 光谱特征提取; 类星体; 正常星系

中图分类号: TN911.7 **文献标识码**: A **DOI**: 10.3964/j.issn.1000-0593(2011)10-2856-05

引 言

天体光谱主要是由连续谱、吸收谱线、发射谱线和噪声组成。其中, 噪声的来源一般是天光背景、大气吸收和仪器设备的不理想性, 有时宇宙射线也会成为干扰源。另外, 天体光谱一般是上千维空间里的一个向量, 但是天文学家通过认证少数几条特征谱线即可识别出光谱的类型, 赵梅芳等研究表明可通过截取并分析几个特征波段对窄线活动星系核和宽线活动星系核光谱进行分类(例如, 特征谱线 H_α, H_β, [OⅢ]或[NⅡ]周围)[1], 李乡儒等通过提取一维线性判别特征即可实现星系与类星体, 以及类星体与Seyfert2光谱的分类。上述已有研究表明, 对于基于光谱的天体识别来说, 原始观测数据中往往存在许多数据冗余。而噪声和冗余的存在往往会造成自动分类系统的过学习和计算负担, 并表现为系统的精度和效率不能满足需要。同时, 一个光谱自动分类系统一般包括数据采集、数据预处理、特征提取、识别子系统和后处理等五个部分[2]。其中, 特征提取和识别子系统是光谱分类系统中两个不可分割的部分, 它们相辅相成, 协同完成分类目标: 特征提取模块不仅需要提取与分类目标有关的信息, 尽可能抑制其他与当前任务无关的信息(相对于当前目标来说, 这些都是噪声), 而且需要把信息转化成适合分类器特点的表达方式, 它直接关系到分类器的泛化性能(对新样本的识别能力)和设计的难度。所以特征提取是光谱数据自动分类研究中的一个关键内容。

本工作的部分结果报告可见文献[3]。

1 相关研究

目前光谱自动分类研究中常使用的特征提取方法有主成分分析(PCA), 小波变换(Wavelet), 均值漂移(Mean shift), 以及有监督相关向量机(RVM), 支持向量机(SVM), 覆盖算法和线性判别分析(LDA)等方法。上述工作简要综述如下:

(1)PCA

在星系和类星体光谱识别研究方面, Gaspar[4]采用PCA方法进行特征提取, 并对红移值已知的星系光谱的自动分类作了研究。李乡儒等结合Galaxy和Qso的识别问题探讨了光谱自动处理中的流量标准化问题, 在该工作中首先使用PCA进行特征提取和特征选择, 然后在4维PCA空间中运

收稿日期: 2011-01-03, **修订日期**: 2011-04-10

基金项目: 国家自然科学基金项目(61075033)资助

作者简介: 李乡儒, 1972年生, 华南师范大学数学科学学院副教授 e-mail: xiangru.li@gmail.com

用 k 近邻法针对不同标准化方式进行了研究。

(2)Wavelet

邢飞等通过 Mallat 小波分解算法，并将分解后的高频子空间作为噪声成分丢弃，实现光谱信号的特征提取，然后使用支持向量机方法研究了恒星光谱的分类。刘蓉等基于小波变换技术研究了星系光谱的分类[5]，在该研究中首先使用 Spline2 小波对光谱信号进行分解，并以第四层小波系数作为光谱谱线信息的描述，进而在谱线描述空间中使用主成分分析方法对数据进行约简单，最后利用 Fisher 线性判别分析方法实现正常星系和活动星系光谱的分类。张怀福等通过使用小波包提取光谱特征研究了活动天体和非活动天体光谱的分类[6]，他们首先使用 Daubechies 3 小波对光谱进行三层小波包分解，并以第三层中各个频率子空间中的能量累积和作为光谱分类特征的描述，形成一个 8 维小波包特征空间，然后在该空间中使用 SVM 方法对红移未知的活动天体和非活动天体光谱进行分类。

(3)Mean shift

均值漂移算法是一种局部模式搜索方法，他可用于信号滤波、目标跟踪、图像分割等领域[7]。段福庆和刘蓉等研究了该方法在光谱滤波和特征谱线提取方面的应用[8,9]，研究结果表明该方法能够去除脉冲噪声，抑制非脉冲噪声、光背景噪声和随机噪声，较强的谱线保护能力，整体上优于小波硬阈值法、高斯滤波和中值滤波方法。

(4)有监督特征提取

李乡儒等在结合类星体光谱、星系光谱、Seyfert 1 光谱和 Seyfert 2 光谱的分类问题研究了基于 LDA 和 RVM 的光谱特征提取，并探讨了光谱识别中有监督特征提取的必要性和重要性。

