网格聚类分析天文光谱数据

陈淑鑫[1,2] 孙伟民[2] 王丽丽[3]
(齐齐哈尔大学机电工程学院 齐齐哈尔 161006)[1]
(哈尔滨工程大学理学院纤维集成光学教育部重点实验室 哈尔滨 150001)[2]
(德州学院信息管理学院 德州 253023)[3]

摘 要 应用互联网+融合信息技术,天文大数据研究实现了海量观测数据及次生数据的高效存储、检索、数据分析及信息挖掘。现结合我国自主知识产权的大科学工程LAMOST望远镜巡天第四期(DR4)发布的经定标后的光谱数据,运用R语言中RFITSIO软件包读写光谱专用文件FITS格式,读取LAMOST发布的恒星天文数据,结合统计学和数据挖掘方法设计了有监督的网格聚类验证方案,处理并识别光谱数据,经降维提取光谱特征,归一化连续谱,保留吸收谱线特征,再划分网格聚类波长定标中心,利用相似度量函数来描述识别观测光谱数据。

关键词 聚类分析,FITSio,光谱数据,LAMOST
中图法分类号 TP319 **文献标识码** A

Analysis of Astronomical Spectral Data Based on Grid Clustering

CHEN Shu-xin[1,2] SUN Wei-min[2] WANG Li-li[3]
(College of Mechanical and Electrical Engineering of Qiqihar University, Qiqihar 161006, China)[1]
(Key Lab of In-fiber Integrated Optics Ministry Education of China, Science College, Harbin Engineering University, Harbin 150001, China)[2]
(School of Information Management, Dezhou University, Dezhou 253023, China)[3]

Abstract The efficient analysis and processing of astronomical data is supported by Internet plus integration information technology. The massive observation data and secondary data have achieved efficient storage, retrieval, data analysis and information mining. Combined with the fourth (DR4) released star spectral flowed calibration data of the large scientific engineering LAMOST telescope survey which has owned China's independent intellectual property rights, taking the astronomical data released by LAMOST as an example, RFITSIO software package of R language programming platform is used to read and write the spectrum documents of special FITS format. With the statistical data and data mining method, the verification scheme of supervised grid clustering was designed, by processing and identifying the spectral data. The spectral characteristics were extracted by dimensionality reduction. The characteristics of the absorption spectra were retained by normalized continuous spectrum, then the center grid of clustering wave length scale was divided, and the similarity measure function was used to describe the observed spectrum.

Keywords Cluster analysis, Flexible image transport system input output, Spectrum data, Large sky area multi-object fiber spectroscopy telescope

1 引言

随着互联网+数据科学时代天文大数据信息的飞速增容,天文学在宇宙中寻求特殊的、未知的天体是人类探索宇宙奥妙所追求的目标,因此需要从飞速扩容的天文大数据信息数据库中,提取未知、潜在且具有研究价值的信息模式。当前大数据研究应用颇有价值的领域融合了数据库、机器学习、统计学等诸多理论和技术,相关天文大数据科学研究在大视场、大规模光学光谱观测中诠释了宇宙中星系的形成及演化等问题。天文光谱分析的相关重要课题相继开展,研究者们收集天体发射到地球的辐射数据信息,使得现拥有的恒星、星系及类星体等光谱数据量急速扩增,其通过分析挖掘光谱数据信息的相关性得出天体位置、宇宙分布等。大部分研究成功地实现了光谱的恒星大气物理参数的自动测量。我国自主研发的大型多目标光纤光谱天文望远镜LAMOST[1]郭守敬望远镜,即大天区面积多目标光纤光谱望远镜(Large Sky Area Multi-Object Fiber Spectroscopy Telescope, LAMOST),它是目前世界上最大口径、最多观测目标、最广视场范围、最高天体光谱获取率的光纤光谱望远镜,现居国际领先的科学技术,所获取的光谱大数据信息达10^7数据量级。LAMOST拥有完整的自动化观测、数据处理和存储的软件系统[2],已发布的DR4巡天采集的数据增至760多万条恒星光谱,其中大部分是主序星的光谱。本文基于光谱恒星大气物理参数自动测量方法的研究,表示物理特性、化学成分以及运动和发展的规

