第62卷 第5期 天 文 学 报 Vol.62 No.5
2021年9月 ACTA ASTRONOMICA SINICA Sept., 2021

doi: 10.15940/j.cnki.0001-5245.2021.05.002

基于卷积神经网络的恒星光谱型和光度型的分类模型*

洪舒欣[1,2] 邹志强[1,2†] 徐灵哲[3]

(1 南京邮电大学计算机学院 南京 210023)
(2 江苏省大数据安全与智能处理重点实验室 南京 210023)
(3 中国科学院国家天文台南京天文光学技术研究所望远镜新技术研究室 南京 210042)

摘要 恒星光谱分类是天文学中一个重要的研究问题. 对于已经采集到的海量高维恒星光谱数据的分类, 采用模式匹配方法对光谱型分类较为成功, 但其缺点在于标准恒星模版之间的差异性在匹配实际观测数据中不能体现出来, 尤其是当需要进行光谱型和光度型的二元分类时模版匹配法往往会失败. 而采用谱线特征测量的光度型分类强烈地依赖谱线拟合的准确性. 为了解决二元分类的问题, 介绍了一种基于卷积神经网络的恒星光谱型和光度型分类模型(Classification model of Stellar Spectral type and Luminosity type based on Convolution Neural Network, CSSL_CNN). 这一模型使用卷积神经网络来提取光谱的特征, 通过注意力模块学习到了重要的光谱特征, 借助池化操作降低了光谱的维度并压缩了模型参数的数量, 使用全连接层来学习特征并对恒星光谱进行分类. 实验中使用了大天区面积多目标光纤光谱天文望远镜(Large Sky Area Multi-Object Fiber Spectroscopy Telescope, LAMOST)公开数据集Data Release 5 (DR5, 用了其中71282条恒星光谱数据, 每条光谱包含了3000多维的特征)对该模型的性能进行验证与评估. 实验结果表明, 基于卷积神经网络的模型在恒星的光谱型分类上准确率达到92.04%, 而基于深度神经网络的模型(Celestial bodies Spectral Classification Model, CSC_Model)只有87.54%的准确率; CSSL_CNN在恒星的光谱型和光度型二元分类上准确率达到83.91%, 而模式匹配方法MKCLASS仅有38.38%的准确率且效率较低.

关键词 恒星: 基本参数, 方法: 数据分析, 深度学习: 卷积神经网络

中图分类号: P152; **文献标识码**: A

1 引言

从古至今, 人类对宇宙的探索就从未停止过, 从人类观察天象、天体开始算起, 天文学已经有5000多年的历史了. 随着科技的进步, 当前的大型天文望远镜已经可以同时观测数千个天体目标, 每个观测夜晚可以采集到数万条天体光谱. 我国的郭守敬望远

2020-12-15收到原稿, 2021-02-02收到修改稿

*国家自然科学基金项目(U1931207)资助

†zouzq@njupt.edu.cn

镜(Large Sky Area Multi-object Fiber Spectroscopic Telescope, LAMOST)可以同时获得4000条天体光谱, 是当今世界上光谱获取率最高的天文望远镜[1]. LAMOST光谱巡天主要包含两个部分: LAMOST河外巡天(LEGAS)和LAMOST银河系巡天(LEGUE)[2]. 2019年3月, LAMOST Data Release 6 (DR6)数据集对海内外研究者正式公布, 该数据集一共包含了4902个观测天区, 收集了1125万条光谱数据, 可供专家学者分析和研究. 面对海量的天文光谱数据[3], LAMOST的数据处理管道软件(Pipeline)对所有光谱进行了分类. 然而该Pipeline采用的是基于交叉相关的模版匹配方法, 尽管较好地完成了光谱分类任务, 但该方法运算量大、对模板的依赖大, 在处理低质量光谱时有一定的困难, 迫切需要研究出能够高效准确地识别、分析和处理天文光谱数据的自动化方法.

本文分为以下几个部分: 第2节主要介绍了天体分类的相关研究. 第3节描述了实验使用的数据集和相应的数据预处理, 并详细介绍了使用的理论和设计的模型. 第4节介绍了实验验证与性能评估, 将实验结果和其他模型进行了对比讨论. 第5节是我们的结论和对未来工作的展望.

2 相关工作

天文光谱分类是天文研究的重要组成部分, 恒星是构成星系和宇宙的基本单元. 恒星的分类, 不仅有助于了解恒星物理学, 而且可以推动对银河系的整体结构和演化的研究. 当前对恒星光谱分类的方法主要有两大类: 模式匹配方法和机器学习方法.

2.1 模式匹配方法

模式匹配方法是通过肉眼比较光谱与少量标准恒星样本来完成的, 目前最通用的恒星分类方式是由美国天文学家摩根和基南在20世纪40年代提出的MK (Morgan-Keenan)分类系统, 它的判别依据是恒星光谱中的特征谱线、谱带以及这些谱线和谱带的相对强度. 尽管已经通过开发自动软件努力使MK分类过程自动化(例如Gray等[4]), 但是由于巡天项目采集到的恒星光谱信噪比范围很广, 使得光谱特征不像少数观测良好的高质量标准光谱那样清晰, 所以观测到的恒星光谱很难与标准光谱进行匹配. 而且标准恒星样本通常非常明亮且位于太阳附近, 而新的光谱研究可以探测深空目标, 因此当前的标准恒星库并不完整[5].