另外，赵梅芳等在文献[1]中，采用 k 近邻方法研究了红移已知的窄线活动星系核和宽线活动星系核光谱数据的识别，在该研究中，首先根据给定的红移，将光谱移回静止状态，然后根据窄线活动星系核和宽线活动星系核分类相关的特征谱线方面的知识，截取部分波段的流量数据，最后使用 k 近邻方法对光谱数据进行识别。这可以归为一种基于局部特征的方法，这类方法在计算机视觉中得到了深入的研究和广泛的应用[10][11]。

2　特征分析的空间变换与分割法

特征提取是对信号的测量指标进行整合、重组和取舍的过程，其目的是去除冗余、噪声、并将信号转化为利于后续处理的表达方式。该过程可以分解为两个环节，数据表达空间的转换和特征的选择。空间转换是实现将数据的表达从一个空间转换到另一个空间

$$R^n \rightarrow R^m = L(\nu_1, \nu_2, \cdots, \nu_m) \tag{1}$$

它实现了对数据信息的重组，其中 $\nu_1, \nu_2, \cdots, \nu_m$ 是空间 R^m 的某个基向量，$L(\nu_1, \nu_2, \cdots, \nu_m)$表示由 $\nu_1, \nu_2, \cdots, \nu_m$ 生成的线性空间，m 和 n 是非负整数。在这个过程中实现了对原有测量指标的分解、重组，以方便后续步骤中剔除数据冗余和测量指标间的耦合。需要指出的是，不同的特征提取方法，往往有不同的特征空间 R^m，或确定了特定的基向量。为了方便后文的阐述，我们将 R^n 和 R^m 分别称为数据空间和特征空间。特征选择过程实际上是按照某个准则将特征空间的给定基向量进行排序和分组

$$\{\nu_1, \nu_2, \cdots, \nu_m\} \rightarrow \{\nu_{(1)}, \nu_{(2)}, \cdots, \nu_{(m)}\} \tag{2}$$

$$\{\nu_{(1)}, \nu_{(2)}, \cdots, \nu_{(m)}\} = \{\nu_{(1)}, \nu_{(2)}, \cdots, \nu_{(k)}\} \bigcup \{\nu_{(k+1)}, \cdots, \nu_{(m)}\} \tag{3}$$

由此可得到特征空间的一个分解

$$R^m = F + RN = L(\nu_{(1)}, \nu_{(2)}, \cdots, \nu_{(k)}) + L(\nu_{(k+1)}, \cdots, \nu_{(m)}) \tag{4}$$

其中，F 是最终的特征子空间，RN 代表冗余数据和噪声形成的子空间。为了后文阐述方便，称 F，RN 和该特征分析框架分别为特征子空间，干扰子空间和空间变换与分割(space transformation and partition，STP)法。另外，由于只有特征子空间是后续处理中被使用到，而噪声和冗余数据则需要抛弃，所以在一些特征提取算法中只显式地计算特征子空间 F 及其基向量$\{\nu_{(1)}, \nu_{(2)}, \cdots, \nu_{(k)}\}$。

例如，在主成分分析中，变换空间 $R^m = L(\nu_1, \nu_2, \cdots, \nu_m)$，是由观测空间 R^n 旋转得到，且 $n = m$；在 PCA 空间 R^m 中，不同数据成分之间互相独立，$\{\nu_{(1)}, \nu_{(2)}, \cdots, \nu_{(m)}\}$ 是通过对$(\nu_1, \nu_2, \cdots, \nu_m)$ 按照各个成分对观测数据的描述能力排序得到。在 SVM 和 RVM 中，数据表征空间的变换通过 $\{\nu_i = k(x_i, \cdot), i = 1, \cdots, m\}$ 实现，$m = l$，其中，l 是训练数据的数量，$D = \{x_1, x_2, \cdots, x_l\}$ 是训练数据，$k(\cdot, \cdot)$ 是核函数；对于特征的选择，在 SVM 方法中通过结构风险最小化实现，在 RVM 中，通过 Bayesian 分层框架实现。在 LDA 中，通过最大化类间距离和最小化类内距离直接计算特征空间 F。

