**基于卷积神经网络的天文光谱数据分类方法**

智源杯天文数据算法挑战赛 第7名 剁椒鱼头真不辣

* 李鑫烨 南京大学 18120165301 [161220070@smail.nju.edu.cn](mailto:161220070@smail.nju.edu.cn)
* 曹佳涵 南京大学 15366101383 [cjh327@126.com](mailto:cjh327@126.com)

**摘要** 本次比赛给出近100万个天体的光谱数据，要求选手根据这些光谱数据，把未知天体分成行星（star），星系（galaxy）和类星体（qso）三类。本队方案包括对数据进行归一化、滤波等预处理操作，对少数类别数据进行增强，分类模型经过探索GBDT和卷积神经网络，最终采取通过数据增强、网络结构等维度训练不同神经网络并融合的策略。该方案最终在决赛测试集上达到0.98448的Macro-F1 score，在480支参赛队伍中排名第7名。

**关键词** 天体分类；卷积神经网络；数据挖掘

一、简介

在大规模光学光谱观测和大视场天文学研究方面，郭守敬望远镜（LAMOST，大天区面积多目标光纤光谱天文望远镜）居于国际领先的地位。它的视场和口径规模都居世界领先地位。焦面上有4000根光纤可以同时获得4000个天体的光谱，所以也是世界上光谱获取率最高的望远镜。

作为世界上光谱获取率最高的望远镜，LAMOST 每个观测夜晚能采集万余条光谱，使得传统的人工或半人工的利用模板匹配的方式不能很好应对，需要高效而准确的天体光谱智能识别分类算法。

本次比赛给出近100万个天体的光谱数据，要求选手根据这些光谱数据，把未知天体分成行星（star），星系（galaxy）和类星体（qso）三类。

二、相关背景

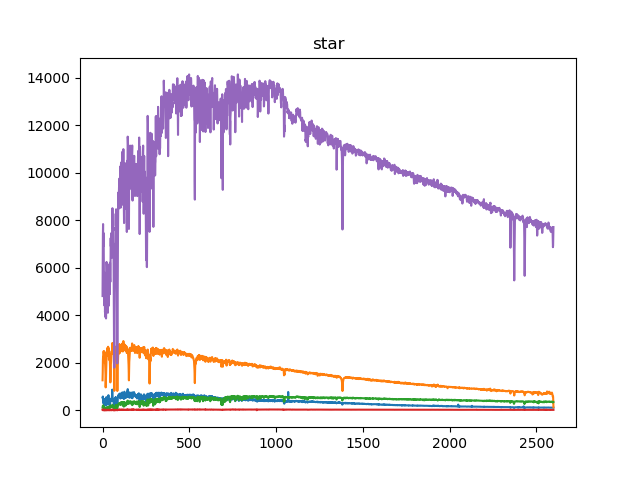
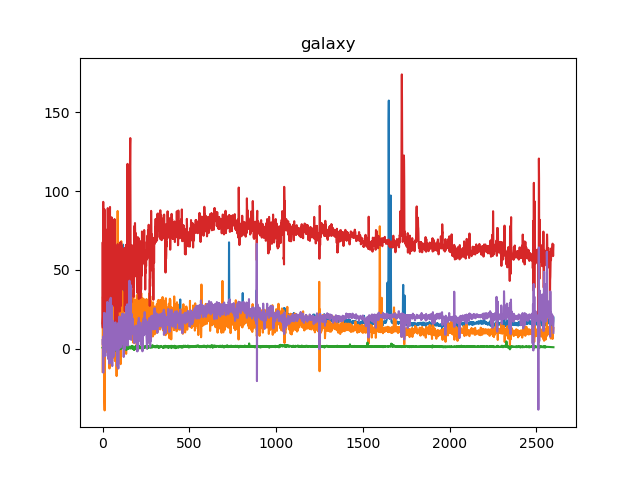
随着科学技术的发展和观测设备不断升级，天文数据呈现爆炸式的增长。人工智能（AI）技术能够辅助天文学家们处理分析海量天文数据，发现新的特殊天体和物理规律。天体光谱数据的智能处理正由传统机器学习方法逐步转向深度学习，主要采用基于计算机视觉的技术手段。 参考文献[1]提出了使用 5 层卷积神经网络估计大气参数的方法。 参考文献[2]提出使用自编码算法的神经网络对斯隆数字巡天（SDSS）光谱进行恒星大气物理参数的估计。 参考文献[3]提出使用深度神经网络模型并构造分类器对光谱进行分类。深度学习方法较机器学习在处理天体光谱数据上的精度、鲁棒性和泛化性都有明显提升。

