第 33 卷 第 3 期 计 算 机 工 程 2007 年 2 月
Vol.33 No.3 Computer Engineering February 2007
·人工智能及识别技术· 文章编号：1000—3428(2007)03—0177—03 文献标识码：A 中图分类号：TN911.7

LAMOST 光谱数据自动处理软件系统

杨金福[1]，吴福朝[1]，罗阿理[2]，赵永恒[2]
(1. 中科院自动化所模式识别国家重点实验室，北京 100080；2. 中国科学院国家天文台，北京 100012)

摘 要：介绍了 LAMOST 光谱数据自动处理软件系统的设计及实现。阐述了系统实现中所使用的核心算法：基于覆盖算法的光谱分类方法，基于小波变换的晚型恒星识别算法和基于均值漂移的红移求取算法。在美国 SDSS 的天体数据库上测试表明，该软件系统具有较快的处理速度，并且能够获得较高的分类正确率和红移计算精度，可以满足大型巡天计划的实际需求。
关键词：大天区面积多目标光纤光谱天文望远境；系统设计；覆盖算法；小波变换；均值漂移

Auto-processing Software System for LAMOST Celestial Spectra

YANG Jinfu[1], WU Fuchao[1], LUO Ali[2], ZHAO Yongheng[2]
(1. National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100080;
2. National Astronomical Observatory, Chinese Academy of Sciences, Beijing 100012)

【Abstract】 This paper introduces the design and implementation of an automatic processing software system for LAMOST celestial spectra. Some key algorithms in the system such as covering algorithm, wavelet transform, mean shift, etc., are also presented. The experiment results over the datasets of sloan digital sky survey (SDSS) show that this system has good characters in processing speed, classifying accuracy and redshift measuring precision. And the implemented system is workable for the large survey projects.
【Key words】 LAMOST; System design; Covering algorithm; Wavelet transform; Mean shift

大天区面积多目标光纤光谱天文望远境（LAMOST）项目是国家重大科学工程计划。项目完成后将有大量的光谱数据产出，每个观测夜将获得 1 万~2 万条光谱，在观测期内将达到 10^7数量级的光谱数据。这样一个庞大的天文数据，利用人工方法来处理光谱分类和红移测量显然不能满足实际的需求。因此，需要研究借助计算机来实现对天体光谱自动分类和红移测量，以满足天文学家在此基础上中进行科学研究的需要。

目前，国际上常用的天文软件包有MIDAS、FIGARO、IRAF等，它们都是通过人机交互的方式来完成光谱处理的。虽然利用人工交互的方式处理光谱时可以融入天文学家的专家知识，并根据天文学家的经验做出判断和分析。但面对象LAMOST这样大型巡天计划的海量数据，如果还采用人工交互方式，靠天文工作者逐条分析光谱，是件非常困难的事情。美国的Sloan数字巡天计划(Sloan Digital Sky Survey，SDSS)的光谱自动处理软件算法主要是采用模板匹配和交叉认证的方法[1]。但SDSS与LAOMOST不同，SDSS的光谱是经过流量定标的，而LAMOST的光谱则并没有进行流量定标，因此不能照搬SDSS的方法，必须研究出适合LAMOST巡天项目的自动处理方法，并开发出相应的处理软件。

本文根据LAMOST的实际情况，研究了基于覆盖技术、小波变换、K近邻和均值漂移的光谱分类与红移测量算法[3~5]，并在此基础上设计并开发出天体光谱自动处理软件系统。

1 系统设计

按照天文上划分，天体光谱可以划分为正常星系、恒星、星暴星系、活动星系核等，其中活动星系核又可以分为类星体、SEYFERT1、SEYFERT2、LINER 等。正常星系和恒星的特性表现为光谱的谱线是吸收线，而星暴星系和活动星系核的谱线则是以发射线为主。依据这样的划分方法，我们在系统设计时，分类器结构采用如图 1 所示的树状结构。

