第61卷 第2期 天 文 学 报 Vol.61 No.2
2020年3月 ACTA ASTRONOMICA SINICA Mar., 2020

doi: 10.15940/j.cnki.0001-5245.2020.02.009

一种新的基于2维傅里叶谱图像的恒星光谱特征提取方法和深度网络分类应用*

张静敏[1†] 马晨晔[2] 王 璐[1] 杜利婷[1] 许婷婷[3] 艾霖嫔[1]
周卫红[1,4‡]

(1 云南民族大学数学与计算机科学学院 昆明 650500)
(2 云南农业职业技术学院经济管理学院 昆明 650031)
(3 广州大学物理与电子工程学院 广州 510006)
(4 中国科学院天体结构与演化重点实验室 昆明 650011)

摘要 天体光谱分类是天文学研究的重要内容之一, 其关键是从光谱数据中选择和提取对分类识别最有效的特征构建特征空间. 提出一种新的基于2维傅里叶谱图像的特征提取方法, 并应用于LAMOST (the Large Sky Area Multi-Object Fiber Spectroscopic Telescope)恒星光谱数据的分类研究中. 光谱数据来源于LAMOST Data Release 5 (DR5), 选取30000条F、G和K型星光谱数据, 利用短时傅里叶变换(Short-Time Fourier Transform, STFT)将1维光谱数据变换成2维傅里叶谱图像, 对得到的2维傅里叶谱图像采用深度卷积网络模型进行分类, 得到的分类准确率是92.90%. 实验结果表明通过对LAMOST恒星光谱数据进行STFT可得到光谱的2维傅里叶谱图像, 谱图像构成了新的光谱数据特征和特征空间, 新的特征对于光谱数据分类是有效的. 此方法是对光谱分类的一种全新尝试, 对海量天体光谱的分类和挖掘处理有一定的开创意义.

关键词 **恒星**: 基本参数, **方法**: 数据分析, **技术**: 光谱分析

中图分类号: P144; **文献标识码**: A

1 引言

大天区面积多目标光纤光谱天文望远镜(the Large Sky Area Multi-Object Fiber Spectroscopic Telescope, LAMOST)[1]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. 自2011年启动大视场、多光纤光谱巡天观测, LAMOST已顺利走完8 yr的巡天之路. 2017年6月, LAMOST第1期低分辨率巡天圆满结束. 2017年9月1日开始, LAMOST开始了第2期中分辨率巡天, 目前获得的中分辨率的高质量光谱数据已达140万条左右. 2019年3月27日, 包含先导巡天及前6 yr正式巡天

2019-10-21收到原稿, 2019-12-09收到修改稿

*国家自然科学基金项目(61561053)资助

†1450255119@qq.com

‡ynzwh@163.com

的LAMOST Data Release 6 (DR6)数据集正式面向国内天文学家和国际合作者发布. LAMOST DR6共包含1125万条光谱, 其中高质量光谱达到了937万条, 约是国际上其他巡天项目发布光谱数之和的2倍, LAMOST成为了全世界获取光谱数第1个超过千万量级的巡天项目.

LAMOST的科学目标[2]之一是建立完善正确的恒星演化理论, 因此将恒星按照正确的特征进行分类至关重要. LAMOST巡天项目的快速发展对恒星光谱的自动分类和高效处理提出了更高的要求, 光谱特征提取[3]是恒星光谱自动分类的关键步骤, 信息丰富且无冗余的特征利于构建有效的特征空间进行分类识别. Chen等[4]采用K-means聚类算法, 利用AstroStat软件对LAMOST光谱的线指数进行聚类. 利用线指数分析天文光谱是对低分辨率光谱进行特征提取的有效方法, 线指数能代表恒星的主要光谱特征, 结果表明对线指数聚类可以有效检查数据质量和识别罕见对象. Wang等[5]则提出了一种新的自动光谱特征提取方法, 并基于此采用深度神经网络模型应用到光谱分类研究和缺陷谱恢复研究中.

