一种基于 Hough 变换和神经网络的分层类星体识别方法 [1]

周 虹 黄凌云 [2] 罗曼丽

(中国科学院自动化所国家模式识别实验室 北京 100080)

摘 要 类星体是宇宙中最明亮、密集的天体. 它产生于宇宙诞生早期, 具有重要的研究价值. 观测的类星体光谱由于红移现象, 光谱向长波方向偏移, 因此识别类星体观测光谱中的发射线和确定类星体的红移是类星体识别的主要目标. 类星体光谱固有的高噪声和观测光谱特性, 给类星体识别带来很大困难. 一般来说基于规则的直接匹配方法在类星体识别中效果不佳. 本文介绍一种神经网络和 Hough 变换 (HT) 结合的类星体自动识别方法. 该方法具有简单、快速、高效、鲁棒性强和通用性强等特点.

关键词 类星体, Hough 变换 (HT), 神经网络

中图号 TN-052, TP391.4

1 引 言

类星体是宇宙中已知的最明亮天体, 被认为是活跃的星系中心. 类星体是宇宙诞生早期的产物 [1], 也是了解宇宙起源的重要线索, 具有重要的研究价值. 我国近期内计划研制一台大型天体望远镜, 该望远镜在每个观测夜晚将能收集多达 3 万个光谱数据 [2]. 要实时处理这些数据, 目前的手工方法显然不再适用, 需要研究自动识别方法.

对类星体的识别存在两个困难. 第一是观测到的类星体光谱噪声极其严重. 第二是观测到的光谱和静止光谱之间存在红移现象. 所谓红移现象, 是指由于类星体的高退行速度, 使得观测到的发射线波长比静止发射线波长长, 表征该现象的数值是红移值. 设类星体静止发射线波长为 λ, 该发射线的地面观测波长为 λ', 红移值为 Z, 则如下关系式成立:

$$Z = (\lambda' - \lambda)/\lambda. \tag{1}$$

类星体识别的目的就是确认类星体发射谱和红移值. 目前有关类星体自动识别方法研究的文献很少, 我们根据类星体光谱的特性和类星体识别的目的, 将类星体识别分成下面几个子问题: (1) 寻找类星体观测光谱随红移变化的规律; (2) 特征提取, 去除连续谱, 保留发射谱; (3) 确认发射峰和红移. 在下面的各节中, 我们将对上述三部分进行详细的介绍.

2 类星体静止光谱和观测光谱的特点

静止光谱是指没有发生红移时的类星体光谱, 它的形状主要由连续谱和发射谱组成 [1], 如图 1 所示. 连续谱由类星体的强辐射形成, 在紫外波段, 随着波长加长, 辐射强度逐渐变小. 类星体的组成元素发出的静止波长和辐射强度固定的发射线称为发射峰, 只由发射峰组成的光谱为发射谱. 目前天文学家推测静止发射谱波长范围为 1000 ~ 7000Å. 图 1 中标出的峰值为发射峰, 峰值处标注的是形成该发射峰的元素.

[1] 1998-09-24 收到, 1999-08-09 定稿
国家自然科学基金 (69675009) 资助课题

[2] 通信作者

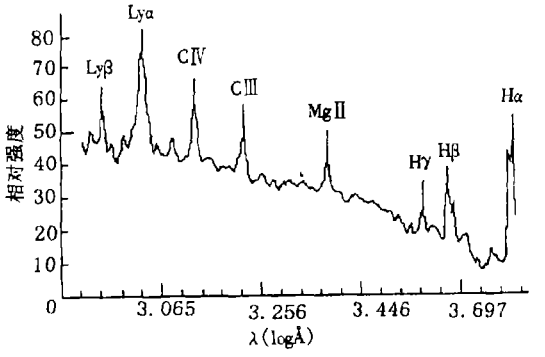

图 1 类星体静止光谱图 (λ 表示波长对数)

