第31卷，第9期
2011年9月

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

Vol. 31, No. 9, pp2582-2585
September, 2011

基于最近邻方法的类星体与正常星系光谱分类

李乡儒[1]，卢 瑜[1]，周建明[2]，王永俊[1]

1. 华南师范大学数学科学学院，广东 广州 510631
2. 潍坊教育学院会计与统计学院，山东 青州 262500

摘 要 随着高质量CCD传感器技术的日渐成熟与广泛应用，以及许多大型巡天计划的相继实施，天体数据量极大，因此天体观测数据的自动识别、分析问题首当其冲。文章在原始测量空间使用最近邻方法(NN)研究了正常星系与类星体光谱的识别问题。正常星系和类星体属于河外天体，一般距离地球较远，其观测光谱会受到许多干扰，所以这两类天体光谱的分类在光谱自动识别研究中具有一定的代表性。同时，采用的NN是模式识别和数据挖掘方面的基准性方法，在许多新方法的评估中，往往以NN方法的性能作为比较对象。从实用价值来说，研究表明，NN方法的类星体和正常星系光谱识别率与文献中复杂方法的最好结果相当，但该文方法不需要进行分类器的训练，利于实时进行增量式学习和并行实现，这对海量光谱数据的快速处理有重要意义。因此，该研究具有重要的理论参考意义和一定的实用价值。

关键词 天体光谱分类；最近邻方法；正常星系；类星体

中图分类号：TN911.7 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2011)09-2582-04

引 言

天体光谱是天体电磁辐射按照波长的有序排列，蕴含着天体重要的物理信息，如天体的化学成份、各元素的丰度、天体的表面温度、光度、直径、质量以及天体的视向运动和自转。天文学家和天体物理学家通过分析天体光谱信息，不仅可以研究宇宙中物质的分布特征，还可以研究天体的形成和随时间的演化等重大科学问题。因而光谱的处理和分析在天文学，尤其是天体物理研究中具有特别重要的意义。

随着高质量CCD传感器技术的日渐成熟和广泛应用，以及许多大型巡天计划的实施，如斯隆数字巡天(sloan digital sky survey，SDSS)计划[1]，2度视场星系红移巡天(two-degree-field galaxy redshift survey，2dF)计划[2]，美国的光谱巡天望远镜(spectroscopic survey telescope，SST)项目[3]，大天区面积多目标光纤光谱天文望远镜(large sky area multi-object fiber spectroscopic telescope，LAMOST)巡天项目[4]和大型综合巡天望远镜(large synoptic survey telescope，LSST)计划等，天体观测数据的获取速度正在大幅度提高，例如，我国正在实施的LAMOST项目，计划每个观测夜将获得数万条光谱数据。在这种情况下，完全依靠天文学家人工进行识别光谱的方式已不能适用，需要研究海量天体光谱的自动识别技术。本文研究了基于最近邻方法的类星体与正常星系光谱分类。

正常星系和类星体属于河外天体，一般来说它们距离地球较远，尤其是类星体，距地球有几十亿光年以上，光线穿过广袤的宇宙空间，会受到许多干扰，如穿越星系、星云时被吸收，产生吸收线，星际物质的反射和折射、大气吸收、天光背景、水气以及其他天体发出的光。例如，LAMOST的目标天体暗达20.5星等(mangitude)，天体本身就很暗，获得的光谱必将有低信噪比的特点。预计，LAMOST观测光谱的信噪比在10～15范围内。当信噪比很低时，噪声影响会造成“假”特征谱线的选取。所以，正常星系和类星体的分类在光谱识别研究中具有一定的代表性。而且，本文采用的最近邻方法是模式识别和数据挖掘方面最经典、最成熟、最有代表性的方法之一，它是模式识别领域中的基准性方法，在许多模式识别方法的评估中，往往以最近邻方法的性能作为比较对象。因此，本研究具有重要的理论参考意义和一定的实用价值。

1 相关研究

由于正常星系和类星体的分类在光谱识别研究中的典型性和代表性，关于这两类天体识别的研究受到了较为广泛的

收稿日期：2010-11-19，修订日期：2011-03-24

基金项目：国家自然科学基金项目(61075033，60805028)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