3　基于曲线拟合与下采样的光谱特征提取

由上述分析可知，特征提取过程实际上可以认为是计算特征子空间的基$\{\nu_{(1)}, \nu_{(2)}, \cdots, \nu_{(k)}\}$。在计算该基的过程中，需要考虑的因素包括有效成分的提取，以及冗余和噪声的剔除。实验研究表明，可以通过曲线拟合的方法消除或减轻噪声的影响。对于拟合曲线，当沿着波长移动的时候，流量是一种渐变过程，在相邻波长之间，流量变化常常非常小，所以，在拟合光谱中往往存在相当严重的数据冗余。因此，本工作通过对拟合结果进行下采样实现冗余的剔除和特征的提取，例如，如果要提取两个特征，特征子空间的基可以表达为

$$\nu_{(1)} = f(\cdot, \lambda_1)$$

$$\nu_{(2)} = f(\cdot, \lambda_2)$$

其中，λ_1 和 λ_2 是两个给定的波长值，$f(\cdot, \lambda)$是一个函数，其作用就是对于给定的一条光谱 x，计算它的拟合曲线，并在给定的波长 λ 处进行采样。所以，曲线拟合与下采样方法的特征子空间是一个泛函空间。对于任何一个光谱 x，提取的特征就是对原始光谱进行拟合后的一系列采样值。该特征提取方法包括两个环节：拟合，采样。通过拟合过程抑制了噪声的影响。采样点的计算一般仅需要局部的某几个数据点的计算，所以通过采样实现特征提取能够提高计算效率。

在实现中，仅仅在少数几个波长位置对光谱流量进行采样，所以不必将整个 f 进行拟合，只需要根据给定波长周围的少量几个观测点估计采样值，这样可大幅度提高特征提取的速度。

4 实验研究

下面通过类星体与星系光谱的识别问题对本文方法做出定量的评估。在本工作研究中，识别方法选择最近邻方法，实验数据选用 Sloan 发布的一维 Galaxy 和 QSO 光谱数据[12] 各 4 000 条，所在天区编号分别是 0267—0276 和 0267—0389，波长范围截取为 380.0～900.0 nm。由于研究表明，在光谱识别中采用对数波长-流量数据格式有较好的效果[13]，所以本研究中采用该数据格式。

4.1 步骤与结果

不同天体之间亮度和距离的差异会导致观测到的光谱数据在流量数量级上有一定差异，而且巡天观测中一般只进行流量的相对定标，所以同一类天体的观测光谱之间可能存在流量数量级的不确定性。观测数据数量级的不确定性会对光谱的识别造成一定的负面影响，所以在光谱识别之前需要进行流量标准化。

综上所述，本研究实验步骤是，首先使用 S_{median}，S_{unit}，S_{mean}，S_{max} 或 S_{min} 方法对光谱数据进行流量标准化，然后将数据表达转换为对数波长-流量格式，继而使用本文方法提取光谱特征，最后在特征空间中使用最近邻方法进行光谱识别。

为了保证实验结果的统计意义，每个实验都独立重复十次，每次均从 3.1 节所述的实验数据库中为每类随机选择 3 000条光谱数据作为训练集，剩余的作为测试集。10 次独立重复实验的平均识别率统计结果如图 1 和表 1 所示。

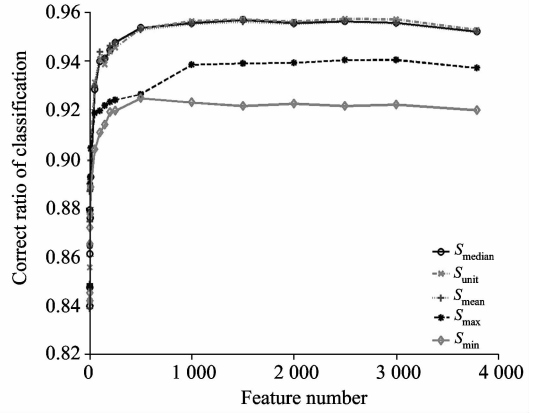

Fig. 1 The quasar and galaxy spectra classification performance based on EFCD feature extraction method and nearest neighbor classifier

4.2 噪声和数据冗余影响的评估

如前所述，特征提取的要点是去除冗余和噪声影响，噪声的存在会降低识别算法的精度，数据冗余不仅会造成效率的降低，也会削弱系统的泛化能力，在本部分中设计实验对噪声和冗余的影响分别作了定量评估。

Table 1 The quasar and galaxy spectra classification performance based on EFCD method and nearest neighbor classifier

