杨海峰◎著

天体光谱数据挖掘与分析

DATA MINING AND ANALYSIS ON CELESTIAL SPECTRA

中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

天体光谱数据挖掘与分析

杨海峰 著

電子工業出版社
Publishing House of Electronics Industry
北京 · BEIJING

内 容 简 介

随着 LAMOST 正式巡天的实施，已成功获取了 600 万条天体光谱及星表，并每天以海量的数字增长着，这给长期传统的人工分析、人眼证认等任务带来了巨大挑战。本书以河外星系和恒星光谱为研究背景，针对天文学研究中稀有天体的特征分析以及天体光谱的分类等任务，将新兴的数据挖掘技术应用到天体光谱规律的发现和研究中，并从天文物理学角度对挖掘结果进一步分析。主要包括特殊、稀有天体的挖掘与分析，光谱分类及后处理方法两个方面的内容。

本书可供从事数据挖掘、机器学习以及天文信息学等相关专业的科研人员参考，也可以作为高等院校计算机、天文学专业的高年级本科生与研究生的学习参考书。

图书在版编目（CIP）数据

天体光谱数据挖掘与分析/杨海峰著. —北京：电子工业出版社，2016.12
ISBN 978-7-121-30768-3

Ⅰ. ①天… Ⅱ. ①杨… Ⅲ. ①天体－光谱学－数据采集－研究 Ⅳ. ①P141.5 ②TP274

中国版本图书馆 CIP 数据核字（2016）第 322749 号

责任编辑：徐蔷薇　　文字编辑：米俊萍
印　　刷：三河市华成印务有限公司
装　　订：三河市华成印务有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　　邮编：100036
开　　本：720×1000　1/16　印张：11　字数：160 千字
版　　次：2016 年 12 月第 1 版
印　　次：2016 年 12 月第 1 次印刷
定　　价：49.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至zlts@phei.com.cn，盗版侵权举报请发邮件至dbqq@phei.com.cn。

本书咨询联系方式：xuqw@phei.com.cn。

前　言

仰望璀璨的星空，辽阔而深邃，自由而宁静，吸引着人们苦苦追寻与不断探索的向往。LAMOST 是一架横卧南北方向的中星仪式反射施密特望远镜，在 5 度视场、直径为 1.75 米的焦面上放置 4000 根光纤，可以同时获得 4000 个天体的光谱，是当前世界上光谱获取率最高的望远镜。随着 LAMOST 正式巡天的实施，已成功获取了 600 万条天体光谱及星表，并每天以海量的数字增长着，这给长期传统的人工分析、人眼证认等任务带来了巨大挑战。而数据挖掘，作为一门新兴的学科分支，涉及人工智能、机器学习、模式识别等多个学科领域，主要任务是从大量的原始数据中提取潜在的、人们感兴趣的知识，其已被广泛地应用于科学、工程、商业等领域。将数据挖掘技术应用于海量的天体光谱数据中，获取潜在的、有意义的天体规律及性质，对更有效地使用巡天数据、进一步深入天文学理论研究都具有比较重要的应用价值。

近年来作者一直从事数据挖掘应用与天体光谱分析交叉领域的研究，在深入了解光谱分析任务、分析当前数据急剧增长特点的基础上，结合计算机技术优势，开展了一系列的研究工作，本书是近年来相关科研成果的总结。全书除绪论主要介绍天体光谱数据的主要特征以及数据挖掘技术的基本理论外，主要内容分为两篇共 6 章，具体章节编排如下。

（一）特殊、稀有天体的挖掘及分析（包括第 2～4 章）。第 2 章针

对星系光谱中呈现的双红移系统，提出了一种基于模糊识别的光谱特征线识别方法，并采用 SDSS DR9 和 LAMOST DR2 的星系光谱数据，系统地搜寻了具有双红移系统的星系光谱，并对其结果进行了光谱及图像分类、特例分析、前景星系消光测量等方面的讨论。第 3 章针对碳星光谱中存在的模板较少从而导致从海量数据中搜寻比较困难的问题，提出了一种新的高效的 PU 学习方法，并选择 SDSS DR10 中的 10 万余条光谱实验验证了该方法的搜寻质量和效率。第 4 章针对 LAMOST 河外星系光谱分辨率及信噪比等特征，修正了[OII]、Hδ、Hα特征线边界，通过测量其等值宽度并按照经典（由 Goto.提出）的判定依据，从 LAMOST DR2 中系统搜寻了 E+A 星系，并对其结果进行了红移分布、空间分布、星等分布特征、图像特征及星族特征等方面的讨论。