在恒星光谱数据的分类研究方面, 刘中田等[6]提出一种恒星自动识别方法对SDSS (the Sloan Digital Sky Survey) DR4的光谱数据进行测试, 最后的测试结果显示M型星的分类精度高达98%; Zhong等[7]提出了一种模板匹配方法自动识别LAMOST晚期K型和M型矮星; 潘景昌等[8]利用贝叶斯方法对恒星光谱进行分类, 提高了恒星光谱数据的处理效率; Liu等[9]则计算了恒星光谱数据的线指数, 基于支持向量机(Support Vector Machine, SVM)进行分类, 结果显示A型和G型星的分类精度在90%以上, K型和OB型星的分类精度则低于50%, 有待提升.

面对快速增长的LAMOST巡天数据, 需要更高效智能的方法和工具去提高LAMOST望远镜的科学产出. 近年来, 机器学习尤其是深度学习在天体光谱分类上得到了广泛应用, 并取得一系列研究成果. Bai等[10]综合Pan-STARRS (Panoramic Survey Telescope And Rapid Response System) 1和WISE (Wide-field Infrared Survey Explorer)发布的ALLWISE数据, 应用机器学习对GAIA (Global Astrometric Interferometer for Astrophysics) DR2中85613922个数据对象进行分类识别, 并将分类结果与Simbad数据库进行了交叉匹配, 最终得到的分类准确率为91.9%. 为测量500 m口径球面射电望远镜(FAST)反射器上2226个节点的位置, Zhang等[11]采用带候选区域的卷积神经网络对图片中的节点进行检测, 实验结果表明识别率高达91.5%, 高于传统边缘检测的识别率. 石超君等[12]基于卷积神经网络构建了一个分类器, 对F型和K型星光谱自动分类, 并与SVM、误差反向传播算法对比, 采用交叉验证方法验证分类器性能. 可见在图像视觉领域广泛应用的卷积神经网络同样适用于天体光谱分类任务.

本文采用了深度学习中一种经典且有效的深度卷积网络Inception v3, 卷积神经网络(Convolutional Neural Network, CNN)[13]是一种具有局部连接、权重共享、汇聚特性的深层前馈神经网络, 一般由卷积层、汇聚层和全连接层交叉堆叠而成. 在CNN的卷积运算中, 1维卷积经常用于信号处理, 2维卷积则常用于图像处理, CNN在图像的分类识别上已经取得了长足的发展. 由于恒星光谱数据是1维的, 为了充分利用CNN在图像处理上的优势, 提出一种新的基于2维傅里叶谱图像的特征提取方式, 通过短时傅里叶变换(Short-Time Fourier Transform, STFT)将原始1维光谱数据变换成2维傅里叶谱图像,

由于变换形成新的能量分布, 构建了新的图像特征. 为验证此特征提取方法的性能, 选取30000条LAMOST DR5中的F、G和K型3种恒星光谱数据, 将原始恒星光谱数据变换成2维傅里叶谱图像, 并采用CNN中的Inception v3网络进行图像分类, 得到了较高的恒星光谱数据分类精度.

文章安排如下: 第2节介绍新的光谱数据特征生成方法; 第3节中我们采用CNN方法对新生成的特征图像进行了分类研究和实验; 第4节对分类实验结果进行了分析讨论; 第5节为结语.