本文基于光谱数据的特点，对光谱数据进行归一化和滤波等预处理，在分类时将卷积神经网络应用到了该问题上，并对网络结构进行优化，通过模型结构、训练增强方式等多个维度进行模型融合。该方法应用在智源杯天文数据算法挑战赛中，在480支参赛队伍中排名第7名。

三、主要思路

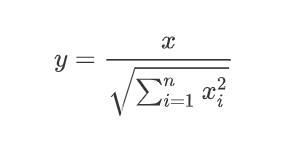
天体光谱数据是天体的点源光经过色散形成的、分布在不同波长下流量强度的序列。根据其吸收线、发射线的位置、强弱、宽度等性质，天文学家可以判断该天体的所属类别。可以说天体光谱的在一维坐标系下的“样子”决定了它的类别，这种“样子”决定“类别”的任务天然适合于卷积神经网络。传统的卷积神经网络预训练模型都是针对二维图像数据构建并训练得到的，并不适用于一维天体光谱数据。 针对光谱数据的特点，我们采用一维卷积神经网络，为了加深网络层数，我们在卷积层后加入Batch Normalization层来抑制过拟合。

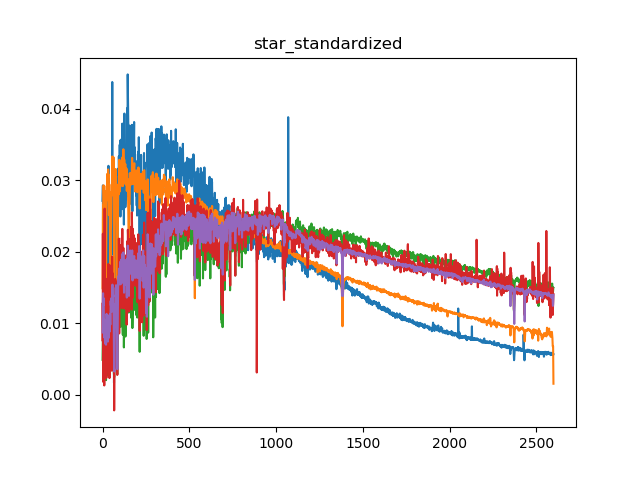
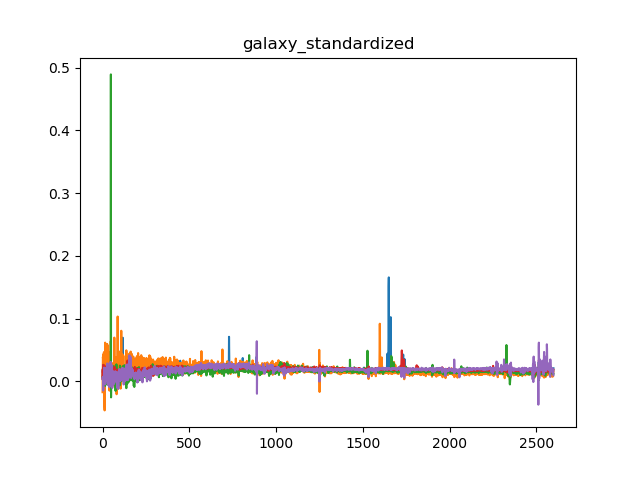
四、数据探索

 每个样本是2600维的光谱序列数据。每个类别随机选取了部分数据进行可视化，以star类和galaxy类为例：

**（一）归一化**

我们发现，不同类别星体的光谱总体上形状呈现出较大的差异，波峰波谷的位置是用来区别类别非常显著的特征，对于同一类别星体的光谱，其绝对大小也相差很大，这是因为我们在观测天体光谱时，它们的亮度由于距离地球远近不同而强弱不同，导致光谱流量量纲（即数量级）存在很大差异。也就是说，同一类型的天体光谱也会因为这些原因使得光谱能量大小差异巨大，从而为光谱分类带来很大难度。因此要对数据进行归一化。