（二）光谱分类及后处理方法研究（包括第 5～7 章）。第 5 章针对巡天数据分析中最基本的光谱型分类问题，提出了一种基于贝叶斯支持微量机的光谱自动分类方法，选择 SDSS DR10 的 M 型恒星光谱，实验验证了该方法在光谱子型的分类上具有较高的准确率及效率，同时对预处理过程中噪声、归一化方法、特征提取方法对分类结果的影响进行了讨论。第 6 章针对恒星光谱分类任务，提出了一种基于分类模式树的恒星光谱分类规则挖掘方法。采用 SDSS 恒星光谱作为实验数据，验证了该方法的正确性，而且具有较高的分类正确率。第 7 章针对采用数据挖掘方法提取的光谱分类规则中存在的冗余严重影响分类效率和质量的问题，提出了基于谓词逻辑、集合运算的两种分类规则后处理方法，从而减小了分类器的大小。采用 SDSS 恒星光谱数据，实验验证了这两种方法在不降低分类准确率的前提下，可以有效提高分类效率。

本书的完成得到了太原科技大学人工智能实验室、计算机科学与技术学院、中科院国家天文台各位老师的大力支持，特别是张继福教授、罗阿理研究员为本书提出了许多宝贵的建议，在此一并致以诚挚的谢意。

本书所涉及的部分研究工作得到了国家自然科学基金项目（项目编号：61272263，61572343）、山西省科技攻关项目（项目编号：2015031009）和太原科技大学博士启动基金（项目编号：20162007）的资助，在此向相关机构表示深深的感谢。

由于作者水平有限，书中难免有不妥之处，欢迎各位专家和广大读者批评指正。

编　者

2016 年 11 月

目　录

第一篇　特殊、稀有天体的挖掘与分析

第二篇　光谱分类及后处理方法

第1章 绪 论

1.1 天体光谱

仰望璀璨的星空，辽阔而深邃，自由而宁静，吸引着人们苦苦追寻与不断探索的向往。自有人类文明以来，天文学就有着非常重要的地位，从盘古开天地、女娲补天等具有神话故事的宇宙演化观，到《步天歌》“北斗之宿七星明……”对星象变化的认识，再到屈原《天问》中对未知的星空提出的系列问题，中国古天文对世界天文学的发展起着重要的作用。那么，宇宙是如何形成和演化的？和银河系类似的星系在宇宙演化过程中是什么角色？它们又是怎么形成与演化的？人类是通过何种方式去认识宇宙的奥秘的？……伴随着这些有趣的问题，天文学作为一种特殊的基础学科，得到了推动与发展。近年来，随着各大光谱巡天项目的陆续实施，观测得到的各种波长范围、各种分辨率、各种类型天体的光谱急剧增长，为天文学研究提供了充足的样本，天体光谱已成为人类认识宇宙重要的手段之一，图 1.1 为一类 F 型恒星光谱示例。

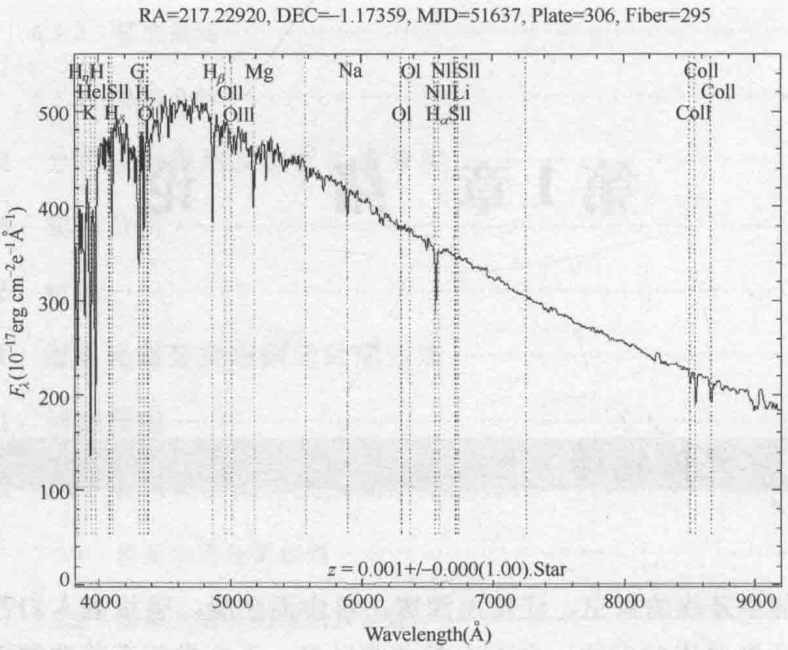

图 1.1　F 型恒星 1D 光谱

1.1.1 LAMOST 光谱巡天[1~3]

LAMOST（大天区面积多目标光纤光谱天文望远镜，也称为郭守敬望远镜）是一架横卧南北方向的中星仪式反射施密特望远镜，坐落在河北省承德市兴隆县的兴隆观测基地上（东经 7 小时 50 分，北纬 40 度 23 分，海拔 960 米），图 1.2 为 LAMOST 望远镜实景图。