2.2 机器学习方法

机器学习[6]方法可以通过学习大量光谱的特征训练识别光谱的能力, 从而对无标签的光谱进行自动分类. 随着机器学习研究与应用的深入, 越来越多的机器学习方法被用于天文光谱数据的分析与处理[7–8]. Liu等[5]将支持向量机(SVM)应用到恒星光谱分类中, 发现A和G型恒星的分类完备性高达90%; 而OB和K型恒星的分类, 由于完备性低至约50%, 导致大约40%的OB型和K型恒星分别被误分类为A型和G型恒星. 此外, Elting等[9]和Saglia等[10]描述了SVM在星系-类星体分类中的应用. 但是SVM算法对大规模样本数据训练比较困难, 且在解决多分类问题时存在一定的局限. Folkes等[11]在1996年使用人工神经网络(ANN)模型对星系样本进行了分类. Bailer-Jones等[12]在1997年将ANN用于恒星光谱分类, 对矮星和巨星在光度类型分类上准确率达到95%.

由于机器学习算法用来提取特征的层数较少, 无法提取恒星光谱的更高层特征, 所以机器学习的一个分支—深度学习[13–14]也已经应用到恒星光谱分类问题中[15–16]. Fabbro等[17]、Zou等[18]构建了深度神经网络, 使用卷积[19–20]的方法来对光谱数据进行分类, 得到了比一般神经网络更好的效果. Hon等[21]使用一维卷积神经网络对红巨星进行分类, 准确率达到99%. Liu等[22]建立了9层卷积网络, 对F型、G型和K型星的分类准确率分别达到90%、93%和97%. 随着模型深度的加深, 提取的特征更深入, 然而这样会带来梯度弥散或梯度爆炸的问题, 残差网络[23]的提出解决了这一问题, 可以构建极深的神经网络, 并极快地加速网络训练, 戴加明等[24]已将深度残差网络用于星系形态分类. 注意力机制[25–26]可以提高模型在训练中的针对性, 它能够自动关注对分类有利的特征, 而忽略无效特征. 2020年, Zou等[27]提出了基于残差和注意力机制的卷积网络, 对天体光谱的分类准确率高达98.92%.

从以上的讨论可以看出, 当前恒星光谱分类算法仍存在一些问题. 模式匹配方法受限于目前的标准恒星库, 对质量较差的恒星光谱或者与模板没有类似谱线特征的恒星都无法进行分类判断. 传统的机器学习算法难以处理高维的海量恒星光谱数据, 且对特征的提取不够深入. 由于神经网络可以挖掘数据深层的隐式特征, 所以一般的神经网络可以实现恒星光谱型分类的任务, 但是, 目前仍然缺乏同时兼顾恒星光谱型和光度型的二元分类模型, 故本文提出了一种基于卷积神经网络的恒星光谱型和光度型分类模型(Classification model of Stellar Spectral type and Luminosity type based on Convolution Neural Network, CSSL_CNN). 不同于一般的分类模型, CSSL_CNN分类模型针对海量高维的恒星光谱数据, 构建了卷积神经网络来提取恒星光谱的深层特征, 利用注意力模块学习重要的光谱特征, 并设计了池化层对高维特征进行压缩降维, 将处理后的数据输入Softmax分类器, 最终通过训练分类器使得我们的CSSL_CNN模型实现了对恒星在光谱型和光度型两个维度上进行分类的功能.

3 方法

本节的组织结构如下: 3.1节介绍了本文工作的整体框架, 3.2节详细介绍了使用的数据集和对数据进行的预处理步骤, 3.3节详细介绍了本文CSSL_CNN分类模型的组成结构和各部分的作用.

3.1 整体框架

图1描述了我们工作的整体流程, 包含以下3个部分:

(1)恒星光谱数据的获取及预处理. LAMOST给出的天文光谱数据中包含了恒星、星系、类星体和未知4大类, 本文的研究对象是其中的恒星, 因此要从下载的数据集中筛选出我们需要的恒星光谱数据. 由于天文光谱数据具有高维、分布不均等特点, 所以要对收集到的恒星光谱数据进行压缩、归一化等操作, 对标签进行One-hot编码以保持其独立性, 为后续的模型提供优良的数据集;

(2)分类模型的构建与训练. 面对高维非线性的恒星光谱特征, 采用了卷积神经网络模型对步骤(1)中处理好的恒星光谱数据进行深层特征提取, 使用反向传播算法训练和优化模型;

(3)分类模型的验证与评估. 使用训练好的模型对无标签的恒星光谱进行分类, 并且对分类结果进行评估.