关注。李乡儒等[5,6]结合 Galaxy 和 QSO 的光谱识别问题探讨了光谱自动处理中的流量标准化问题和光谱数据的表达问题，由于光谱数据的维数很高，为了降低计算量，在该工作中首先使用主成分分析(PCA)进行特征提取和特征选择，然后在 4 维 PCA 空间中运用近邻方法针对不同标准化方式进行了研究；结合类星体光谱、星系光谱、Seyfert 1 光谱和 Seyfert 2 光谱的分类探讨了光谱识别中有监督特征提取的必要性和重要性。许馨等通过将 Fisher 判别分析方法与核技巧结合起来研究了恒星、星系和类星体的光谱分类问题，并称该方法为广义判别分析(generalized discriminant analysis，GDA)方法。在该方案中，首先通过核方法诱导出非线性映射将光谱数据映射到某个高维特征空间 F 中，然后在空间 F 中进行线性判别分析。杨金福等提出了一种基于核技巧的覆盖算法(核覆盖算法)，并研究了它在正常星系、恒星、星暴星系和活动星系核光谱分类中的应用。该算法将核技巧与覆盖算法相结合，并在特征空间中抽取支持向量。赵梅芳等比较系统地研究了活动星系核和星系的分类[7]，研究了基于自适应径向基神经网络的类星体和星系光谱识别，她们首先通过单位化实现光谱标准化，去除流量数量级不确定性对识别的影响，然后使用主成分分析方法进行特征提取，实现对光谱数据的约简，最后使用自适应径向基神经网络方法对光谱数据进行分类，在此方法中，神经网络能够根据待处理问题的情况自适应地增加神经元个数，给识别系统的训练带来一定的方便；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[8]。

2　识别方法

在本研究中，我们选用最近邻方法，它作为模式分类领域一种简单而有效的分类方法，有着广泛的应用。对于一个两分类问题，假设有训练数据 $\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$，和待识别样本 $x\in R^N$，且 $\mathrm{dist}(x,x_{i_0})=\min\{\mathrm{dist}(x,x_1),(x,x_2),\cdots,(x,x_n)\}$，其中 $x_i\in R^N$ 是观测数据，$y_i\in\{1,-1\}$ 是类别标示。按照最近邻方法，则样本 x 被决策为来自第 i_0 类。虽然在各种识别问题的解决方案中，最近邻分类器很少能算得上是巧妙和高效的方法，但和其他某些方法不同，我们可以应用最近邻方法来解决广阔领域的各种问题(实际上，它可能是唯一一种几乎什么问题都能解决的一般性方法)。如果要解决的问题实例不太多，而且最近邻方法可以用一种能够接受的速度对问题求解，那么设计一个更高效算法所需花费的代价很可能是不值得的。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，图象识别，机器人定位，目标识别，检索，雷达信号识别和机械故障诊断等[9,10]；而且，在很多模式识别方法的评估中，往往以最近邻方法的性能作为比较对象。

3　类星体与正常星系光谱分类

3.1　数据及其表达

Galaxy 和 QSO 一般红移较大、与地球距离比较远，由此导致这两类天体的观测光谱受到的噪声影响较大，相应的分类问题在光谱识别研究中具有一定的代表性。因此，我们在本工作中研究了这两种天体光谱的识别问题。对这两类天体，我们从 SDSS 发布的光谱数据库中分别随机选择 4 000 条数据[11]。波长范围截取为 380～900 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，所以本研究中采用这种数据格式[5,6]，并为每条光谱均匀采样 3 791 个点(此即原始测量数据)。

3.2　实验步骤与结果

由于不同天体在亮度和距离一般均有差异，其结果是观测到的天体光谱流量的数量级会有较大的不确定性，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一致性。也就是说，同一类天体的光谱，会由于天体在亮度和距离方面的差异而使观测到的光谱流量成为一个随机变量，这对光谱的自动识别会造成严重的负面影响，所以在光谱识别之前需要进行流量标准化[5,6]。为此，提出了如下的流量标准化模型