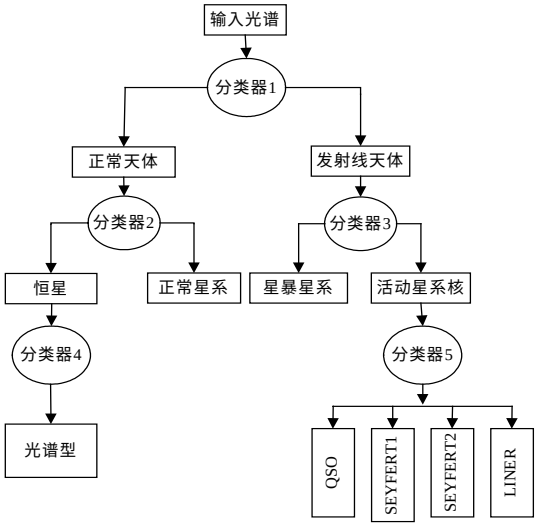

图 1 分类器设计

首先利用分类器 1 将光谱分为正常天体和发射线天体(其中正常天体包括正常星系和恒星，发射线天体包括星暴星系和活动星系核)，然后对正常天体利用分类器 2 分为恒星和正常星系，对活动天体利用分类器 3 分为星暴星系和活动星

基金项目：国家“ 863 ”计划基金资助项目(2003AA133060)；国家重大科学工程 LAMOST 计划基金资助项目
作者简介：杨金福(1977－)，男，博士生，主研方向：模式识别，机器学习；吴福朝，研究员；罗阿理，副研究员；赵永恒，研究员
收稿日期：2006-04-10 **E-mail：**yangjf@nlpr.ia.ac.cn

系核，最后利用分类器 4 将恒星分为 O、B、A、F、G、K 和 M7 种光谱型，用分类器 5 将活动星系核分为类星体、SEYFERT1、SEYFERT2 和 LINER。

在分类器结构图中，把前两级（分类器 1、2 和 3）称为粗分类过程，后一级（分类器 4 和 5）称为细分类过程。粗分类的算法是采用基于覆盖技术的分类算法，而细分类中分类器 4 采用小波变换的方法，分类器 5 则采用 K 近邻方法。

根据分类结果，利用基于均值漂移谱线提取和红移测量算法来计算光谱的红移值。整个系统的处理流程如图 2 所示。(1)从磁盘读取 fits 格式一维光谱数据，经过预处理后调用粗分类算法模块对其进行分类；(2)判断结果，如果是恒星或者是活动星系核（AGN）则调用细分类模块进行进一步的细分；(3)调用谱线提取和求红移模块计算光谱的红移值；(4)保存结果并返回。

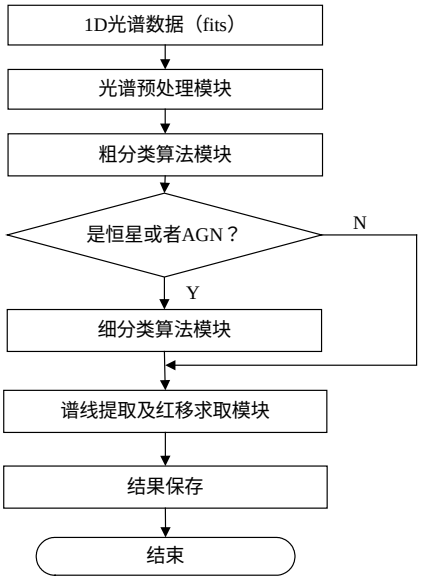

图 2 处理流程

2 核心算法

2.1 基于覆盖技术的分类算法

覆盖技术的主要思想是利用覆盖规则对训练样本进行训练，得到每一类的支撑向量（代表点），分类时我们只需考虑待识别样本与每一类支撑向量之间的距离即可。这里只简单介绍基于类内最大距离的覆盖技术，详细介绍见文献[2,3]。