LAMOST 由反射施密特改正板 MA（大小为 5.72 米 × 4.40 米，由 24 块对角线长 1.1 米，厚度为 25 毫米的六角形平面子镜组成）、球面主镜 MB（大小为 6.67 米 × 6.05 米，由 37 块对角线长为 1.1 米，

厚度为 75 毫米的六角形球面子镜组成）和焦面构成。球面主镜及焦面固定在地基上，反射施密特改正板作为定天镜跟踪天体的运动，望远镜在天体经过中天前后时进行观测。天体的光经 MA 反射到 MB，再经 MB 反射后成像在焦面上。焦面上放置的光纤，将天体的光分别传输到光谱仪的狭缝上，然后通过光谱仪后的 CCD 探测器同时获得大量天体的光谱。LAMOST 所应用的薄镜面主动光学加拼接镜面主动光学技术，在曝光 1.5 小时内可以观测到暗达 20.5 等的天体，使其成为大口径兼大视场光学望远镜的世界之最。同时，采用并行可控的光纤定位技术，在 5 度视场、直径为 1.75 米的焦面上放置 4000 根光纤，可以同时获得 4000 个天体的光谱，成为当今世界上光谱获取率最高的望远镜。

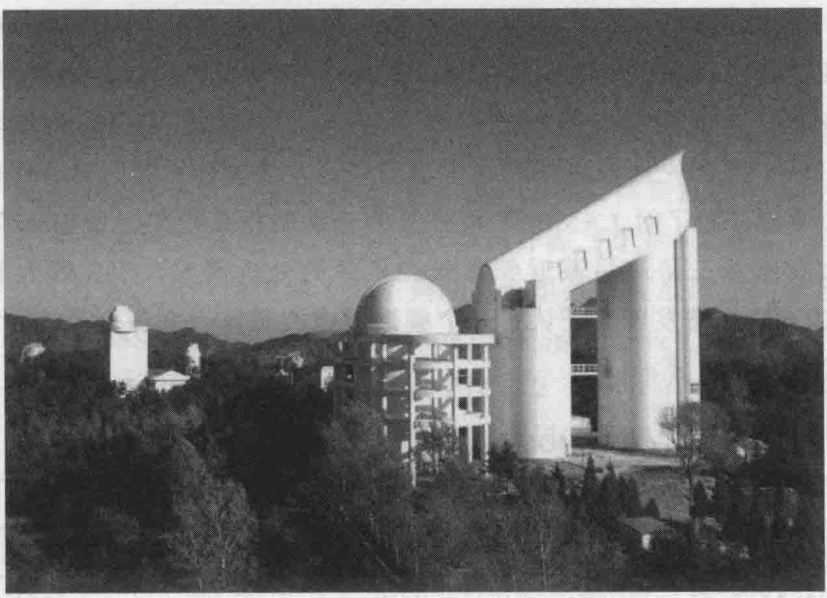

图 1.2　郭守敬望远镜实景图

LAMOST 巡天两个重要组成部分是 LEGUE（LAMOST Experiment for Galactic Understanding and Exploration survey）和 LEGAS（LAMOST ExtraGAlactic Survey），LAMOST 河外巡天又包括河外星系巡天及类星体巡天两部分，而河外星系的选源主要有以下几部分：北银冠天区（主要是 SDSS legacy 中由于光纤碰撞导致错过的那些天体，星等 r< 17.75），南银冠天区（星等 r < 18，对于一些蓝星系 r < 18.8，目前观测天区范围为 45° < ra < 60°，0.5° <δ<9.5°），与红外巡天（如 IRAS，WISE，HERSCHEL）交叉的亮红外星系，以及 LCSSPA（位于南银冠的两个 20 deg^2 完备小天区内）。

LAMOST 自 2008 年获得首条光谱以来，经过两年的任务观测及为期一年的先导巡天，于 2012 年 9 月开始正式巡天，截止到 2016 年 1 月，LAMOST DR3 的观测任务已基本结束。表 1.1 列出了 LAMOST 河外源及 pipeline 分类为“Unknown”的统计情况，LAMOST 河外星系及类星体的观测光谱数已超过 8 万条，同时被 pipeline 分类为“Unknown”的光谱中也不乏有价值的星系及类星体光谱。同时越来越多的学者开始了对 LAMOST 河外天体光谱的相关研究，如 Huo 等人对仙女座及三角座星系近邻背景类星体的分析研究，Shi 等人利用 LAMOST 光谱发现并证认了一个频谱射电类星体，并识别了 20 个双峰发射线星系等。