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

其中，$x=(x_1,x_2,\cdots,x_n)^{\mathrm{T}}$ 是原始观测光谱，x' 是流量标准化后的光谱，$\sigma(x)$ 是标准化因子，它是一个标量[5,6]。文献[5,6]通过定义不同的标准化因子 $\sigma(x)$ 给出了多种流量标准化方法，并推荐使用 $S_{\max}$，S_{median}，S_{mean} 和 S_{unit} 四种标准化方法，其中的流量标准化因子分别定义如下

$$\sigma_{\max}(x)=\max(x_1,x_2,\cdots,x_n)=x_{(n)} \tag{2}$$

$$\sigma_{\mathrm{median}}(x)=\mathrm{median}(x_1,x_2,\cdots,x_n)=\begin{cases}x_{([n+1]/2)} & n\text{ 为奇数}\\ (x_{(n/2)}+(x_{(n/2+1)}))/2 & n\text{ 为偶数}\end{cases} \tag{3}$$

$$\sigma_{\mathrm{mean}}(x)=\sum_{i=1}^{n}x_i/n \tag{4}$$

$$\sigma_{\mathrm{unit}}(x)=\sqrt{\sum_{i=1}^{n}x_i^2} \tag{5}$$

综上所述，本研究实验步骤是，首先对光谱数据进行流量标准化，然后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表 1 所示。

3.3　分析与比较

表 2 和表 3 分别是使用 PCA＋KNN，LDA，PCA＋SVM 和 GDA 方法对 Galaxy 和 QSO 光谱分类的结果[5,6]。由表 1、表 2 和表 3 的结果可见，本文方法的分类正确率 95.74%明显高于 LDA，PCA＋SVM 和 GDA，与 PCA＋KNN 方法虽然各有高低，但相差相对不明显。

Table 1 Correct recognition ratio of QSO and Galaxy spectra based on nearest neighbor method. The nearest neighbor classifier is designed in the spectrum observational space, which is not processed by the usual filters or feature extraction methods

Standardization method	S_{unit}	S_{mean}	S_{median}	S_{max}
Correct ratio/%	95.265 0	95.185 0	95.205 0	93.710 0

Table 2 The classification results of Galaxy spectrum and QSO spectrum based on PCA+KNN, LDA and PCA+SVM. In [5, 6], the $\sigma_{unit}(x)=\sqrt{\sum_{i=1}^{n}x_i^2}$ is implemented mistakenly by $\sum_{i=1}^{n}x_i^2$, we corrected it in this work

Standization method	S_{unit}	S_{mean}	S_{median}	S_{max}
PCA+KNN/%	95.74	94.44	93.73	94.75
LDA/%	92.41	92.16	91.57	91.57
PCA+SVM/%	90.94	92.38	91.73	93.41

Table 3 The classification results of Galaxy spectra and QSO spectra based on GDA method

Spectrum class	Galaxy	QSO	average
Correct ratio/%	90.65	94.0	92.325

基于 PCA+KNN, LDA, PCA+SVM 和 GDA 方法的光谱识别系统实现过程，可分为学习过程和生产过程两个阶段，例如，PCA+KNN 方法需要在学习阶段根据训练数据获取主成分子空间，在 PCA+SVM 在学习阶段还需要在主成纷子空间中判别支持向量和设计非线性分类器，在 GDA 和 LDA 中需要在学习阶段根据已观测数据计算最易于分类的特征方向。如果在生产过程中，又得到了一部分监督数据 S_{app}，若要将 S_{app} 添加到训练集中，以使分类器有不断学习的能力，则需要在扩展的训练集上重新进行学习过程，设计新的分类器，这增加了分类系统使用的难度和时间消耗。但是，在本文方法中则完全没有分类器学习阶段，在生产阶段可根据问题的发展，通过直接调整训练集，而分类器则不需要进行任何变更，所以该方法便于分类系统的扩展和演化，有更好的环境适应性。

另外，如果将训练数据集记为 S_{tr}，则可通过将它分为若干个子集 S_{tr_1}，S_{tr_2}，…，S_{tr_k}，并在各个子集上分别搜索最近邻，实现本文方法的空间并行化，以适应于大数据量的快速处理，而 PCA+KNN，LDA，PCA+SVM 和 GDA 方法则比较难于实现生产过程的空间并行化。