设 A 和 B 分别为两类不同的样本，基于类内最大距离算法来求解支撑向量的过程如下：

初始，置 $k=0$；$S^A=\phi$，$S^B=\phi$；$A_0=A$，$B_0=B$；其中 S^A，S^B 分别是 A 和 B 两类的支撑向量集。

(1)计算集合 A_k 的直径 ρ_k，并令 $d(X_k,X_k^{'})=\rho_k$，$X_k,X_k^{'}\in A_k$；

(2)计算 $d_k=d(X_k,B),d_k^{'}=d(X_k^{'},B)$；令：

$\overline{S_k}=\{X\mid d(X,X_k)<d_k/2\}$

$\overline{S_k^{'}}=\{X\mid d(X,X_k^{'})<d_k^{'}/2\}$

$S_k=\overline{S_k}\cup\overline{S_k^{'}}$，$A_{k+1}=A_k-S_k$，$S^A\leftarrow S^A\cup S_k$；

(3)若 $A_{k+1}=\phi$，输出 S^A，终止算法；若 $A_{k+1}\neq\phi$，则：

1)若 $A_{k+1}\neq\{X\}$(即不为单点集)，则置 $k\leftarrow k+1$，转步骤(1)；

2)若 $A_{k+1}=\{X\}$(即为单点集)，则令 $X_{k+1}=X$，计算：$d_{k+1}=d(X_{k+1},B)$，并令：

$S_{k+1}=\{X\mid d(X,X_{k+1})<d_{k+1}/2\}$

$S^A\leftarrow S^A\cup S_{k+1}$；

输出 A 的支撑向量集 S^A，终止算法。

类似地可以计算 B 的支撑向量集 S^B。

2.2 基于小波变换的恒星分类算法

小波分析是 20 世纪 80 年代中期发展起来的一门新兴学科，它在时-频域同时具有良好的局部化性能，因此已经广泛地被用于信号处理领域。基于小波变换的恒星分类算法过程如下：

(1)对恒星光谱进行 3 阶小波分解，取第 2 和第 3 阶的系数进行重构，并归一化，得到谱线特征；

(2)利用第(1)步得到的谱线特征与恒星谱线特征模板求相关；

(3)设置阈值，并根据求得的相关值进行类别判断。

2.3 基于均值漂移的红移测量算法

均值漂移是模式识别中的一种经典方法，它的主要作用是在特征空间求取模式点，也即局部密度最大点[5]。

首先利用均值漂移的方法来提取特征谱线，然后再通过模板匹配来进行谱线证认并确定红移值。利用均值漂移总是指向局部密度最大点这一性质，通过使用均值漂移过程迭代逼近连续谱，以达到光谱去噪的目的。基本过程如下：

(1)归一化光谱后，采用均值漂移去噪得到谱线光谱；

(2)对谱线光谱设置局部阈值提取出特征谱线，得到一个特征波长序列，用 $\{\lambda_i,i=1,2,\dots,N\}$ 表示。

(3)跟据红移公式 $\lambda=(1+z)\lambda'$，其中 λ 为观测波长，λ' 为静止波长，寻找特征波长序列 $\{\lambda_i,i=1,2,\dots,N\}$ 与静止波长序列 $\{\lambda_i',i=1,2,\dots,M\}$ 的匹配即可得到红移值。

3 系统测试

LAMOST 投入运行以后，每天能产生万余条光谱，所以我们的系统需要保证能够在 24h 内能处理所有产出的光谱数据。为了使软件获得较高的运算效率，我们在系统实现时，核心算法是采用 ANSI C/C++语言编写，用户界面则是采用 QT 语言编写，软件系统运行环境是 Linux 操作系统。图 3 是软件系统的部分界面情况。