2 恒星光谱数据的2维傅里叶谱图像变换

2.1 短时傅里叶变换

作为频域分析的基本工具, 傅里叶变换是信号在时域和频域运算的沟通桥梁, 但信号在傅里叶变换后丢失了时域信息. STFT在傅里叶分析的基本变换函数前乘上一个时间有限的时限函数, 即窗函数, 从而把一个较长的时间信号分成相同长度的更短的信号段, 在每个信号段上进行傅里叶变换, 实现了时域上的局部化. STFT定义了一个有效的时间和频率分布类, 利用STFT可以通过时间窗内的1段信号表示某一时刻的信号特征. 对信号$Z(u)$进行STFT的公式如下:

$$\mathrm{STFT}_{Z(u)} = \int_{-\infty}^{+\infty} [Z(u)g(u-t)]\mathrm{e}^{-\mathrm{j}\omega u}\mathrm{d}u\,, \tag{1}$$

其中, $Z(u)$为时间段u上的源信号, $g(u-t)$为以时刻t为中心的窗函数, 当t取不同值时$g(u-t)$在信号$Z(u)$上滑动, j是虚数单位, ω是信号函数中的基频. STFT先把一个函数和窗函数相乘, 再进行傅里叶变换, 并通过窗函数的滑动得到一系列的频谱函数, 将这些结果依次展开可以得到一个2维时频图. 本文利用STFT的这一特性将1维的恒星光谱数据变换成2维傅里叶谱图像.

2.2 恒星光谱数据和预处理

本文实验样本为取自LAMOST DR5数据库中的30000条恒星光谱数据(信噪比大于20), 这些数据在LAMOST 5D pipeline下被归类为F、G和K型光谱. 其中, 每种恒星光谱数据各10000条. 这些恒星光谱数据的波长范围在3700–9100 Å之间, 每1条恒星光谱数据包含了恒星在不同波长下的一系列辐射强度值, 即流量强度值.

同一条恒星光谱数据在不同波长下的流量值可能存在巨大的差异, 影响后续的数据处理和分析, 因此需要对原始数据进行归一化处理, 以消除流量值变化区间处于不同数量级的问题. 本文采用以下归一化方法:

$$b_{\mathrm{norm}} = \frac{b}{b_{\max}}\,, \tag{2}$$

其中, $b=(b_1,b_2,\ldots,b_n)$表示一条原始恒星光谱数据, $b_1,b_2,\cdots,b_n$表示给定波长下所对应的n个流量值, b_{norm}表示归一化后的恒星光谱数据, $b_{\max}$表示$b_1,b_2,\cdots,b_n$中的最大值, 即在某一波长下的流量值达到最大. 归一化处理后, 恒星光谱数据中的所有流量值均映射到[0, 1]之间, 数量级相同, 且仍然保留了原始光谱数据中各特征间的相对大小关系.

在LAMOST DR5数据库中, 每1条光谱都有唯一的天体编号OBSID, 一条F型星光谱数据(OBSID: 492302245)在归一化处理前后的对比图如图1所示.