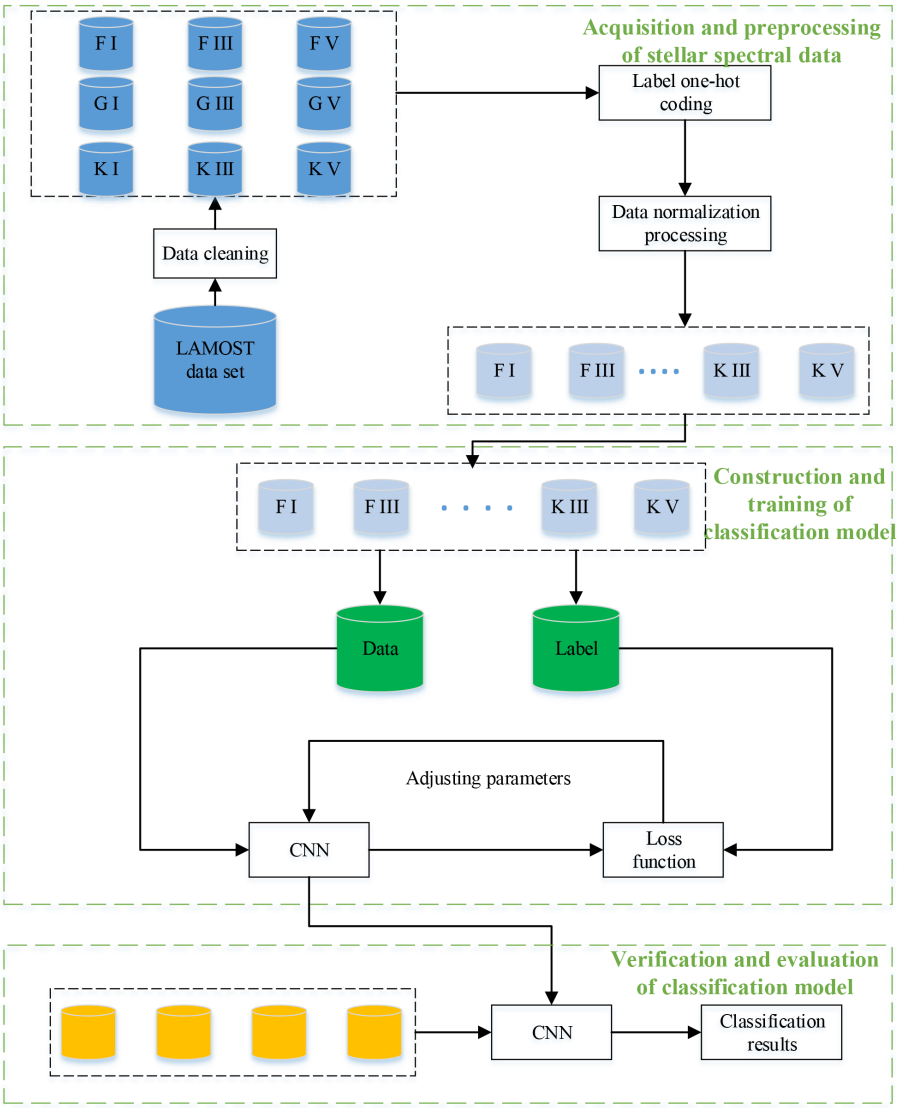

图 1　整体框架

Fig. 1　Overall framework

3.2　数据获取和预处理

我们使用的数据集来自LAMOST DR5数据集, 可以在官网[1]中的Fits Download里进行查询和下载. 数据集里共包含71282条恒星光谱数据, 每条数据都有两个标签, 在光谱型上可以是F型、G型或K型, 在光度型上可以是巨星(记为I)、亚巨星(记为III)或矮星(记为V). 所有的数据共可分为9类: (F, I)、(F, III)、(F, V)、(G, I)、(G, III)、(G,

[1]http://dr5.lamost.org/

V)、(K, I)、(K, III)、(K, V). 每条光谱包含了3910个左右的特征, 波长覆盖了3699–9100 Å的范围, 图2是一个G型恒星的光谱示意图, 其中横坐标是光谱波长, 纵坐标是对应位置的通量.

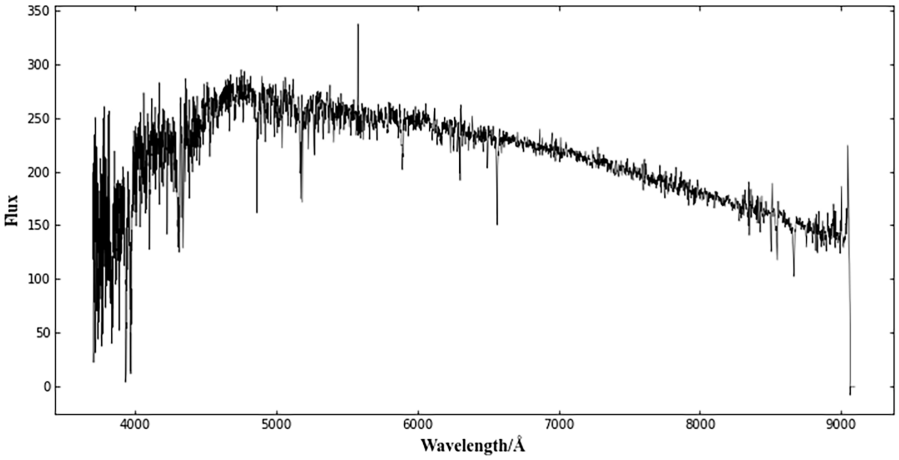

图 2　G型恒星光谱图

Fig. 2　Spectrogram of a G-type star

考虑到(F, I)、(F, III)、(F, V)、(G, I)、 ···、(K, V)标签是离散、无序的, 如果对这9个标签直接进行数字化处理, 即仅用0、1、2、3、···、8这样的数字来分别代表(F, I)、(F, III)、(F, V)、(G, I)、···、(K, V), 则破坏了这9种类型之间的独立性, 因为数值大小会影响到权重矩阵的计算, 而One-hot编码采取M位状态寄存器来对M个状态进行编码, 每一个状态都有它独立的寄存器位, 并且在任意时候只有一位是有效的. 因此, 本文使用了One-hot编码, 既可以保留这种编码的独立性使距离计算更加合理, 又能使损失函数的计算变得非常方便. 在经过One-hot处理后, 9种标签的编号分别为(100000000, 010000000, 001000000, 000100000, ···, 000000001).