表 1.1　LAMOST 巡天前三年观测天体光谱数统计

	Pilot	DR1	DR2	DR3
STAR	807 575	2 317 365	3 779 674	5 268 687
GALAXY	2 723	9 359	37 401	54 022
QSO	621	4 396	9 129	13 353
Unknown	65 496	177 859	305 142	372 078

1.1.2 SDSS 光谱巡天

SDSS（Sloan Digital Sky Survey，斯隆数字化巡天）开始于 2000 年，是最有影响力的巡天项目之一，现已进入第四期巡天任务，旨在获取海量测光及光谱数据，以研究宇宙大尺度结构、星系的形成与演化等天体物理学领域的重大前沿课题。SDSS 项目使用的是位于美国新墨西哥州阿帕奇波因特天文台（Apache Point Observatory）的 2.5 米望远镜，该望远镜配备有一台 120 兆像素的成像用相机（一次覆盖 1.5 平方度，用于测光巡天）和一对连接了 640 根光纤的光谱仪（用于光谱巡天）。每个光谱观测天区（spectral plate）指的是一个打了 640 个孔（对应观测目标及定标星）的焦面金属板，覆盖约 7 平方度。光谱的波长覆盖范围是 3800～9200Å，分辨率$\lambda/\Delta\lambda$在 1850～2200。

SDSS 巡天的重要目标是星系及类星体红移巡天，前两期的巡天任务（Legacy 巡天）对北银冠 7500 平方度天区以及南银冠三个 strip 超过 760 平方度的区域进行了观测，获得河外源（星系+类星体）光谱超过 100 万条；第三期实施了旨在刻画亮红星系及类星体空间分布的巡天项目 BOSS（Baryon Oscillation Spectroscopic Survey），获得了 150 万条红移 $z < 0.7$ 的亮星系及 16 万条红移 $2.2 < z < 3$ 的类星体；第四期河外源的观测分别向更深、更细两个方向进行了扩展，设计了两个子巡天项目：eBOSS（the Extended Baryon Oscillation Spectroscopic Survey，见图 1.3）和 MaNGA（Mapping Nearby Galaxies at APO，见图 1.4），截止到 DR12 数据发布，已获得星系光谱数据总量达 2 599 191 个。SDSS 巡天的光谱分辨率、波长覆盖范围等特征与 LAMOST 很相似，数据及其相应的处理技术对我们的相关研究具有重要的借鉴作用。

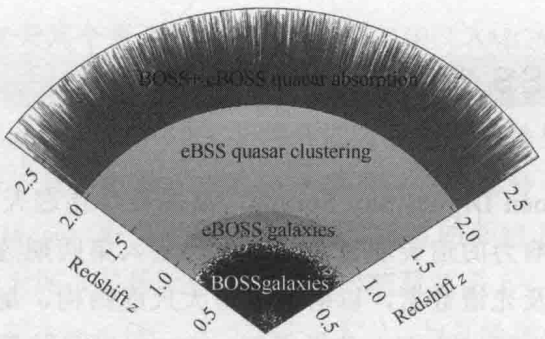

图 1.3　eBOSS 河外源观测深度示意图

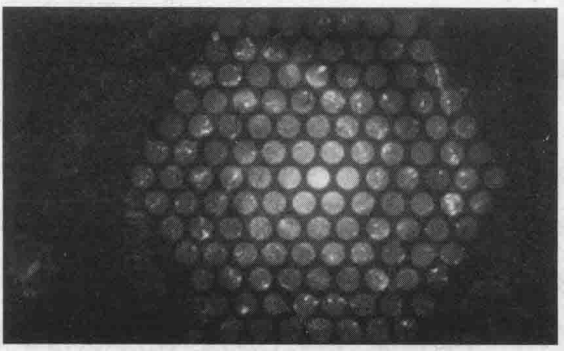

图 1.4　MaNGA IFU 略图

1.1.3　光谱分析

由于每种原子都有自己的特征谱线，因此可以根据光谱来鉴别物质和确定它的化学组成，这种方法叫做光谱分析。光谱分析在科学技术中有广泛的应用，历史上，通过光谱分析还帮助人们发现了很多新元素。19 世纪初，在研究太阳光谱（见图 1.5）时，发现它的连续光谱中有许多暗线。最初不知道这些暗线是怎样形成的，后来人们了解了吸收光谱的成因，才知道这是太阳内部发出的强光经过温度比较低