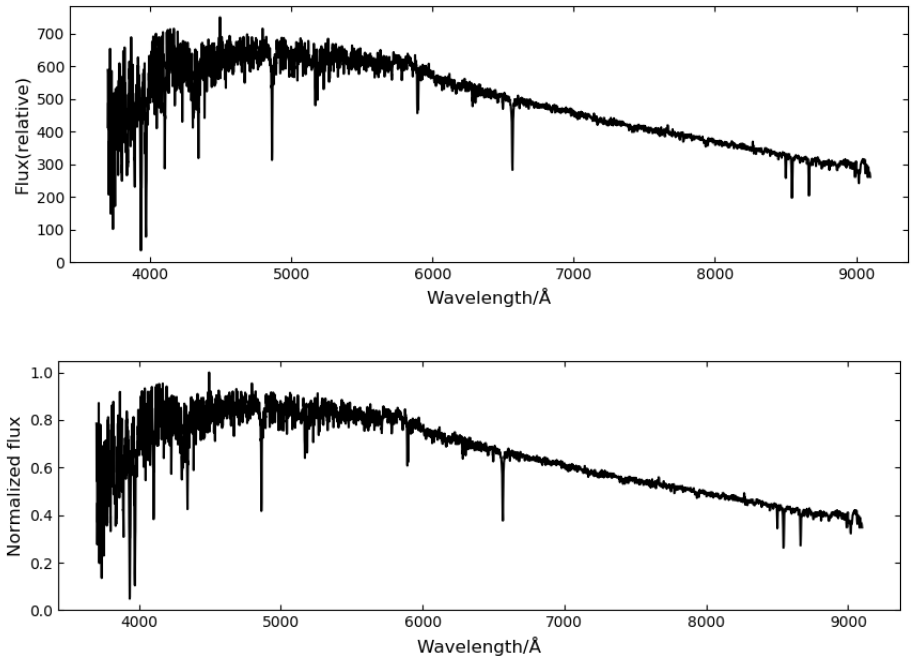

图 1 原始恒星光谱数据和归一化后的恒星光谱数据

Fig. 1 Original stellar spectral data and normalized stellar spectral data

2.3 光谱数据2维傅里叶谱图像的生成

在图像分类上, CNN通过卷积提取对分类识别有效的图像特征. 对于1维恒星光谱数据, 要想充分利用CNN的优势, 就不能直接作为数据输入. 对此, 提出一种新的基于STFT的特征提取方式, 利用STFT的时频解析性质, 把归一化后的1维恒星光谱数据变换成2维傅里叶谱图像.

对于恒星光谱数据, 将不同波长下所对应的流量强度值按照波长大小有序排列, 呈现出的是一幅不断上下波动的能量分布图. 基于此, 将恒星光谱数据看作某种意义上的能量信号, 并进行STFT. 实验利用Python中核心是STFT的specgram函数来得到恒星光谱数据的2维傅里叶谱图像.

Python中的specgram函数中的主要参数如下:

$$\text{specgram}(x, \text{window, noverlap, nfft, fs}, \cdots)\,, \tag{3}$$

其中, x是信号, 为1维数组或序列, window是窗函数, noverlap是帧重叠点数, nfft是傅里叶点数, fs是采样率. 利用specgram函数可计算并绘制x中数据的2维谱图. 其关键是利用窗函数将x中的数据分成nfft个数据段, 并分别计算每部分的频谱, 最后以彩图的形式绘制出频谱图. 其中, window窗口将应用于每段数据, 每段数据的重叠量用noverlap指

定. 对于1维恒星光谱数据, 在经过归一化处理后通过specgram函数(采样率设为2)可变换成2维傅里叶谱图像, 如图2所示.