根据参考文献[2]，我们希望模型能够通过学习光谱能量的分布或是光谱上的特征模式来对不同星体分类，所以应当排除这种能量量级带来的影响。与传统输入特征归一化的方式不同，这里采用“流量归一化”，即每个样本根据自己的能量大小进行归一化，而不是在整个数据集上归一化。因此我们采用单位化标准因子进行归一化，公式为：

归一化后的效果（以star类和galaxy类为例）：

归一化前不同光谱的流量不在同一数量级，归一化后不同光谱的流量统一，归一化有利于卷积神经网络更快速地学习不同天体光谱之间的特征差异，使神经网络关注的重点不包含流量因素，有利于神经网络训练速度加快、精度提升。

**（二）主成分降维**

主成分分析可以在保留大部分信息的前提下降低维度，并且过滤掉部分噪声，最终选择750个主成分代替原有2600维度的数据，从而达到降低数据维度，提升计算效率的目的。

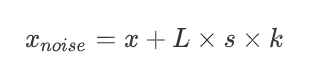
**（三）Savitzky-Golay滤波**

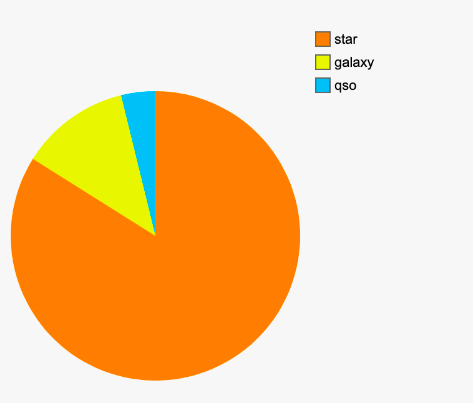
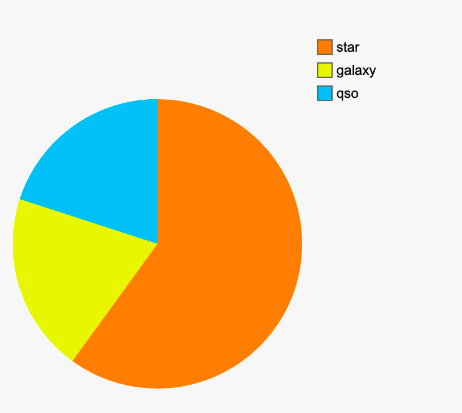
为了平滑噪声，我们对原始光谱数据进行Savitzky-Golay滤波。Savitzky-Golay滤波拟合法是根据时间序列曲线的平均趋势，确定合适的滤波参数，用多项式实现滑动窗内的最小二乘拟合；利用Savitzky-Golay滤波方法（基于最小二乘的卷积拟合算法）进行迭代运算，模拟整个时序数据获得长期变化趋势。

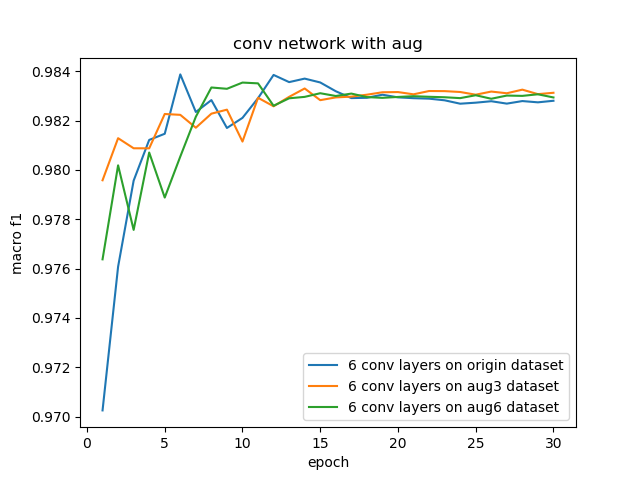
我们发现窗口宽度显著影响滤波效果，较小的窗口宽度能有效过滤噪声，提升预测精度，而过大的窗口会抹除数据的细节，损失有用的信息，造成预测精度的降低。经过小数据集的验证，最终选择窗口宽度为7.