总之，最近邻方法的类星体和正常星系光谱识别率与复杂、巧妙的方法 PCA+KNN，LDA，PCA+SVM 和 GDA 的最好结果相当，但本文方法不需要进行分类器的训练，利于实时进行增量式学习和并行实现，这对海量光谱数据的快速处理有重要意义。所以，本文研究具有一定的实用价值。

4 结 论

本文使用最近邻方法研究了正常星系与类星体的识别问题。正常星系和类星体属于河外天体，一般均有较大的红移影响，且他们的观测光谱数据往往会受到许多噪声干扰，所以这两类天体光谱的分类在光谱识别研究中具有一定的代表性。同时，本文采用的最近邻方法是模式识别和数据挖掘方面最经典，最成熟、最有代表性的方法之一，它是模式识别领域中的基准性方法，在许多模式识别方法的评估中，往往以最近邻方法的性能作为比较对象。而且本工作研究表明，最近邻方法的类星体和正常星系光谱识别率与文献中复杂方法的最好结果相当，但本文方法不需要进行识别器的训练，利于实时进行增量式学习和并行实现，这对海量光谱数据的及时处理有重要意义。因此，本文工作在光谱识别方法研究中将具有重要的理论参考意义和一定的实用价值。

References

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

[2] Shanks T, Boyle B J, Croom S M, et al. ESO Astrophysics Symposia: Mining the Sky. Berlin: Springer, 2001. 143.

[3] Ramsey L W, Sebring T A, Sneden C A. Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series. Hawaii: Society of Photo-Optical, 1994. 31.

[4] CHU Yao-quan(褚耀泉). Journal of University of Science and Technology of China(中国科学技术大学学报), 2007, 37(6): 591.

[5] 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.

[6] Li Xiangru, Hu Zhanyi, Zhao Yongheng, et al. Chinese Astronomy and Astrophysics, 2008, 32(1): 13.

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

[8] Li Xiangru, Hu Zhanyi. International Journal of Computer Vision, 2010, 89(1): 1.

[9] Boiman O, Shechtman E, Irani M. IEEE Conference on Computer Vision and Pattern Recognition, 2008. 1.

[10] TU Zhi-song, HAO Wei, LI Ling-jun, et al(涂志松, 郝 伟, 李凌均, 等). Coal Mine Machinery(煤矿机械), 2009, 30(8): 237.

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

Galaxy/Quasar Classification Based on Nearest Neighbor Method

LI Xiang-ru[1], LU Yu[1], ZHOU Jian-ming[2], WANG Yong-jun[1]
1. School of Mathematical Sciences, South China Normal University, Guangzhou 510631, China
2. Weifang Educational College, Qingzhou 262500, China

Abstract With the wide application of high-quality CCD in celestial spectrum imagery and the implementation of many large sky survey programs (e.g., Sloan Digital Sky Survey (SDSS), Two-degree-Field Galaxy Redshift Survey (2dF), Spectroscopic Survey Telescope(SST), Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST) program and Large Synoptic Survey Telescope (LSST) program, etc.), celestial observational data are coming into the world like torrential rain. Therefore, to utilize them effectively and fully, research on automated processing methods for celestial data is imperative. In the present work, we investigated how to recognizing galaxies and quasars from spectra based on nearest neighbor method. Galaxies and quasars are extragalactic objects, they are far away from earth, and their spectra are usually contaminated by various noise. Therefore, it is a typical problem to recognize these two types of spectra in automatic spectra classification. Furthermore, the utilized method, nearest neighbor, is one of the most typical, classic, mature algorithms in pattern recognition and data mining, and often is used as a benchmark in developing novel algorithm. For applicability in practice, it is shown that the recognition ratio of nearest neighbor method (NN) is comparable to the best results reported in the literature based on more complicated methods, and the superiority of NN is that this method does not need to be trained, which is useful in incremental learning and parallel computation in mass spectral data processing. In conclusion, the results in this work are helpful for studying galaxies and quasars spectra classification.

Keywords Spectra classification; Nearest neighbor method; Galaxy; Quasar(quasi-stellar object, QSO)

(Received Nov. 19, 2010; accepted Mar. 24, 2011)