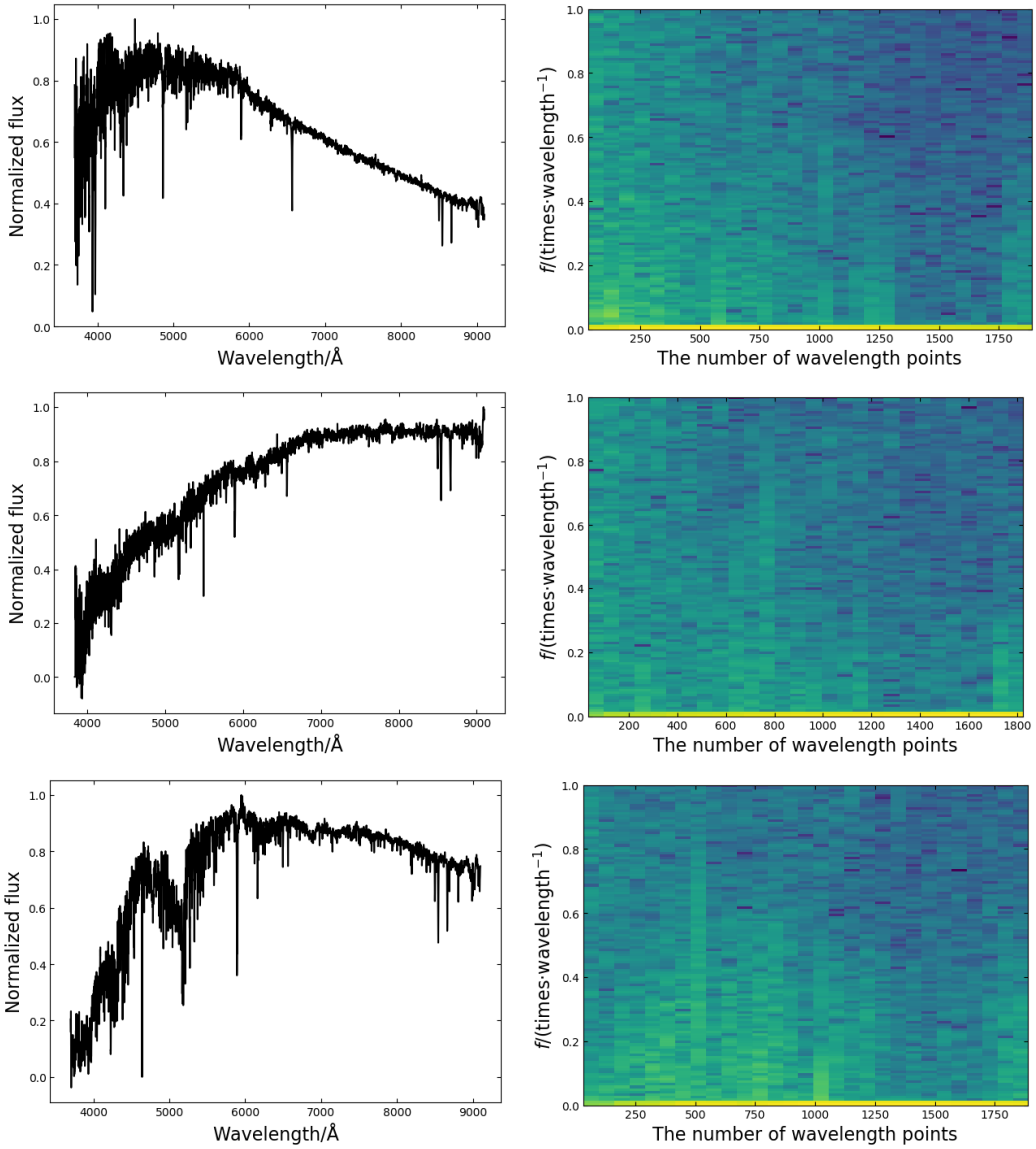

图 2 归一化后的1维恒星光谱数据和生成后的2维傅里叶谱图像, f是流量值在单位波长内变化的次数.

Fig. 2 Normalized one-dimensional stellar spectral data and generated two-dimensional Fourier spectral image, f means the changing times of fluxes in a unit wavelength.

图2中第1列是归一化后的恒星光谱数据, 第2列是对上述归一化后的恒星光谱数据进行STFT后所生成的对应2维傅里叶谱图像, 可以看到其中出现了新的特征. 谱图像是恒星光谱的一种新的谱线能量分布, 用2维平面表达3维信息; 与STFT传统的时频域应用不同, 谱图像的横坐标是和LAMOST观测谱线波长范围(3700–9100 Å)对应的

连续谱采样点, 原始的谱线数据约为4000个采样点, 图中所示为进行了2次采样后采样点(约2000个)的谱线能量分布图(2次采样是为了降低数据维度, 也可以不进行2次采样), 为处理方便, 采样点编号从0开始; 纵坐标是流量值在单位波长内变化的次数, 2维平面上的坐标点值为光谱在某一波长下的能量强度, 能量值的大小通过颜色来表示, 颜色越深越亮表示该点能量越强. 图2中3条恒星光谱数据的OBSID分别为: 492302245 (上)、15005130 (中)、546402094 (下).