类星体距地球有几十亿光年，光线穿越宇宙空间时会受到许多噪声源的干扰，因此观测光谱存在严重的噪声。如图 2 和图 3，图中标注的峰，如 Mg Ⅱ、 Hα 、 Hβ 是识别出的真实发射峰，其它峰均为由噪声形成的假峰，且数目远远多于真峰。图 2 的红移为 0.5420，图 3 的红移为 0.4400 。

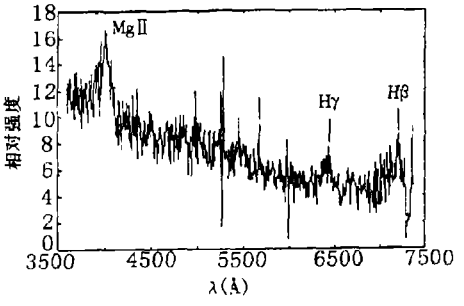

图 2 典型类星体观测光谱
(λ 表示波长)

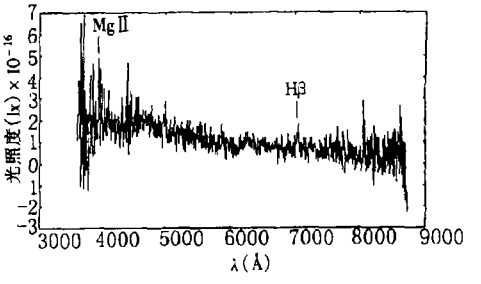

图 3 噪声严重的类星体观测光谱
(λ 表示波长)

观测光谱有如下特征：(1) 由于红移，观测谱线波长比静止谱线波长大，发射峰间的间隔成比例增加；(2) 受观测仪器的限制，观测范围在 3700 ~ 7800Å ，为整个类星体波长的一部分；(3) 观测到的发射谱受噪声影响严重，谱线少，谱线强度变化大，连续谱的形状变化大。

3 模拟静止和红移后的发射谱

基于天文学家提供的类星体主要发射线表 (含主要发射线的波长，相对强度和相对宽度等信息)，我们构造了静止和红移后的发射谱。用高斯轮廓模拟发射峰，模拟方程为

$$f(x) = He^{-[(x-x_0)/w]^2}, \tag{2}$$

其中 x_0 为发射线的波长， H 为谱线的相对强度。类星体静止发射峰有两种：宽线和窄线。宽线时 w 取 100Å ，窄线时 w 取 10Å 。每个发射峰用 100 个点模拟，再将各模拟发射峰按波长迭加，得到类星体的静止发射谱，如图 4 所示。

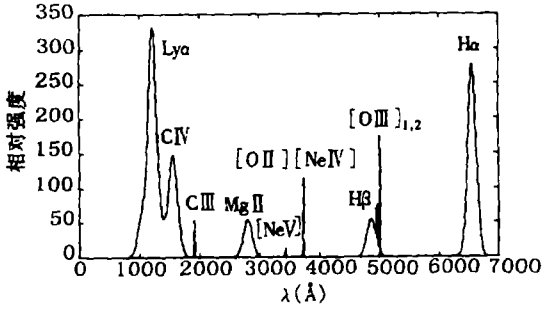

图 4　模拟的类星体静止发射谱

假设红移值为 Z，由红移公式 $\lambda' = (1+Z)\lambda$[1] 分别求出每个模拟发射峰红移后的状况，再将各红移后的模拟发射峰叠加，取可观测范围 3700 ~ 7800Å，得到不同红移下的模拟观测发射谱. 图 5 为红移为 1 时的模拟观测发射谱，图 6 为红移为 2 时的模拟观测发射谱.