不同于一般的数据, 恒星光谱数据之间的差异极大, 不同波长对应的光强差别很大, 因此必须要对恒星光谱数据进行归一化处理, 否则模型将会把关注点放在高强度的特征上, 而忽视强度低的特征. 归一化就是将我们需要的数据经过处理后限制在0∼1的范围内, 在后续使用梯度下降法求解最优化问题时, 归一化后可以加快梯度下降的求解速度, 即提升模型的收敛速度. 此外, 这里的归一化处理还可以消除特征数据之间的量纲影响, 解决特征指标之间的可比性问题. 常用的归一化方法有min-max标准化、z-score标准化和L2范数归一化等, 我们采用的是L2范数归一化, 这是因为L2范数是对向量各元素的平方和求平方根, 它可以防止过拟合, 还可以保证天体光线在传播过程中的衰减不会影响到模型的学习[27]. 具体的形式化描述为:

$$x_i' = \frac{x_i}{\|x_i\|_2}\,, \tag{1}$$

其中, x_i表示第i条光谱数据, $\|x_i\|_2$表示第i条光谱数据的L2范数, x_i'表示归一化后的第i条光谱数据.

3.3 基于卷积神经网络的CSSL_CNN分类模型

为了快速高效地处理海量高维的恒星光谱数据, 本文提出了一种基于卷积神经网络的CSSL_CNN分类模型, 下面将对这个模型的各部分进行详细说明.

CSSL_CNN分类模型是一种有监督分类模型, 它通过学习大量有标签的恒星光谱数据获得识别光谱类别的能力, 从而对给出的无标签恒星光谱数据进行高准确率的分类.

我们构建的CSSL_CNN分类模型由4个卷积层、4个注意力块、2个池化层、2个全连接层组成, 如图3所示.

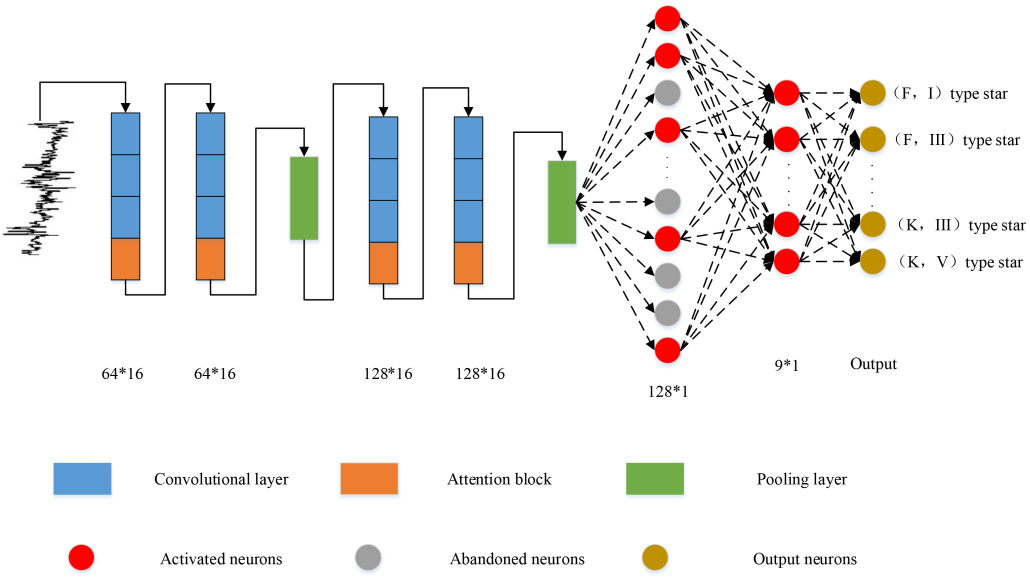

图 3 CSSL_CNN分类模型结构

Fig. 3 Structure of CSSL_CNN

图3中卷积层中的卷积核数量逐层增加, 前两个卷积层中各有64个长度为16的卷积核, 后两个卷积层中各有128个长度为16的卷积核, 从而可以提取到更深层的特征. 卷积层后的池化层压缩了特征, 提取出主要特征. 最后有两个全连接层, 分别包含128和9个神经元, 以0.5的概率丢弃掉其中部分神经元, 防止过拟合. 最终根据全连接层的计算结果输出恒星分类结果.

在输入时, 首先要提供input_shape, 每条恒星光谱包含的特征个数不一致, 在3910左右, 为保持数据长度一致, 且考虑到实际情况, LAMOST的光谱在减天光时, 在红端有些天光残留较多有些残留较少, 这会对模型训练产生较大影响. 如果保留这些数据, 它们带来的干扰可能比从它们中获得的特征还要大, 所以在实验中统一保留前3600个特征.