3 基于2维傅里叶谱图像的深度卷积光谱分类

3.1 卷积神经网络

大数据时代的来临推动了深度学习以及神经网络的快速发展, 作为深度学习中的一种神经网络, CNN主要应用于图像以及视频分析等任务, 尤其是在处理诸如图像这种2维结构的数据上, CNN取得了巨大的成功. 经典的CNN网络有LeNet-5、AlexNet、GoogLeNet等. Inception网络最早的v1版本就是非常著名的GoogLeNet, 并赢得了2014年ImageNet图像分类竞赛的冠军.

深度学习的优势在于非线性关系的探索以及表现上, 理论上足够深层次的神经网络可以拟合任何复杂的关系函数. 因此一般来说, 提升网络性能最直接的方式就是增加网络的大小, 即增加网络的深度或者宽度, 但这样的简单处理容易致使网络陷入过拟合、梯度消失、网络计算量增大等问题. 对此, 深度卷积网络中的Inception网络构建了Inception模块来提升训练效果. 在Inception模块中, 一个卷积层实现了多个不同大小的卷积并行操作, 在相同的计算量下能提取到更多的特征, 更高效地利用了计算资源, 提升了训练效果. Inception网络有多个改进版本, 其中较有代表性的是Inception v3网络. Inception v3网络将大的卷积核替换成多层的小卷积核, 以减少计算量和参数量, 同时保持感受野不变. Inception v3的网络结构如图3所示.

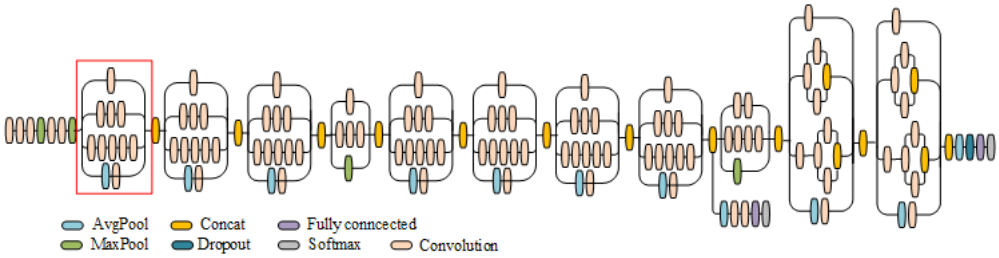

图 3 Inception v3结构

Fig. 3 Structure diagram of the Inception v3

Inception v3模型共46层, 由11个Inception模块以及少量汇聚层堆叠而成, 包含卷积层(Convolution)、最大池化层(MaxPool)、均值池化层(AvgPool)、连接层(Concat)、屏蔽层(Dropout)、激活层(Softmax)、全连接层(Fully connected), 图3中方框所标注出来的结构就是一个Inception模块. 在卷积神经网络中, 不同的卷积核相当于不同的特征提取器, Inception v3网络以并联的方式将不同的卷积层结合起来, 即同时使用不同尺寸的卷积核, 最后将得到的矩阵拼接起来, 从而更加灵活地进行特征提取.

3.2 实验设计和结果

通过STFT将30000条恒星光谱数据生成2维傅里叶谱图像后, 谱图像形成了新的特征. 为验证该特征提取方法的性能, 采用深度卷积网络中的Inception v3网络对2维谱图像进行分类识别. 其中, 学习率为0.05, 迭代次数为4000, 验证集3000条, 测试集3000条. 训练过程中, 每次随机训练100个样本, 每训练100次在验证集上进行一次数据验证, 不同迭代次数下的验证集分类准确率如图4所示.