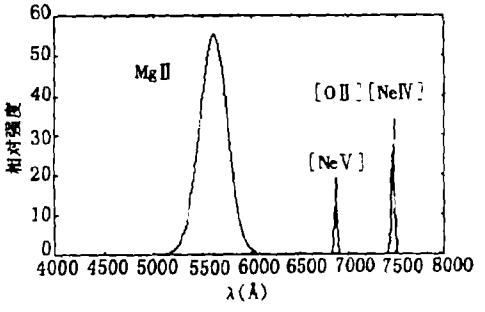

图 5　红移为 1 时模拟的观测发射谱

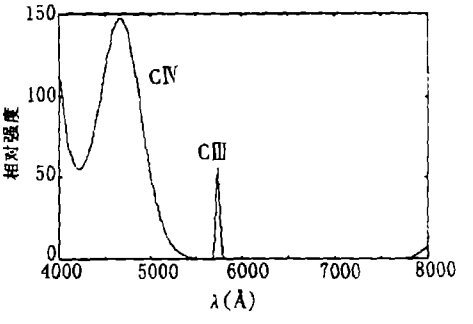

图 6　红移为 2 时模拟的观测发射谱

比较静止和红移后的模拟发射谱，得到发射谱随红移变化的规律为：

当 $Z \leq 0.2$ 时，观测谱中只有由 Hβ、[OIII]$_1$ 和 [OIII]$_2$ 合成的峰和 Hα 形成的峰；

当 $0.2 < Z \leq 0.7$ 时，由 MgII 形成的峰逐渐变大，由 Hβ、[OIII]$_1$、[OIII]$_2$ 合成的峰仍存在，而由 Hα 形成的峰逐渐消失；

当 $0.7 < Z \leq 1.2$ 时，仅 MgII 形成的峰比较明显，其它峰较小，该段的识别难度最大；

当 $1.2 < Z \leq 2$ 时，由 CIII 形成的窄峰出现，由 SiIV + OIV、CIV 和 HeII 形成的峰逐渐变大，而由 MgII 形成的峰逐渐消失；

当 $2 < Z$ 时，由 OVI、HI 和 NV 形成的峰出现。

上述规律以及模拟的发射谱为识别和验证工作提供了依据，特别是模拟的静止谱已成为识别发射峰的模板. 这一模拟有助于今后对类星体性质和自动识别的研究.

4 基于多尺度形态滤波和自适应节点选取的峰值提取

选取合适的节点集合，通过样条函数逼近，以分段直线代替连续谱曲线，可以取得连续谱的参数描述. 这对于提取发射谱、定量描述连续谱特征以及推理都有重要作用. 连续谱逼近过程分为初选节点、粗逼近连续谱、粗分割凸区、重选节点、细逼近连续谱以及细分割凸区等步骤. 其中“粗”与“细”的差别在于逼近时选用的数据集合不同，前者使用所有的数据，而后者只用粗逼近过程中认为是连续谱上的点.

样条逼近要取得满意的结果，节点选取应合适．节点数太多则逼近结果对噪声敏感；否则凸区的宽度太大．为了得到合适的节点集合，应采用自适应过程．

首先采用空间选择性滤波方法，在一系列尺度上分别滤波，随后对所有尺度下的结果逐点相乘，综合各个结果，真特征得到加强，而伪特征得到减弱．然后划分出一些凸区，通过分裂、合并，得到满足发射区必要条件的区域．

再用分段样条直线逼近连续谱，除发射区外，每个波长处的辐射强度对于类星体连续谱逼近都有一定贡献．实验证明，在逼近过程中阶次选取为一阶，已可以获得比较满意的逼近效果．

然后通过分割发射区提取类星体的发射谱．提取出的类星体光谱的凸区分为三类：第一类高而宽，可能是真的发射区；第二类高度中等但较宽，可能是几个发射区叠加而成，也可能是虚假的发射区；最后一类是窄发射线或噪声．由于发射区与连续谱相交，故分割的目的是获取交点．本系统采用多次逼近策略得到连续谱的近似值．对每个选取出的凸区，保留其特征参数，如峰的高度、位置、面积、起始波长和终止波长等．

5 基于 Hough 变换和神经网络的红移确定和反推发射谱法

神经网络和 Hough 变换相结合的识别系统结构如图 7 所示．

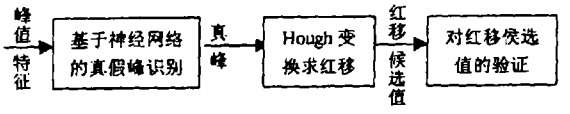

图 7　确定红移结构图

5.1 基于 BP 多层神经网络的真假峰识别

真假峰识别单元的任务是确定提取的凸区对应的是真峰还是假峰，当输入凸区被认为是真峰时，该单元并不能确定输入峰对应何种元素的发射峰，也就是说，真假峰识别是一个（0/1）分类器．由于关于真假峰的专家知识较少，很难通过规则区分，故选择神经网络解决此问题比用顺序推理规则更合理．