卷积层用来对特征进行提取和映射. 卷积的过程基于一个卷积核, 卷积核相当于一个过滤器, 完成恒星光谱分类的特征提取. 卷积核内的每一个值都是我们需要训练的神经元参数, 起初会有个初始值, 在训练网络的过程中, 网络会通过反向传播不断地更新这些参数值, 直到寻找到最佳参数值. 当卷积核扫描整个数据时, 卷积核的参数值是固定不变的, 即所有数据都共享相同的权值, 这大大削减了卷积核中的参数个数. 由于恒星光谱

数据是一维数据, 所以使用一维卷积.

由于神经网络的权值、偏置都是线性变化, 而线性模型的表达能力不够, 所以需要使用激活函数来加入非线性因素. 我们使用的激活函数是ReLU函数, 表达式如下所示:

$$\mathrm{ReLU}(x) = \max(x, 0)\,, \tag{2}$$

其中, max是取最大值的函数, x表示上一层网络的输出.

注意力机制能够使模型聚焦于对分类任务有重要作用的特征上. 人脑在进行阅读、看图时, 会自动忽略低价值的信息, 优先获取大脑认为有用的信息, 即重要的局部特征. 类似地在识别光谱数据时, 由于光谱信息受天光背景影响较大, 特别是某些特定波段, 因此模型需要识别出这些受影响更大的波段, 适当降低权重, 并且更加关注能提供重要特征的波段. 在卷积运算中, 每个卷积滤波器生成一个信道, 传统的卷积神经网络认为每个信道携带的信息对后续分类的贡献相同, 然而实际上, 有些特征对分类有很大的贡献, 有些特征只有很小的贡献, 甚至会产生负面影响. 因此, 我们引入注意力块来为不同的信道分配权重, 重要的信道可以得到更高的权重, 使得模型能抓住重点进行学习, 从而优化分类结果.

卷积操作后, 很多特征信息被提取出来, 但是相邻区域会有相似特征信息, 如果全部保留就造成了信息冗余, 增加了计算难度, 这时候就需要进行池化操作来降维, 压缩数据和参数数量, 降低过拟合风险.

由于恒星光谱数据的特征维度很高, 所以要用池化层来压缩输入的特征, 使特征矩阵变小, 一方面能够简化网络运算的复杂性, 另一方面又可以提取出恒星光谱数据的主要特征. 在本实验中, 采用的池化方法是Max Pooling, 在一个局部区域中, 取该区域的最大值来代替该区域, 滑过整个区域之后, 就得到一个比之前小很多的矩阵.

经过前面多次卷积、激励和池化的操作之后, 将提取好的特征输入全连接层. 全连接层在卷积神经网络中起“分类器”的作用, 它将前面高度抽象化后的特征进行了整合, 将学到的特征表示映射到样本空间中, 使用分类器Softmax对各种分类情况都计算出一个概率, 最终输出分类结果. Softmax是将N分类的问题, 转化为一个$N \times 1$维的向量, 向量的每行代表属于该类别的概率, N个概率和为1, 最终输出最大概率对应的类别. Softmax的表达式如下所示:

$$\mathrm{Softmax}(m_j) = \frac{\mathrm{e}^{m_j}}{\sum_{l=1}^{n} \mathrm{e}^{m_l}}\,, \tag{3}$$

其中, n表示类别的个数, m_j表示第j个节点的输出值, m_l表示第l个节点的输出值, e是自然常数.

在全连接层之前, 若神经元数目过大, 学习能力强, 也许会出现过拟合问题. 因此, 可以引入dropout操作, 来随机舍弃神经网络中的部分神经元, 解决此问题.

模型训练使用的优化器是Adam优化器, 学习率是0.001. 在训练过程中, 采用了反向传播(Back Propagation, BP)算法. BP算法的核心是对整个网络所有可能的路径重复使用链式规则, 将输出与真实标签之间的损失值一层一层反向传递回输入层, 这样可以让每个层神经元间的权值得到修正, 从而最大限度减小误差. 反向传播在经过全连接层时, 会根据全连接层的权重分配误差, 权重越大的, 反向传播误差就会分配到越多. 反向传播

算法强大的地方在于它是动态规划的, 可以重复使用中间结果计算损失函数对于每个权重的梯度, 从而对每层的参数进行更新, 直到输出中的偏差减小到可接受范围内或者到达预设的训练次数为止.

实验中使用的损失函数是交叉熵损失函数, 表达式如下所示:

$$\mathrm{loss}_k = -\sum_{c=1}^{n} y_{kc} \lg(P_{kc}) \,, \tag{4}$$

其中, y_{kc}是第k个数据的真实标签, P_{kc}是第k个数据属于类别c的预测概率, 即Softmax求出的值.

为了使损失函数尽可能的小, 我们使用一种叫梯度下降的技巧来调整神经元之间的权重, 从而达到最佳的结果. 在微积分里梯度的几何意义就是函数变化最快的方向, 所以沿着梯度向量相反的方向, 就可以更快地找到损失函数的最小值.

4 实验验证与性能评估

4.1 实验设置

我们在一台CPU型号为Intel® Core (TM) i7-9750H CPU (2.60 GHz), 运行内存为8 GB, SSD (固态硬盘)大小为512 GB的电脑上进行了实验.

CSC_Model (Celestial bodies Spectral Classification Model)和CSSL_CNN训练过程中, 训练集和测试集的比例设为8:2, 批处理大小为16, 训练轮数为10.