本文受国家自然科学基金项目(U1631239),黑龙江省自然科学基金资助项目(F2015203),黑龙江省教育厅基本科研业务专项(135109219),齐齐哈尔大学教育科学研究项目(2016072)资助。
陈淑鑫(1978—),女,博士生,副教授,主要研究方向为数据处理,E-mail:shuxinfriend@126.com。

律。通过网格的聚类算法分析获取数据知识信息，着力解决对数据流的聚类问题，以在 LAMOST 数据发布恒星光谱的高分辨率光谱参数为基础，利用 R 语言的 RFITSIO 软件包对光谱大数据进行图形化分析，从数据挖掘算法中提取特征信息和知识，进而为更深度学习海量数据提供理论方法，从中发现未知天体和新天文现象。

2 天文光谱谱线

天文大数据时代里计算科学蓬勃发展，研究者依据空间属性的存在、天体空间位置和距离的概念以及相邻天体之间存在的一定的相互作用，表现了天文数据的复杂性与非线性关系属性。恒星光谱的通常形态为一个连续谱（黑体辐射谱）上叠加多种吸收线。由于不同元素的原子在电子跃迁时吸收的光子频率的差异，吸收线的强度可用来表示其对应元素在该恒星上的丰度。天文光谱谱线是在观测恒星光谱分析中反映恒星组成的重要指标和中间参数。

2.1 LAMOST 高维光谱数据

大数据时代海量天文光谱的数据挖掘包含信息提取、数据自动处理等关键技术，光谱间的流量相似度分析是天文学光谱数据挖掘的重要数据处理环节。LAMOST 巡天采集数据恒星光谱大部分是主序星的光谱。在第 4 期数据发布中（LAMOST DR4）共发布 FGK 恒星光谱的恒星参数星表由 650 多万条（DR3）增至 760 万，发布的参数以高分辨率光谱为基础，且以构建光谱库 Elodie 为基准[2]。上述数据所获取的光谱噪声较大，同时存在流量定标误差，造成光谱畸变，增加了恒星大气物理参数的测量难度。

2.2 天文数据 FITS 格式

20 世纪 80 年代普适图像传输系统 FITS（Flexible Image Transport System，原意为灵活图像传输系统）[3]格式被国际天文联合会（International Astronomical Union，IAU）正式公布为国际标准，成为天文学领域应用最广泛的数据格式，其可灵活地定义描述数据的参数，被保存在世界各地的数据中心，每条行记录表示图像的某一信息。FITS 文件头由多个长为 80 的行的记录数据组成，包含 2880 字节，逻辑记录为 36 个行记录数据[4]。左对齐的关键字（keyword）是长为 8 个字符的字符串，由大写英文字母、数字、下划线“_”或连字符“-”组成，中间不能有空格，若长度不满 8 字符则在末尾用空格填充。其独立于观测的硬件设备所获取的每条 FITS 数据文件，能够描述数据定义和数据本身编码，用于天文数据的传输、分析和存储。

2.2.1 R 软件数据挖掘

大数据时代亟待处理和分析的数据日益增长，1993 年开发的开源工具 R 语言程序是用于统计分析和数据处理的强大工具。用户通过免费的 Comprehensive R Archive Network（CRAN）公共库共享使用新的统计包。采纳优化内存方法提升 R 程序的性能技术，利用外部数据处理系统的并行计算能力完成数据分类、数据挖掘、分析数据，分析发现有价值的规律和概念；快捷地设置监督聚类中心，将实测数据按照所测量的参数进行聚类，从而对与每个类中心参数相近的实测数据与该网格点对应的光谱进行分析来判断其差异。