神经网络在应用中必须解决以下问题 [4]：

(1) *神经网络的结构、层数和每层的节点数的选择*　目前常用的适用于各种专门问题的神经网络模型就有十几种，各模型又有不同的改进算法．真假峰的识别问题属于有监督的两类分类问题，适合用前向多层神经网络 (MLP) 实现．理论上三层 MLP 可以对任何可分集合进行分类，所以我们选用三层 MLP 进行真假峰的识别．

三层 MLP 包括输入层、隐层和输出层，每一层都和下一层全连接．输入层选择能描述发射峰性质的特征如发射峰的波长、高度、面积和反映发射峰间位置关系的发射峰与光谱中最大峰的距离为输入参数．如果神经元选择 sigmoid 函数的 MLP 有 N 个单元，由输入单元定义了一个 N 维空间．选择任意隐层单元或输出单元，该单元确定了一个 N 维空间上的超平面，平面的输出一边为 0，另一边为 1，权值决定了这个超平面在输入空间的位置．一般要求这个超平面穿过输入空间的原点，而这个限制对很多问题不合适，所以我们增加了一个输入恒为 1 的输入节点，使超平面不必穿过原点．除该点外，为使其它输入特征在调节权值时具有相同的贡献，我们将它们归一化到 $[0, 1]$．

隐层节点数的选择目前无普适方法而依赖于训练集的大小、噪声量、求解问题的复杂度等因素．通常有两种方法，一是从大到小，二是从小到大．本系统根据删选法，从大到小地选择合适的节点数．先选择 2^5 为隐层节点数初值，然后在输入样本集相同的条件下进行反复测试，最后在实验中得到当节点数为 10 ~ 20 时，效果较好，凭经验取 15．输出为一个节点，具有 0 / 1 值表示真峰或假峰．

神经元隐层和输出层都选用 sigmoid 型非线性函数.

(2) *学习和训练问题*　BackPropagation(BP) 是一种适用于函数映射和分类的 MLP 的有监督学习方法，使用梯度下降法最优化误差函数. 但 BP 的收敛很慢，且对于学习率 η 和权的初值敏感. 特别是 η 的选择对收敛速度影响很大. η 太小网络收敛慢，反之则权和误差函数发散. 因此需要改进学习算法和训练方法加快收敛速度. 当误差函数有许多局部和全局最优点时，η 在训练过程中往往变化很大. 目前有很多调节 η 的算法，如高阶梯度法、加入动量法和动态启发式调节 η 等 [5] .

本系统采用快速 BP 学习算法，加入动量因子和动态调节 η 方法加快收敛速度. η 初值取 2，增量为 1.01 倍，减量为 0.99 倍，动量取 0.9，平均误差和目标取 0.02 . 所有这些参数初值的选取都是先根据经验取一初值，然后调节该值，在其它条件不变的情况下，选较优的. 然后整体调节，选择整体性能较优的参数初值组合. 初值的选取问题主要凭经验和测试，目前还没有完善的理论指导.

为加快学习速度，训练集和训练方法的选择非常重要. 批处理训练虽然能加快收敛速度，但是容易进入局部极小. 实验表明，批处理训练不适合本问题. 我们选用在线训练方法，即每一样本训练一次后调节权值，然后依次反复训练. 另外，采用分类错误率下降法修改权值.

(3) *推广问题*　目前收集到的类星体光谱数据有限，在训练集有限的情况下提高测试集的识别率是个重要问题. 通常认为选用较小的网络利于训练和推广，本系统用删除法优化网络结构，即先用大网络结构，然后删除一个或若干个隐层节点，再训练，再删除，$\cdots$，最后在保证网络性能的前提下选择较小的网络. 同时用加噪声法训练网络，交替输入典型和易混样本.