4.2 实验结果和讨论

在这一部分中, 两种经典的方法被用来做对比, 我们做了以下实验:

实验1: 在光谱型分类上, 使用CSC_Model[18]和CSSL_CNN进行对比实验;

实验2: 在光谱型和光度型二元分类上, 使用MKCLASS程序[4]和CSSL_CNN进行对比实验.

4.2.1 实验1: 光谱型分类

在光谱型分类上, 本文对比了CSC_Model, CSC_Model是自编码器加多层感知机的模型, 其中自编码器用于提取特征和降维, 多层感知机用于分类. 使用F、G、K、M型恒星光谱各10000条进行了实验, accuracy指标用于对模型的性能进行评估.

CSC_Model和CSSL_CNN在恒星光谱型分类上的实验结果如表1所示.

表 1 光谱型分类实验结果

Table 1 Experimental results of spectral classification

Model	Accuracy (%)	Spend time/min
CSC_Model	87.54	25
CSSL_CNN	92.04	303

CSC_Model和CSSL_CNN的训练过程分别见图4和图5.

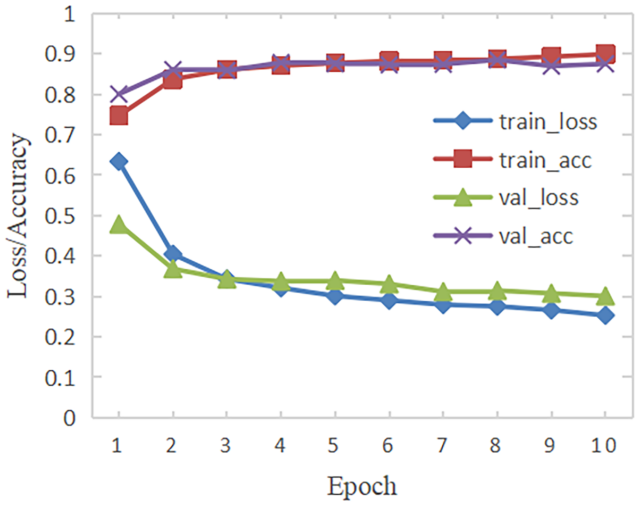

图 4 CSC_Model光谱型分类训练过程

Fig. 4 Training process of spectral classification of CSC_Model

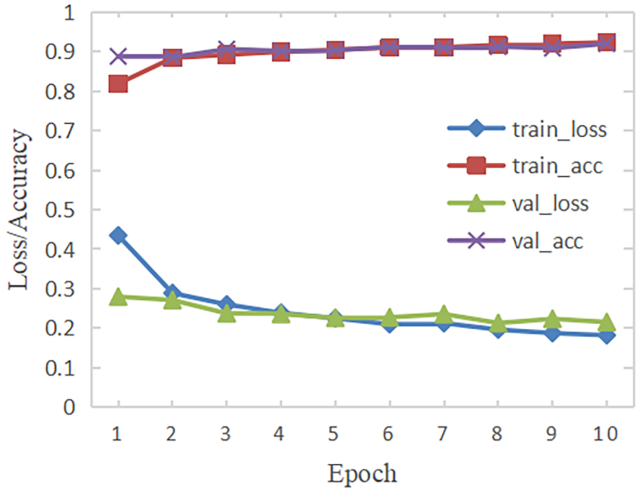

图 5 CSSL_CNN光谱型分类训练过程

Fig. 5 Training process of spectral classification of CSSL_CNN

CSC_Model花费时间25 min, CSSL_CNN需要303 min. 图中train_loss指的是训练过程中模型的损失值, train_acc指的是训练过程中模型的分类准确率, val_loss指的是测试过程中模型的损失值, val_acc指的是测试过程中模型的分类准确率. 横坐标Epoch代表训练的轮数, 纵坐标代表每次训练的准确率(Accuracy)和损失值(Loss). 从图中可以看出, 随着训练次数的增加, 损失逐渐减小准确率逐渐上升, 最终平缓地趋于拟合. CSC_Model的准确率最终达到87.54%, CSSL_CNN的准确率最终达到92.04%. CSC_Model中的自编码器将特征从3000压缩到750, 大大减少了训练时间, 但是也因此丢失了部分重要信息, 导致分类准确率不如CSSL_CNN高.

4.2.2　实验2: 光谱型和光度型二元分类

在光谱型和光度型分类上, 本文对比了MKCLASS软件, MKCLASS是用C语言实现的自动对恒星光谱进行分类的程序, 它本质上是模式匹配方法. 使用带有二元标签的恒星光谱数据共71282条进行实验, 数据集构成见表2.

表 2　数据集构成

Table 2　Data set composition

Spectral \ Luminosity	I	III	V
F	4043	2829	10000
G	10000	10000	10000
K	10000	4410	10000

MKCLASS和CSSL_CNN在恒星光谱型和光度型二元分类上的对比实验结果如表3所示.

表 3　二元分类实验结果

Table 3　Experimental results of spectral and luminosity classification

Model	Accuracy (%)	Spend time/min
MKCLASS	38.38	1070
CSSL_CNN	83.91	538

从表中可以看出, MKCLASS在LAMOST DR5数据上分类准确率较低, 这是因为MKCLASS在面对低信噪比的恒星光谱或者通量出现负值的光谱时就无法给出分类结果. MKCLASS使用脚本来对恒星光谱进行批处理, 花费时间约1070 min.