2.2.2 LAMOST 光谱数据 FITS 格式

LAMOST 光谱数据中已发布 FITS 文件命名格式为“spec-MMMMM-YYYY_spXX-FFF.fits”，其中“MMMMM”代表当地修正的儒略日，“YYYY”代表计划标识的字符串，“XX”代表光谱仪的数字编号，“FFF”代表所采集到光谱的光纤编号，扩展名为“.fits”。此外，LAMOST 还设计以 HMS（时分秒）为单位的 RA（赤经）值为关键字“HHMMSS.ss”，以 DMS（度分秒）为单位的 Dec（赤纬）值为关键字“DDMMSS.ss”。主要的数据数组有五行数据和一个 NAXIS1（FITS 数组的维数）列。五行数据分别是流量、倒方差、波长、Andmask 和 Ormask。注明其中“倒方差”的不确定性（$1/\sigma^2$）用来估计每个像素的信噪比（流量 *（倒方差）$^{0.5}$），以及每个像素 Andmask 和 Ormask 屏蔽的 6 位标志位质量情况。由于 LAMOST 是多次曝光合并的光谱，因此数据中的 Andmask 是指多次曝光各个像素 mask 的并集，而 Ormask 是指多次曝光各个像素 mask 的交集，这两个像素屏蔽位如表 1 所列。

表 1 FITS 数据的像素屏蔽位说明

Bit	Keyword	备注
1	BADCCD	CCD 中坏的像素
2	BADPROFILE	坏的轮廓提取
3	NOSKY no	此波长的天光背景
4	BRIGHTSKY	太高的空间级别
5	BADCENTER	CCD 的光纤轨迹
6	NODATA	坏数据

从 DR1 中重新组合主要的 FITS 头文件的关键字，在 FITS 基本头文件的单元可选择符合扩展以及其他可选择的特殊记录。

2.2.3 RFITSIO 软件包应用

美国马里兰大学 Andrew Harris 教授用标准 R 语言程序编写出 RFITSIO 软件包[5]，便于读、写国内外天文学界普适图像传输系统 FITS 文件所有类型的扩展文件[4]（包括 Bintable二进制列表、ASCII 列表 Table 以及图像扩展文件）。采用的 R 语言包是从相应的 CRAN 镜像站点下载的并将其放入库中，加载 FITSio 包读取实验编程及程序运行，载入美国的天文学家安德鲁·哈里斯研发的 FITSio_2.0_0.zip 软件包中。从这个包所包含的功能中来读取单一 FITS 头文件数据单元（HDUs）的图像和扩展的二进制表，以及一个写入图像文件中 readFITS 自动识别图像（多维数组）和扩展二进制表，返回数据、头文件和扩展信息的列表。FITS 中 readFrame 函数能从 R 语言数据框架里返回单一的二进制表头文件数据单元。这两个函数均能选出较大的第 n 个头文件数据单元。在 FITS 头文件中修改和编辑关键字的值所对应的功能，其中 newKwv 为头文件创建关键字＝值/注释行；addKwv 为 header＝value/comment 添加到标题；delKwv 从头中删除 keyword＝value/comment；modVal 修改 header 中 keyword＝value/comment 中的值；addComment 向标题添加 COMMENT 行；addHistory 将 History 行添加到标题。

3 网格聚类算法

由于天文数据的结构及数据背景意义的多样性，实验需要寻找数据间的相似度，设置合理的数据分类，进而发现数据中隐含知识的有用信息。对于大规模天文光谱数据库的高效网格聚类算法，将天文数据多分辨率数据结构作为处理数据模型和光谱数据的特点，在聚类过程中建立 3 个特征值为基础的立方块分布，每一维数据的单元数目影响聚类计算时间，

相关性分类缩短了聚类过程花费的时间，实现了更快的流量定标特征提取。为快速地对大样本 LAMOST 巡天大数据光谱完成挖掘分析[6]，高性能服务器在存储环境下结合海量数据本身的敏感性、时效性、空间相关性等特征，所选用的数据样本要完备，否则得到的规则推广性会很差。按照数据挖掘的方式和目的[7]可分为有监督学习和无监督学习。本文研究采用有监督学习完成分类。