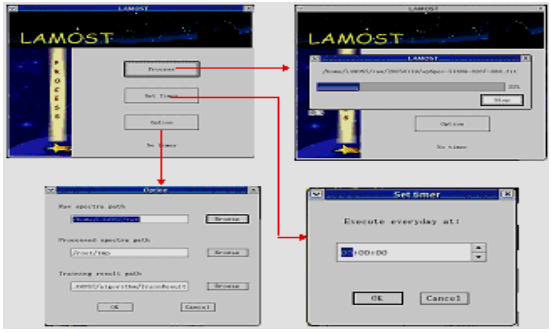

图 3 系统处理界面

由于现在 LAMOST 还没有产出真实光谱数据，因此本文系统测试所采用的数据是 SDSS 的 DR3 数据，数据量约 50 万条光谱。测试机器的硬件配置：CPU P4 2.8GHz，DDR 512MB，80G HDD。软件环境是 Linux Fedora Core3 操作系统，测试时读取的光谱数据是通过局域网访问文件系统服务器的数据。

SDSS 数据的光谱分类结果为星系、类型体、恒星、晚型恒星。而我们的系统分类更细，分为正常星系、O 型恒星、B

型恒星、A 型恒星、F 型恒星、G 型恒星、K 型恒星、M 型恒星、星暴星系、类星体、SEYFERT1、SEYFERT2、LINER。为了能使我们的分类结果与 SDSS 结果比较，需要将分类结果对应起来：SDSS 的星系包括正常星系、星暴星系、SEYFERT1、SEYFERT2 和 LINER；SDSS 的恒星和晚型恒星对应 O 型恒星、B 型恒星、A 型恒星、F 型恒星、G 型恒星、K 型恒星、M 型恒星。与 SDSS 比较的测试结果如表 1 所示。LAMOST 系统详细分类结果如表 2 所示。

表 1 与 SDSS 比较的测试结果

	星系	类星体	恒星	红移误差 0.01
SDSS	374 802	50 963	71 111	--
LAMOST 分类正确数	352 106	47 552	60 071	--
LAMOST 正确率	94.37%	83.20%	90.02%	86.91%

表 2 LAMOST 分类的测试结果

类型	正常星系	星暴星系	类星体	SEYFERT1	SEYFERT2	LINER	O 型恒星
数目	212 331	73 873	57 154	15 775	56 099	12 873	416
类型	B 型恒星	A 型恒星	F 型恒星	G 型恒星	K 型恒星	M 型恒星	
数目	2 616	9 530	15 118	11 218	4 078	23 754	

可以看出，我们实现的 LAMOST 光谱数据自动处理软件系统具有较高的分类正确率和红移测量正确率，红移误差为 0.01。测试 SDSS DR3 的约 50 万条数据所耗时间为 2 577min。这样的处理速度对每天处理 2 万条 LAMOST 数据的任务是完全可以胜任的。

4 结论

本文介绍了基于覆盖技术和小波变换的天体光谱分类方法，以及基于均值漂移的红移测量算法。在此基础上我们开发了 LAMOST 光谱数据自动处理软件系统，该系统的核心算法是基于统计学习的方法，因此更适用于 LAMOST 未定标的光谱数据。通过对 SDSS DR3 的光谱数据（约 50 万）测试，结果表明该系统具有较快的运算速度，且能获得较高的分类正确率和红移测量精度，可以满足国家重大科学工程 LAMOST 的实际要求。该系统的研究开发完成对我国天文学的研究工作起到积极的推动作用。

参考文献

1 SDSS. Spectroscopic Redshift and Type Determination[Z]. http://www.sdss.org/dr4/algorithms/redshift_type.html#finalz.

2 Jinfu Yang, Wu Fuchao, Luo Ali, et al. Automated Classification for Celestial Spectra Based on Cover Algorithm[J]. Pattern Recognition and Artificial Intelligence, 2006, 19(3): 368-374.