CSSL_CNN在恒星光谱型和光度型的二元分类上训练过程如图6所示.

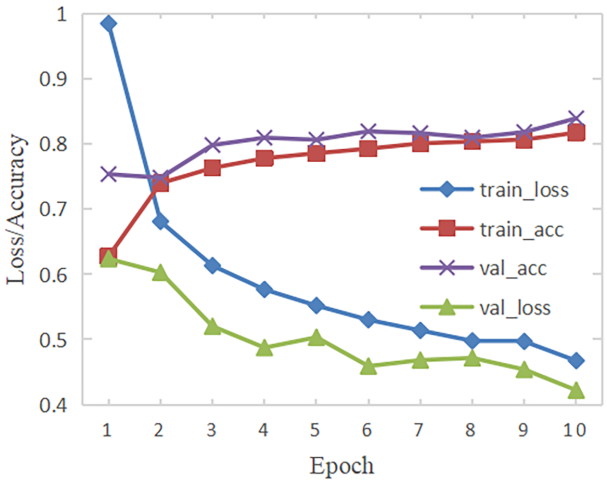

图 6　CSSL_CNN二元分类训练过程

Fig. 6　Training process of spectral and luminosity classification of CSSL_CNN

训练一开始损失值较大, 说明模型的预测结果与真实标签有较大差别. 随着训练轮数的增加, 损失值有了大幅度的降低, 准确率则一直平稳上升, 最终达到83.91%.

CSSL_CNN分类结果的混淆矩阵如图7所示.

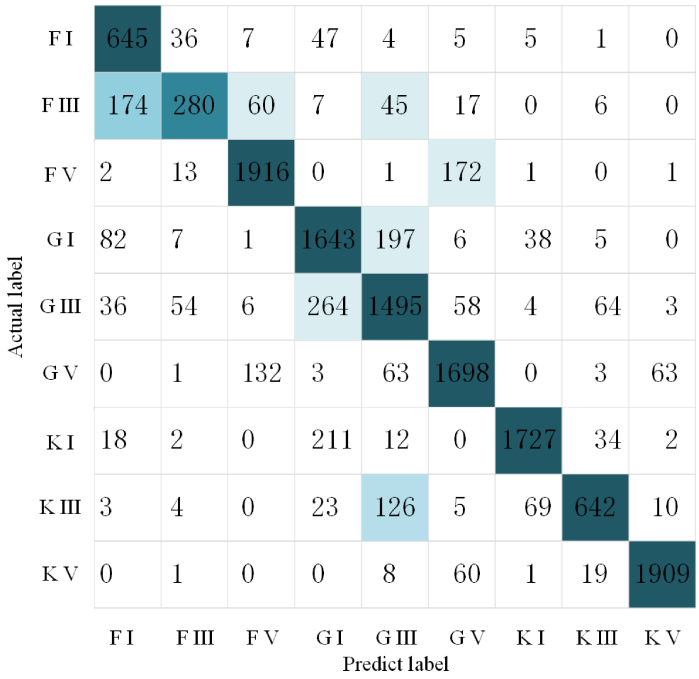

图 7　CSSL_CNN二元分类的混淆矩阵

Fig. 7　Confusion matrix of spectral and luminosity classification of CSSL_CNN

可以看出绝大部分预测结果都与标签一致, 通过分析, 可以知道分类出错的两大主要原因分别是:

(1)相邻的光谱型之间特征差异不大, 由于光谱型是连续变化的温度序列, 相邻两个类型分类错误其实不一定是错的, 比如A9和F2很接近, 比A1和A9还接近, 如果把F2分成A9在某种程度上不算错误. 从混淆矩阵中可以看出, 有很多分类错误的都是光度级没错, 光谱型错误判断成了相邻光谱型, 比如F型亚巨星(F III)有7.64%错误分类成G型亚巨星(G III), F型矮星(F V)有8.17%错误分类成G型矮星(G V), K型亚巨星(K III)有14.29%错误分类成G型亚巨星(G III);

(2)巨星(I)和亚巨星(III)之间的特征差异也相对较小. 从混淆矩阵可以看出, 光度类型上分类错误的大部分是巨星和亚巨星相互出错, 比如F型亚巨星(F III)有29.5%错误分类成F型巨星(F I), 有10.2%错误分类成F型矮星(F V), G型巨星(G I)有9.95%错误分类成G型亚巨星(G III), G型亚巨星(G III)有13.3%错误分类成G型巨星(G I).

5　结束语

随着大型天文望远镜收集到海量的天文光谱, 传统的模式匹配方法出现了效率低, 标准星少等问题, 使用模式匹配方法对海量天文光谱进行分类变得不切实际. 传统的机

器学习方法也存在对大规模数据训练困难、训练深度不足等问题. 随着深度学习在各个领域的应用越来越广泛, 本文利用深度学习的方法, 在天文学领域进行了应用研究, 优化了一般的深度学习模型, 提出了一种基于卷积神经网络的二元分类模型. 该模型同时兼顾了恒星光谱型和光度型的分类, 不仅避免了对标准恒星库的依赖, 而且可以提取数据中隐藏的深层特征, 最后在LAMOST DR5数据集上进行了验证实验, 实验结果表明, 相对目前的研究结果, 我们的方法能够取得较高的分类准确率. 但是我们的CSSL_CNN模型仍然存在一些不足, 由于每条光谱的特征很多, 导致训练时间较长, 以后可以通过一些方法对特征进行进一步的筛选和压缩. 还可以把天文学家的专业知识, 加入到我们的模型中, 以取得更高的分类准确率.