5.2 Hough 变换求红移

HT 是一个强有力的形状分析工具，抗噪声能力强 [6,7] . 本系统运用 HT 技术确定红移，效果明显.

由第一节可知，红移公式为

$$\lambda_i + Z\lambda_i - \lambda'_j = 0. \tag{3}$$

式中 λ_i 为静止发射线波长；λ'_j 为观测谱中提取的发射峰的波长；Z 为红移值.

以 Z 为参数进行 Hough 变换. 在参数空间中，将连续的 Z 值分成若干个均匀小区间，每个区间对应一个一维数组中的元素，这个数组叫累加数组 (accumulator array)，数组中的元素叫计数器 (accumulator) . 当由每对 (λ_i, λ'_j)，$(i \in [1,n]$，$j \in [1,m])$ 算出的 Z 值落在某个区间内时，与该区间对应的累加数组中的元素增加一定值. 将所有的 (λ_i, λ'_j) 对变换、累加后，取累加数组中所有大于某一确定阈值的计数器对应的值为红移参考值.

为了有助于参数空间中峰的增强，本系统采用了加权 HT(WHT) 方法. 用静止发射线的相对强度和观测峰的相对强度比 E_i/e_j . $(i \in [1,n]$，$j \in [1,m])$ 对计数器的增量进行加权，公式如下：

$$t = \left\{ \frac{AE_i/e_j, \quad E_i \le e_j}{Ae_j/E_i, \quad E_i \ge e_j} \right\}, \qquad A\text{为常数}, \quad t\text{为增量}. \tag{4}$$

由于目前的观测数据有两种格式，难以准确估计发射线的相对强度，实验证明 WHT 对系统的改进效果不明显.

当观测光谱中发射峰较少时 (如只含 2 ~ 3 个发射峰)，经 HT 后累加数组中的峰不明显. 另外提取的发射峰的波长偏移会使“票数”在累加时分散到相邻的单元. 上述两种原因都会导致累加数组中的峰发散，使得直接取累加数组中峰值的方法难以得到真正的红移值.

另外当噪声干扰严重时，类星体的光谱由许多小峰组成并含有大量假峰，这些假峰会使累加数组中对应错误红移的计数单元出现尖峰，如图 8 所示。

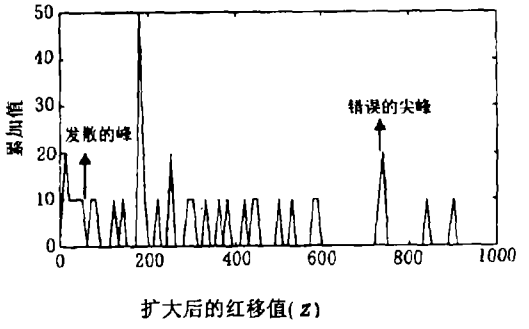

图 8 累加矩阵中的峰 (z 为扩大 1000 倍后的红移值)

尖峰和峰发散现象严重影响了 HT 的效果。峰发散问题是 HT 经常遇到的问题，文献中有几种 HT 后处理方法可以使峰突出 [6,7]，如 2D 核方法 [8]。本系统根据累加数组的特性采用线性核的方法进行 HT 后处理，实验证明该后处理方法有一定的改进效果。

5.3 红移侯选值验证

从观测光谱提取的发射峰经 HT 后得到若干个红移侯选值，这些红移候选值必须经过验证才能确定哪个值更接近真正的红移值。设选出的红移候选值为 Z^*，根据发射谱随红移变化的规律，取模拟的静止发射谱中经红移 Z^* 后可以观测到的部分为红移后的发射谱，将其与观测发射谱进行相似度比较，取相似度最大的模拟的红移后的发射谱对应的红移值为最后的红移值

5.4 统计和反馈学习

统计和反馈学习部分是一个完整的自动识别系统不可缺少的部分，它担任着不断完善和提高系统性能的任务。本系统对红移的正确识别率，正确红移对应累加数组中峰值的比率，红移参考值中包含正确红移的比率和可信度等分别做了统计。并将统计结果反馈到各个系统单元，特别是验证部分。