3 Zhang L, Zhang B. A Geometrical Representation of McCulloch-pitts Neural Model and Its Applications[J]. IEEE Transactions on Neural Networks, 1999, 10(4): 925-929.

4 Liu Zhongtian, Zhao Ruizhen, Zhao Yongheng, et al. A Wavelet Transform Based Method for the Automatic Detection of Late-type Stars[J]. Spectroscopy and Spectral Analysis, 2005, 25(7):1158-1161.

5 Comaniciu D, Meer P. Mean Shift: A Robust Approach Toward Feature Space Anaysis[J]. IEEE Trans. on PAMI, 2002, 24(5): 603-619.

（上接第 150 页）

间结果，这样会使得 PTK 的计算性能得到 4~5 倍的改善，MIC 的计算性能利用中间结果也会有 1.5~2 倍的改善。所以大量 PTK 的计算虽然加重了 CPU 的负担，但影响并不严重。

为了验证该方法的有效性，我们通过 a、b、c 3 个进程对上述过程进行了模拟，其中 a 进程完成 STA 的功能，b 进程完成 AP 的功能，c 进程完成攻击者的功能。在运行中，a、b 进程进行四次握手过程，c 进程向 a 进程发送大量带有不同随机数的 Message1 和 Message3，由于 c 进程不知道 STA 和 AP 所共享的 PMK，因此无法通过 MIC 校验，最终无法和 STA 进行数据通信，而 AP 向 STA 发送 Message3 后通过 MIC 校验，双方装载本次会话临时密钥 PTK 后可以进行数据通信。采用了本文所提出的改进方法，c 进程无法成功进行 DoS 攻击，但是 STA 端的运算量有了相应的提高，CPU 性能虽然有所下降，但对网络性能未造成严重的影响。

本文所提出的改进方法是在增加 CPU 负担和存储器资源消耗之间的一个折中方案。如果网络环境中接收到的 Message 绝大多数都是期望的、合法的，那么请求者就可以存储接收到的 Anonce 和生成的 PTK，用其来对收到的 Message3 中的 MIC 进行校验。本文所提出的基于同一个随机数 Snonce 重新计算 PTK 的方法适用于遭受恶意攻击的情况下，所以这个方法能够较好地解决四次握手过程中存在的安全问题。

4 结束语

本文对 IEEE802.11i 协议中四次握手过程进行了详尽的分析，针对其中存在的安全隐患和可能受到的攻击提出了一种 STA 重复使用同一个 Snonce 计算 PTK 的改进方法。通过该方法可以避免存储器资源被攻击者耗尽所造成的 DoS 攻击，同时利用一些已经计算得到的中间结果减轻了 CPU 的计算负担，使 CPU 的性能只是受到有限的影响。通过这个改进方法可以很好地解决四次握手过程中存在的安全弱点，避免遭受严重的 DoS 攻击，以较小的代价获得更高的无线网络安全性。

参考文献

1 Peikari C, Fogie S. 无线网络安全[M]. 周　靖, 译. 北京: 电子工业出版社, 2004.

2 Stallings W. 无线通信与网络[M]. 北京: 清华大学出版社, 2003.

3 刘乃安. 无线局域网(WLAN)——原理、技术与应用[M]. 西安: 西安电子科技大学出版社, 2004.

4 He Changhua, Mitchell J C. Security Analysis and Improvements for IEEE 802.11i[EB/OL]. 2004. http://www.isoc.org/isoc/conferences/ndss/05/proceedings/papers/NDSS05-1107.pdf.

5 He Changhua, Mitchell J C. 1 Message Attack on the 4-Way Handshake[EB/OL]. 2004. http://theory.stanford.edu/~changhua/11-04-0497-02-000i-1-message-attack-on-4-way-handshake.doc.

6 Baghaei N. IEEE 802.11 Wireless LAN Security Performance Using Multiple Clients[C]//Proc. of the 12th IEEE International Conference on Network, 2004: 299-303.