3.1　有监督聚类分类

有监督的学习中分类算法包括两部分数据：训练数据和测试数据。将事先给出的若干类光谱数据作为训练集，其光谱的类别由理论模板给出，再利用训练集训练得出分类器，新的待分类光谱数据进入该分类器得到一个分类。该方法主要分析巡天大数据搜寻天体的已知类型光谱，执行训练集中的数据分类任务，根据训练样本的属性值提取出每类的准确描述或模型，然后将所有训练样本存储在服务器的模式空间中，倘若有新样本出现再进行泛化。

3.2　K-means 最近邻分类方法

基于网格的聚类算法较适用于有限空间范围内的大数据量、高密度的数据集，通过分析数据来获取信息知识，着力解决对任意形状的数据流聚类问题，从中发现特征量之间的内在结构并探索提取其内在的联系。天文大数据获取数据分布是用基于模板匹配的恒星光谱分类处理，最近邻聚类分组方法将模板库作为经过处理后的训练集，对于每条待测光谱，首先计算出该待测光谱与训练集（即模板库中的光谱）中每条光谱的距离，匹配相关距离最小的模板光谱即为该待测光谱的类型。依据密度来判断聚类检测孤立点，以损失少部分数据信息为代价来提高计算效率，具有较低的时间复杂度，其计算量只与网格中的单元数目有关，执行速度取决于网格分割的时间，并不是依赖于数据集的大小。

3.3　恒星光谱相似度数据分析

光谱的谱线是观测天文数据中的重要特征[7]，分析恒星光谱相似度度量之前，需要完成光谱预处理，对模板光谱和实测光谱的处理步骤为：连续谱归一化，流量归一化和发射谱线检测，最后再进行模板匹配。

3.3.1　处理连续谱归一化频率

获取到的观测光谱主要包含连续谱谱线和噪声，由于所采集数据标量不一致，同时存在观测光谱和理论光谱之间连续谱的差异，需将采样频率设为 1，即使频率的范围变得非常大，数据处理时仍旧很不方便。为此实现统一标准便于比较各个频率的分布情况，能有效防止数据溢出。本实验采用线性函数转换，如式(1)所示，归一化后的频率转换到[0，1]区间。

$$F_y=(F_x-F_{\mathrm{Min}})/(F_{\mathrm{Max}}-F_{\mathrm{Min}}) \tag{1}$$

其中，F_x 为归一化转换前的数据值，F_{Max} 为聚类样本的最大值，F_{Min} 为聚类样本的最小值，F_y 为归一化转换后的数据值。绘制的光谱图横轴为波长 λ，纵轴为 Flux 光强，经连续谱归一化处理，去掉连续谱的信息，只剩下谱线和噪声信息，能很好地减少强线以及宽线对拟合的影响，归一化连续谱拟合方法更有利于后续的谱线检测。

3.3.2　处理流量归一化

采用中国大天区面积多目标光纤光谱天文望远镜巡天项目实测光谱来构建模型，选取 LAMOST 中 M 型恒星光谱样本，径向速度设为所有光谱径向速度减少至零点，光谱移动到静止波长后，截取相同对数波段的光谱样本范围为 3400Å～9000Å，有 5600 个采样点数，参数的动态范围表面的有效温度 T_{eff}＜100K，重力加速度 logg＜0.3dex，金属丰度[Fe/H]＜0.15dex。按 3 个量的尺度划分光谱网格作为训练样本，对流量归一化的测试样本全部进行聚类。

3.3.3　模板匹配

建立模板训练光谱数据库，通过现有的天体物理测量方法进行精确测定或者由理论模型给出库中光谱的物理参量，完成特征提取的集合，将待测光谱与库中的每条模板光谱进行比较，将待测光谱最相似的模板光谱的参量作为待测光谱的参量，其相似性由某种相似度量函数来描述，如图 1 所示。