6 实验结果

基于 BP 的三层神经网络识别真假峰的实验结果如表 1 。

表 1 BP 神经网络识别真假峰的结果

隐层节点个数	训练次数	训练集大小	训练集识别率	测试集大小	测试集识别率
10	~200	68	63%	290	31%
15	~600	68	72%	290	69%
25	~1000	68	72%	290	69%
50	~1000	68	75%	290	69%
100	>5000	68	82%	290	75%

真假峰识别单元对 Hough 变换求红移的影响的实验结果如下：(1) 未引入真假峰识别单元时，将所有提取的峰作为 Hough 变换的输入，正确红移值在红移侯选值 (对应于 HT 累加数组中最大峰和次大峰的那些红移值) 中的比率为 88%；(2) 引入真假峰识别单元后，只有识别的真峰作为 Hough 变换的输入，正确的红移值在红移侯选值中的比率为 100%；(3) 红移侯选值经过验证后，最后的正确识别率分别为：未引入真假峰识别单元时识别率为

20%，引入真假峰识别单元后识别率为 76%。上述结果表明，基于神经网络的真假峰识别单元有效地改进了 Hough 变换求红移的效果。

7 结　论

本文介绍了一个基于神经网络和 Hough 变换相结合的类星体自动识别系统。本文介绍的方法虽然最终的正确识别率只有 76%，但考虑到类星体的高噪声和观测光谱特性，该识别率可以认为已是一种比较满意的效果。特别值得指出的是，本文引入的基于神经网络的真假峰识别模块对提高系统的整体性能起了至关重要的作用。另外本文提出的分层思路也可推广到其它识别问题中。

参　考　文　献

[1] Robert A M, Steven N S. Encyclopedia of Astronomy and Astrophysics. San Diego: Academic Press, 1989, 571–574.

[2] 中国科学院. LAMOST 项目计划建议书. 1995 年 9 月.

[3] 吴永东，马颂德. 多尺度形态滤波弹性匹配技术在类星体谱线识别中的应用. 中国图形图象学报，1997, 2(1): 1–6.

[4] 焦李成. 神经网络计算. 西安：西安电子科技大学出版社，1993, 第二章.

[5] Janakiraman J, HonaVar V. Adaptive learning rate selection for backpropagation network. Proceeding of SPIE'93 Orlando, Florida: 1993, 1–17.

[6] Illingworth J, Kittler J. A survey of the Hough transform. Computer Vision, Graphics, and Image Processing, 1988, 44(1): 87–116.

[7] LeaVers V F. Survey which Hough transform. CVGIP: Image Understanding, 1993, 58(2): 250–264.

[8] Palmer P L, Petrou M, Kittler J. Hough transform algorithm with a 2D hypothesis testing kernel. CVGIP: Image Understanding, 1993, 58(2): 221–234.

A STRATIFIED APPROACH FOR QUASAR RECOGNITION BASED ON HOUGH TRANSFORM AND NEURAL NETWORK

Zhou Hong　　Huang Lingyun　　Luo Manli

(*Institute of Automation, Academia Sinica, Beijing* 100080)

Abstract　Quasar Objects (QSOs) are detectable at very large distance, with broad, redshifted emission lines, strong ultraviolet and strong time variability of the optical light. QSOs play an important role in the research of the universe. The main purposes of quasar recognition are to identify the emission peaks in an observable quasar spectrum and to determine the observable quasar's redshift value. Due to the inherent extremely noisy characteristics of quasar spectrums and the limitation of observable conditions, automatic quasar recognition is a hard problem to tackle, and the commonly used direct matching approaches based on rules are ineffective. This paper introduces a stratified approach based on Hough transform and neural network which is shown to be simple, efficient, robust and easy to generalize.

Key words　Quasar, Hough transform, Neural networks

周　虹：　女，1973 年生，硕士研究生，研究兴趣：模式识别，人工智能.

黄凌云：　男，1976 年生，硕士研究生，研究兴趣：模式识别，信号与图象处理.

罗曼丽：　女，1938 年生，教授，主要从事模式识别，图象处理，多媒体研究.