致谢 LAMOST是由国家发展和改革委员会资助, 由中国科学院承建的国家重大科学工程项目, 并由中国科学院国家天文台负责运行和管理. 感谢审稿人对文章提出的宝贵建议, 使得文章的质量有了显著的提高.

参 考 文 献

[1] Cui X Q, Zhao Y H, Chu Y Q, et al. RAA, 2012, 12: 1197
[2] Zhao G, Zhao Y H, Chu Y Q, et al. RAA, 2012, 12: 723
[3] Luo A L, Zhao Y H, Zhao G, et al. RAA, 2015, 15: 1095
[4] Gray R O, Corbally C J. AJ, 2014, 147: 80
[5] Liu C, Cui W Y, Zhang B, et al. RAA, 2015, 15: 1137
[6] 周志华. 机器学习. 北京: 清华大学出版社, 2016
[7] 李超, 张文辉, 李然, 等. 天文学报, 2020, 61: 21
[8] Li C, Zhang W H, Li R, et al. ChA&A, 2020, 44: 345
[9] Elting C, Bailer-Jones C A L, Smith K W. AIP Conference Proceedings, 2008, 1082: 9
[10] Saglia R P, Tonry J L, Bender R, et al. ApJ, 2012, 746: 128
[11] Folkes S R, Lahav O, Maddox S J. MNRAS, 1996, 283: 651
[12] Bailer-Jones C A L. PASP, 1997, 109: 932
[13] 付文博, 孙涛, 梁藉, 等. 计算机科学, 2018, 45: 11
[14] 仝卫国, 李敏霞, 张一可. 计算机科学, 2018, 45: 155
[15] 张静敏, 马晨晔, 王璐, 等. 天文学报, 2020, 61: 20
[16] Zhang J M, Ma C Y, Wang L, et al. ChA&A, 2020, 44: 334
[17] Fabbro S, Venn K A, O'Briain T, et al. MNRAS, 2018, 475: 2978
[18] Zou Z Q, Zhu T C, Xu L Z. Classification Model for Celestial Spectra Based on Deep Neural Network. Proceedings of the 2019 4th International Conference on Computational Intelligence and Applications. Nanchang: IEEE, 2019: 68
[19] 周飞燕, 金林鹏, 董军. 计算机学报, 2017, 40: 1229
[20] 张顺, 龚怡宏, 王进军. 计算机学报, 2019, 42: 453
[21] Hon M, Stello D, Yu J. MNRAS, 2017, 469: 4578
[22] Liu W, Zhu M, Dai C, et al. MNRAS, 2019, 483: 4774
[23] He K M, Zhang X Y, Ren S Q, et al. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Nevada: IEEE, 2016: 770
[24] 戴加明, 佟继周. 天文学进展, 2018, 36: 384
[25] Chen L, Zhang H W, Xiao J, et al. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 6298
[26] Hu J, Shen L, Sun G. Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 7132
[27] Zou Z Q, Zhu T C, Xu L Z, et al. PASP, 2020, 132: 044503

Classification Model of Stellar Spectral Type and Luminosity Type Based on Convolution Neural Network

HONG Shu-xin[1,2] ZOU Zhi-qiang[1,2] XU Ling-zhe[3]

(1 College of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210023)
(2 Jiangsu Key Laboratory of Big Data Security and Intelligent Processing, Nanjing 210023)
(3 Department of Telescope's New Technology, Nanjing Institute of Astronomical Optics & Technology, National Astronomical Observatories, Chinese Academy of Sciences, Nanjing 210042)

Abstract Star classification is an important topic in astronomy. For the classification of the massive high-dimensional stellar spectral data that has been collected, the pattern matching method is more successful for spectral classification, but its disadvantage is that the differences between standard star templates cannot be reflected in matching actual observed data. Especially when it comes to the classification of both spectral types and luminosity types, the template matching method often fails. Moreover, the classification of luminosity types based on spectral feature measurement strongly depends on the accuracy of spectral fitting. In order to solve the problem of classification based on spectral type and luminosity type, a Classification model of Stellar Spectral type and Luminosity type based on Convolution Neural Network (CSSL_CNN) is introduced. This model uses a convolutional network to extract features of the spectra, adds attention blocks to focus on learning important features, uses a pooling operation for dimensionality reduction, compressing the number of parameters of the model, and the fully connected layer is used to learn features and classify stars. The Large Sky Area Multi-Object Fiber Spectroscopy Telescope (LAMOST) public data set Data Release 5 (DR5) was used in the experiment to verify and evaluate the performance of the model. We used 71282 spectra from DR5, and each spectrum contains more than 3000 features. The experimental results show that the accuracy of our model reaches 92.04% in classification of spectral types, while a Celestial bodies Spectral Classification Model (CSC_Model) based on the deep neural network only reaches 87.54%, and the accuracy of our model is 83.91% in binary classification of spectral and luminosity types, while MKCLASS, a pattern matching method, only has the accuracy of 38.38%, and its efficiency is much lower.

Key words stars: fundamental parameters, methods: data analysis, deep learning: convolution neural network (CNN)