DEF	CR based on S_{median}/%	CR based on S_{mean}/%	CR based on S_{unit}/%	CR based on S_{max}/%	CR based on S_{min}/%
3 791	95.21	95.19	9 527	9 371	9 199
3 000	95.57	95.53	9 571	9 405	9 221
2 500	95.62	95.62	9 573	9 404	9 216
2 000	95.56	95.52	9 561	9 392	9 225
1 500	95.68	95.61	9 572	9 390	9 216
1 000	95.56	95.51	9 563	9 385	9 230
500	95.36	95.28	9 533	9 264	9 247
250	94.76	94.75	9 455	9 241	9 196
200	94.44	94.62	9 439	9 232	9 190
150	94.10	94.10	93.86	92.18	91.40
100	93.99	94.36	94.07	91.96	91.07
50	92.83	92.94	93.12	91.87	90.38
10	89.25	89.91	90.40	90.43	88.84
5	87.57	88.74	89.10	89.24	87.72
4	87.89	88.66	89.06	88.96	87.18
3	86.11	87.42	87.69	87.80	86.50
2	84.72	84.86	85.54	86.43	84.49
1	83.96	83.87	84.17	84.78	84.16

DEF: Dimension of the extracted feature space, CR: Correct ratio

4.2.1 噪声影响的评估

在噪声影响的评估中，通过将数据维数固定于 3 791 屏蔽掉冗余因素对结果的干扰。为了降低噪声的影响，则对原始观测数据进行了高斯滤波，其尺度参数考虑了 σ=1，2，3 和 4 等四种情况，试验结果如表 2 所示。对于任何一种尺度，滤波器窗口越宽，则会使噪声的平滑能力越强，从而能更好地抑制噪声的影响，在表 2 的试验结果中表现为，每列中的识别率结果随窗宽增加而提高。另外，对于同样的核宽，尺度参数越大，滤波器对噪声的平滑能力亦变强，所以，表 2 中每行的识别率结果整体上亦随尺度因子的增加而提升；但是，当平滑能力超过一定限度后亦会对判别信息有一定削减，例如，(WGW,σ)=(3，5)情况下的识别率比(WGW,σ)=(3，3)的低。

Table 2 The influence of noise on quasar and galaxy spectra classification performance

σ	1	2	3	5
WGW=3	94.33	94.32	94.27	94.25
WGW=5	94.38	94.76	94.85	94.88
WGW=7	94.36	94.97	95.17	95.09
WGW=9	94.36	95.06	95.17	95.32

σ is the scale parameter in Gaussian filter. Flux standization method is S_{max}. The classification correct ratio is presented based on percent. WGW: the width of Gaussian window

4.2.2 数据冗余影响的评估

在数据冗余影响的评估中，通过直接间隔选择原始观测流量值的方式进行特征提取，由于没有进行滤波、拟合或插值估计，所以对噪声影响均未处理。在实验中，每次从原有观测数据中扔掉约 50%的观测数据，识别结果如表 3 所示。由实验结果可见，在原始光谱中确实存在相当严重的数据冗余，例如，在使用 S_{mean} 方法进行流量标准化的情况下，当数据维数由 3 791 降至 118 时，识别率仅仅下降 1.08%，仍然高达 94.11%。

Table 3　Evaluating the redundancy in quasar and galaxy spectra

DEF	CR based on S_{median}/%	CR based on S_{mean}/%	CR based on S_{unit}/%
3 791	95.21	95.19	95.27
1 895	95.18	95.15	95.19
947	95.34	95.32	95.37
473	94.88	94.69	94.49
236	94.44	94.42	94.31
118	93.80	94.11	93.70
59	92.46	92.65	92.72
29	91.29	91.67	91.94
14	89.27	90.00	90.64
7	88.06	89.61	89.24
3	83.77	86.13	86.18
1	74.25	77.34	79.70

The notations are defined in the title of Table 1

5　分析与总结

如前所述，本文推荐的特征提取方法包括拟合和采样两个环节。通过“拟合”平抑了噪声的影响，改善了光谱的可识别性。下“采样”的作用则是两方面的：一方面，它减少了数据冗余，降低了冗余数据对分类器效率和性能的负面影响，另一方面，由于本文方法并未对光谱数据进行重组和分解，不同数据项之间有一定耦合性，所以当冗余性降低到一定程度后，继续进行简单的下采样降维，则会降低识别性能。因此，基于本文特征提取方法光谱识别有以下特点：

(1) 随着特征维数的降低，光谱识别性能先提高后降低(表 1，图 1)。

(2) 在光谱数据量由 3 791 维降至 500 维的过程中，光谱识别性能均比原始光谱的识别能力高。

(3) 即使降低至一维的时候，光谱识别率仍然高达 83.955 0(表 1，图 1)。

(4) 由于本文特征提取方法的基本操作是两个实值数据的加权平均，每提取一个特征，仅需两次乘法和一次加法运算，而传统的数据约减方法，提取特征需要的计算量要大得多，例如，如果使用主成分分析方法对 3 791 维的光谱数据进行特征提取，每提取一个特征需要 3 791 次乘法和 3 790 次加法。所以，本文方法的效率要显著地高，这在大型光谱巡天数据处理中有重要意义。