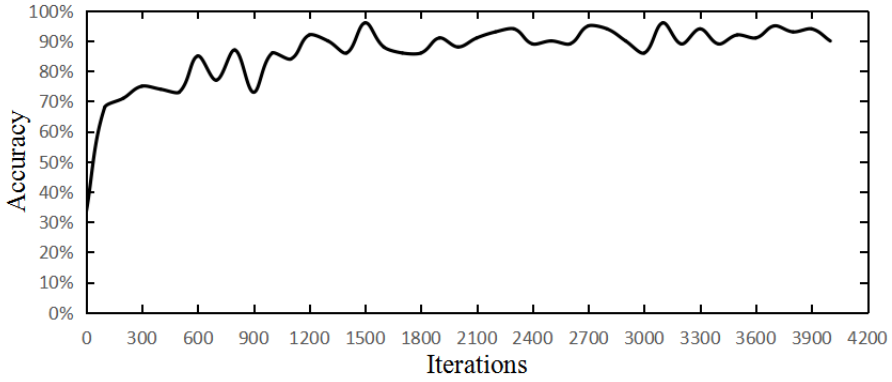

图 4 不同迭代次数下的验证集分类准确率

Fig. 4 Classification accuracy of the validation set under different iterations

由图4可知, 随着迭代次数的增加, 验证集分类准确率也随之上升. 图中显示验证集分类准确率曲线一直呈上升趋势, 从最初的快速增长到趋于平缓, 最后验证集分类准确率稳定在90%以上, 可见随着训练次数的增加, Inception v3网络的训练效果越来越好; 在训练验证交叉进行的过程完成后, 在测试集数据上得到高达92.90%的分类准确率.

4 结果分析

Wang等[5]提出了一种光谱特征提取方法, 对原始数据提取了721个光谱特征, 基于此训练了两种深度神经网络, 分别表示为PILDNN (表示基于伪逆学习算法的深度神经网络)和PILDNN* (表示将每条光谱作为输入向量时分为4个阶段并基于伪逆学习算法的深度神经网络), 选取了LAMOST DR1数据库中的50000条F、G和K型3种恒星光谱进行分类实验. 作为对比, 我们将文献[5]中的实验与本实验分类结果作对比, 分类结果见表1.

虽然本实验与文献[5]中的实验样本数目不同, 但样本数据都是选取自LAMOST数据库, 数据类型都是F、G和K型3种恒星光谱数据. 由表1可知, 基于我们的傅里叶谱图像特征及Inception v3网络的实验在分类精度上有较大提升.

由于光谱数据是典型的高维数据, 在已有的大多数光谱分类算法中, 大多先采用一定的算法(比如利用线指数特征、伪逆学习特征以及采用PCA (Principal Component Analysis)等方法)将光谱数据进行降维、提取特征, 然后再采用相应的算法进行分类等应用. 我们将1维的LAMOST恒星光谱数据进行短时傅里叶变换, 生成一种新的2维特征

谱图像, 再结合CNN在2维图像分类中无与伦比的性能优势, 提出一种全新的光谱图像特征变换提取和构造方法, 实验结果证明了方法的有效性.

表 1 与文献[5]分类结果比较
Table 1 Compared with the classification results in Ref.[5]

The number of sample	The form of sample after feature extraction	Classifier	Spend time/s	Accuracy/%
50000	one dimensional array	PILDNN	226.9495	81.90
50000	one dimensional array	PILDNN*	1103.4251	82.32
30000	two dimensional image	Inception v3	27001.0315	92.90

从原理上分析, 已有的很多算法对数据的处理都是首先进行降维. 降维虽然减少了后续数据处理的复杂性和计算量, 但降维的同时毫无疑问丢失了一定的数据特征信息, 将不可避免地造成分类精度下降. 相反, 我们提出的短时傅里叶变换特征谱图像方法, 对数据进行了升维, 将1维的光谱数据升到2维的傅里叶谱图像. 该光谱特征提取方法不但没有丢失特征信息, 并且由于光谱通过傅里叶变换形成了新的能量分布, 构成了新的适合分类的图像特征, 结合CNN在2维图像分类上的优势, 大大提高了分类精度. 我们提出的新方法为海量天体光谱数据的处理提供了一种新思路, 具有一定的开创意义.