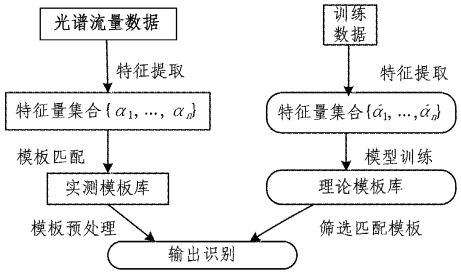

图 1　观测数据输出识别光谱的基本流程

4　处理光谱实验数据

天文学中高维数据挖掘主要采用直接和间接方法，直接对数据提取信息时则采用多种适合处理高维数据的算法。间接将高维数据进行线性变换后，再投影到低维空间，采用相应的高效挖掘算法。目前降维数据方法的关键是如何处理维数无限增大的问题，简化线性变换会掩盖数据原有的信息，为此探索适合的投影方向使数据呈现正态分布。

4.1　读取 FITS 光谱文件

本实验运行环境系齐齐哈尔大学现代教育技术中心云计算中心的超性能计算服务器，总计算能力约 7 万亿次每秒。实验采用 R 语言作为程序编写语言，从 http://dr2.lamost.org/下载 LAMOST 发布第二次巡天共享数据 dr.fits.gz，读取 fits 格式的任意数据文件，如图 2 所示的 spec-56647-M31010N33M1_sp09-107.fits 文件，从中选择具有恒星参数 M1 型星光谱及其参数，实现科学计算、数值计算等操作并绘制图表。

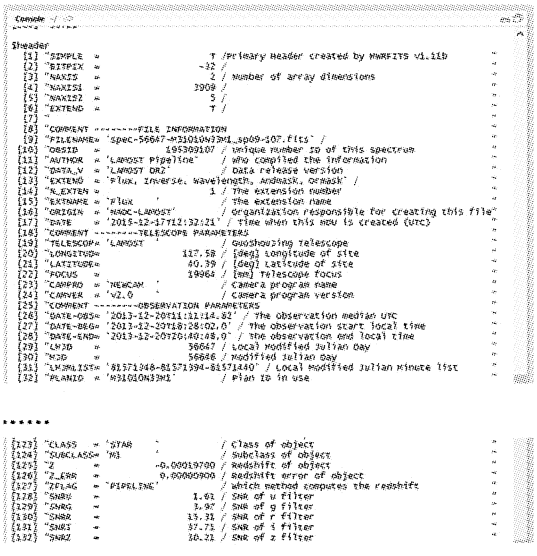

图 2　M1 型星 spec-56647-M31010N33M1_sp09-107.fits 文件信息

4.2 处理数据光谱

实验利用 R 语言动态提取出特征向量矩阵函数 M_STAR $ imDat，将处理数据存储成. csv 格式文件，读取对应参数信息数据列，选取图 2 巡天数据库中的 M1 型星光谱数据文件 spec-56647-M31010N33M1_sp09-107. fits，限定最大数据范围值后利用 plot()函数选取 type="s"参数，如图 3 所示，绘制流量光谱图。归一化后频率如图 4 所示。

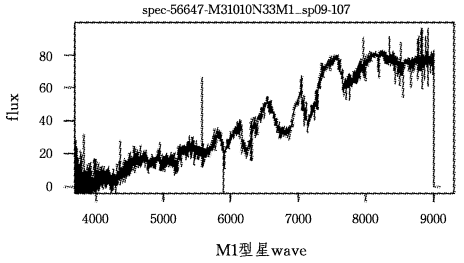

图 3 R 语言读取并绘制的 LAMOST 巡天数据 M1 型星流量光谱图

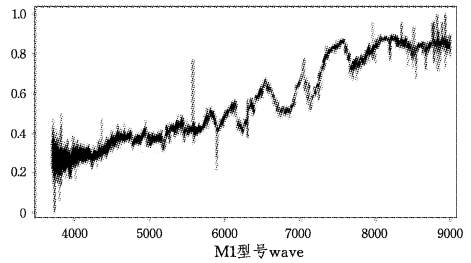

图 4 归一化数据 M1 型星流量光谱图

4.3 网格聚类分析光谱数据

从 LAMOST 巡天大数据中选取 M 型星的 205760 条恒星光谱数据，按 10 种类型进行网格聚类，分析聚类结果将恒星光谱数据聚集到较大的簇中，将拥有相同物理特征的数据聚集在一起，找出不符合恒星光谱数据的分布的光谱，充分保留光谱数据的物理特征，最后进行光谱数据的不同特征聚集离群数据分析。