通过对推荐方法特点的分析可知，本文方法除了可以直接应用外，还可以与其他更精确复杂的方法融合起来使用，即序贯识别：根据上述特点(3)和(4)，可首先使用本文方法进行特征提取并识别，然后根据对识别结果可信度的估计，使用更精确的方法对识别结果可信度低的光谱进行分类处理，通过这种分级识别的方案，可大幅度提高系统的整体效率。

References

[1] ZHAO Mei-fang, WU Chao, LUO A-li, et al(赵梅芳，吴　潮，罗阿理). Acta Astronomica Sinica(天文学报), 2007, 48(1): 1.

[2] LI Xiang-ru, HU Zhan-yi, ZHAO Yong-heng, et al(李乡儒，胡占义，赵永恒，等). Spectroscopy and Spectral Analysis (光谱学与光谱分析), 2009, 29(6): 1702.

[3] Li X, Lu Y, Wang Y. International Conference on Information Science and Technology, 2011. 712.

[4] Gaspar G, Valerie L. Astronomy and Astrophysics, 1998, 332: 459.

[5] LIU Rong, DUAN Fu-qing, LIU San-yang, et al(刘　蓉，段福庆，刘三阳，等). Acta Electronica Sinica(电子学报), 2005, 33(11): 2059.

[6] ZHANG Huai-fu, ZHAO Rui-zhen, LUO A-li(张怀福，赵瑞珍，罗阿理). Journal of Beijing Jiaotong University(北京交通大学学报), 2008, 32(2): 30.

[7] Li X, Hu Z, Wu Fuchao. Pattern Recognition, 2007, 40(6): 1756.

[8] DUAN Fu-qing, ZHOU Ming-quan, ZHANG Jia-cai(段福庆，周明全，张家才). Journal of Jilin University (Engineering and Technology Edition)(吉林大学学报·工学版), 2007, 37(3): 634.

[9] LIU Rong, DUAN Fu-qing, LIU San-yang, et al(刘　蓉，段福庆，刘三阳，等). Journal of Ecttonics&Information Teehnology(电子与信息学报), 2006, 28(2): 312.

[10] Li X, Hu Z. International Journal of Computer Vision, 2010, 89(1): 1.

[11] Li X R, Li X M, Li H, et al. Acta Automatica Sinica, 2009, 35(1): 17.

[12] Abazajian K N, Adelman-McCarthy J K, Agüeros M A, et al. The Astrophysical Journal Supplement, 2009, 182(2): 543.

[13] LI Xiang-ru, HU Zhan-yi, ZHAO Yong-heng, et al(李乡儒，胡占义，赵永恒，等). Acta Astronomica Sinica(天文学报), 2007, 48(3): 280.

A Novel Spectrum Feature Extraction Method

LI Xiang-ru[1], FENG Chun-ming[2], WANG Yong-jun[1], LU Yu[1]
1. School of Mathematical Sciences, South China Normal University, Guangzhou 510631, China
2. Dongchang College, Liaocheng University, Liaocheng 252000, China

Abstract The present focuses on the celestial spectra feature extraction problem, which is a key procedure in automatic spectra classification. By extracting features, the authors can reduce redundancy, alleviate noise influence, and improve accuracy and efficiency in spectra classification. The authors introduced a novel feature analysis framework STP (space transformation and partition), which focuses on four essential components in feature extraction: decompose and reorganize spectrum components, reorganize, alleviate noise influence and eliminate redundancy. Based on STP, we can analyze most of the available feature extraction methods, for example, the unsupervised methods principal component analysis (PCA), wavelet transform, the supervised methods support vector machine (SVM), relevance vector machine (RVM), linear discriminant analysis (LDA), etc. We introduced a novel feature analysis framework and proposed a novel feature extraction method. The outstanding characteristics of the proposed method are its simplicity and efficiency. Researches show that it is sufficient to extract features by the proposed method in some cases, and it is not necessary to use the sophisticated methods, which is usually more complex in computation. The proposed method is evaluated in classifying Galaxy and QSO spectra, which is disturbed by red shift and is representative in automatic spectra classification research. The results are practical and helpful to gain novel insight into the traditional feature extraction methods and design more efficient spectrum classification method.

Keywords Spectrum classification; Feature extraction; Quasar; Galaxy

(Received Jan. 3, 2011; accepted Apr. 10, 2011)