5 结语

本文提出一种新的基于2维傅里叶谱图像的恒星光谱特征提取方法, 通过此方法将1维恒星光谱数据变换成2维傅里叶谱图像, 构造出新的特征空间; 结合CNN在2维图像分类上的优势, 提出一种新的光谱数据分类方法. 针对LAMOST的F、G和K型3种恒星光谱数据, 通过STFT将恒星光谱数据生成2维傅里叶谱图像, 并基于此采用卷积神经网络中的Inception v3网络对得到的2维傅里叶谱图像进行分类, 最后得到的分类精度为92.90%, 结果验证了该特征提取方法的优越性能.

参 考 文 献

[1] 杨海峰. 天体光谱数据挖掘与分析. 北京: 电子工业出版社, 2016: 2-14
[2] 褚耀泉. 中国科学技术大学学报, 2007, 37: 591
[3] 李乡儒. 天文学进展, 2012, 30: 94
[4] Chen S X, Sun W M, Yan Q. RAA, 2018, 18: 73
[5] Wang K, Guo P, Luo A L. MNRAS, 2017, 465: 4311
[6] 刘中田, 邱宽民. 光谱学与光谱分析, 2010, 30: 274
[7] Zhong J, Lépine S, Hou J L, et al. AJ, 2015, 150: 42
[8] 潘景昌, 王杰, 姜斌, 等. 光谱学与光谱分析, 2016, 36: 2651
[9] Liu C, Cui W Y, Zhang B, et al. RAA, 2015, 15: 1137
[10] Bai Y, Liu J F, Wang S. RAA, 2018, 18: 156
[11] Zhang Z S, Zhu L C, Tang W, et al. RAA, 2019, 19: 113
[12] 石超君, 邱波, 周亚同, 等. 光谱学与光谱分析, 2019, 39: 1312
[13] 李玉鉴, 张婷. 深度学习导论及案例分析. 北京: 机械工业出版社, 2016: 38-63

A New Stellar Spectral Feature Extraction Method Based on Two-dimensional Fourier Spectrum Image and Its Application in the Stellar Spectral Classification Based on Deep Network

ZHANG Jing-min[1] MA Chen-ye[2] WANG Lu[1] DU Li-ting[1] XU Ting-ting[3]
AI Lin-pin[1] ZHOU Wei-hong[1,4]

(1 School of Mathematics and Computer Science, Yunnan Minzu University, Kunming 650500)
(2 School of Economics and Business Administration, Yunnan Vocational and Technical College of Agriculture, Kunming 650031)
(3 School of Physics and Electronic Engineering, Guangzhou University, Guangzhou 510006)
(4 Key Laboratory of the Structure and Evolution of Celestial Objects, Chinese Academy of Sciences, Kunming 650011)

Abstract The classification of celestial spectra is one of the important contents of astronomical research. The key is to select and extract the most effective feature for classification from spectra data. In this paper, we propose a new feature extraction method for astronomical spectra based on two-dimensional Fourier spectrum image, and apply the method to the classification study of LAMOST (the Large Sky Area Multi-Object Fiber Spectroscopic Telescope) stellar spectral data. The spectra data are from LAMOST Data Release 5 (DR5). We select 30000 F, G, and K types of spectra data. The short-time Fourier transform (STFT) is used to transform the one-dimensional spectra data into two-dimensional Fourier spectrum images. We classify and test these two-dimensional Fourier spectrum images with a module based on deep convolutional network, and the classification accuracy rate is 92.90%. The experimental result shows that the LAMOST stellar spectra data can be transformed into the two-dimensional Fourier spectrum images by the STFT. These spectral images inform new features, and build a new feature space, which is effective for classification. The method is a fully new attempt in spectra classification, which has certain pioneering significance for the classification and mining of massive celestial spectra.

Key words stars: fundamental parameters, methods: data analysis, techniques: spectral analysis