结束语 天文数据挖掘是从飞速扩容的天文观测大数据信息数据库中提取隐含的、未知的以及具有应用价值的信息模式。中国天文学界研究已从单纯的数据获取提升到引领国际同行共享观测数据的高度[9]，过去的 10 年间我国天文学家增强了观测能力，大型巡天 LAMOST 之后的项目相继建成，如 FAST，HXMT，SVOM 等将投入观测，现已增加获取的天文数据达到 TB 量级，不久将会突破 BP 量级，下一步将融合天文数据科学应用开展相关工作。

(1)在之前已在 FORTRAN 语言、C 语言、IDL 语言、PYTHON 中应用 FITSIO 软件包研究的基础上，再引入 R 语言开发平台，充分发挥强大的统计、分析、数据挖掘的性能优势。

(2)目前天文大数据领域的研究，结合云计算获取海量的数据处理，依靠得力的软件工具读写 FITS 文件完善天文学领域统计分析数据的能力。

(3)R 语言应用于天文大数据挖掘中，利用 RFITIO 软件包对天文数据构造可视化光谱，在低维空间提取样本的主要特征点，高效地获取并挖掘天体信息。

参 考 文 献

[1] CUI X Q, ZHAO Y H, CHU Y Q, et al. The large sky area multi-object fiber spectroscopic telescope (LAMOST)[J]. RAA, 2012, 12(9): 1197.

[2] LUO A L, ZHAO Y H, et al. The first data release(DR1) of the LAMOST regular survey[J]. RAA, 2015, 15(8): 1104.

[3] 柯大荣，赵永恒. 一种图象传输系统及其 FITS 数据基本格式[J]. 现代图书情报技术，1994，10(2)：25-26.

[4] 崔辰州，李文，等. FITS 数据文件的检索和访问[J]. 天文研究与技术，2008，5(2)：117-119.

[5] 郭平，王可，罗阿理，等. 大数据分析中的计算智能研究现状与展望[J]. 软件学报，2015，26(11)：3011.

[6] 孙善武，王楠，欧阳丹彤. 基于聚类分析的业务流程模型抽象[J]. 计算机科学，2016，30(5)：104.

[7] 赵永恒. 大规模天文光谱巡天[J]. 中国科学：物理学力学天文学，2014，44(10)：1041-1045.

(上接第 413 页)

[5] GHANEM M, CHORTARAS A, GUO Y, et al. A grid of infrastructure for mixed bioinformatics data and text mining[J]. Computer Systems and Applications, 2005, 34(1): 116-130.

[6] KARANIKAS H, TJORTJIS C, THEODOULIDIS B. An approach to Text Mining using Information Extraction[C]// Proceeding of the Fourth European Conference on Principles and Practice of Knowledge Discovery in Database. Lyon, France, 2000: 13-16.

[7] HU Q, YU D, DUAN Y, et al. A novel weighting formula and feature selection for text classification based on rough set theory [C]// Proceedings of Natural Language Processing and Knowledge Engineering. 2003: 638-645.

[8] KOSALA R, BLOCKEEL H. Web Mining Research: A Survey [C]// ACM SIGKDD. 2000: 1-15.

[9] LI H, YAMANISHI K. Mining from Open Answers in Questionaire Data [C]// Proc. of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2001: 443-449.

[10] PONS-PORRATA A, BERLANGA-LAVORI R, RUI-SHULCLOPER J. Topic discovery based on text mining techniques [J]. Information Processing and Management, 2007, 43(3): 752-768.