**（四）数据增强**

训练数据中行星（star），星系（galaxy）和类星体（qso）三类数据各占比为83.94%, 12.25%, 3.80%，是一个典型的类别不平衡问题，行星的数量远多于其他两类。不平衡的数据会使卷积神经网络的分类效果下降，因此数据增强是非常必要的。我们通过在星系、类星体数据上增加具有物理意义的噪声进行过采样扩充数据，从而增加星系和类星体所占比例。

参考了阿里云天池天文数据挖掘大赛中银河护卫队的数据增强方案，对原始数据加入不同幅度的正态分布随机数，公式为：

其中x\_noise是加噪的光谱数据，x是原始的光谱数据，L是标准正态分布随机数，s是原始光谱的标准偏差，k是实验得到的折减系数，我们最终取值为k=0.2。

原始训练数据是一个明显的类别不平衡问题。因此在数据增强时，我们有意针对小类别（qso和galaxy）增强更多的样本，来平衡各个类别的数量。上图战术了原始数据集和增强后数据集各类别占比情况，最终增强后三个类别的数量比约为3:1:1，经过实验发现，如下图所示，在单CNN模型上使用增强后的数据集进行训练相较于原始数据集训练过程更加稳定，波动更小，最终收敛后的macro-F1 score得到提升。使用数据增强的另一个优势是模型融合时对不同增强比例的数据集训练可以提供多种差异化的模型，从而显著提升模型融合的效果。

**（五）特征工程**

对预处理后的数据提取出天文学上有意义的特征，具体来说，通过区间统计，提取出每个区间内光谱数据的均值（mean）、方差（variance）峰度（kurtosis）、偏度（skewness）、最小值所在位置（argmin）、最大值所在位置（argmax）、最小值（min）和最大值（max）8个特征，选取窗口大小为50，从而将2600维的光谱数据转化为416维的特征。

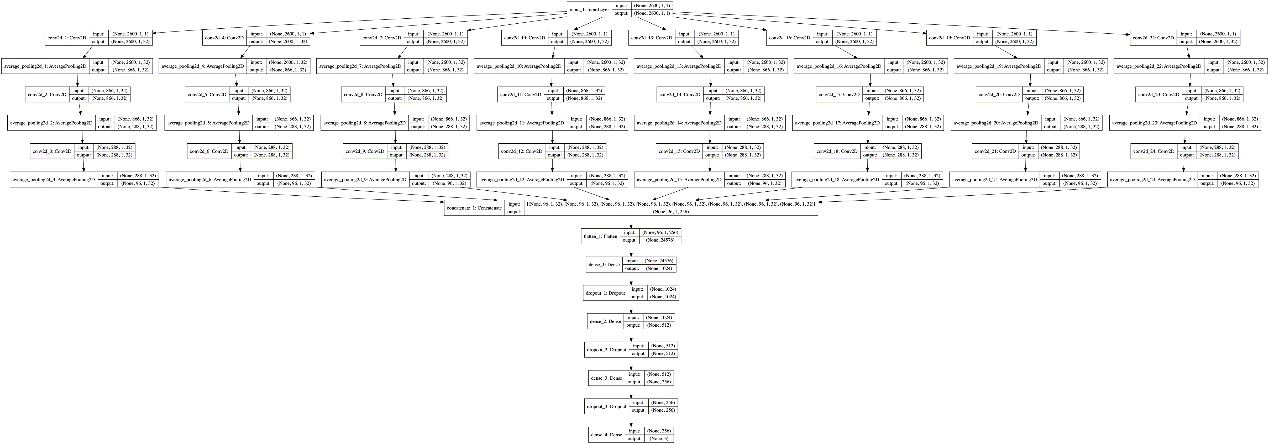
五、实验和试错过程

关于分类模型的选择，我们先后对两种主流模型进行了相关的实验和探索，分别是GBDT梯度提升决策树和CNN卷积神经网络。

**（一）GBDT探索**

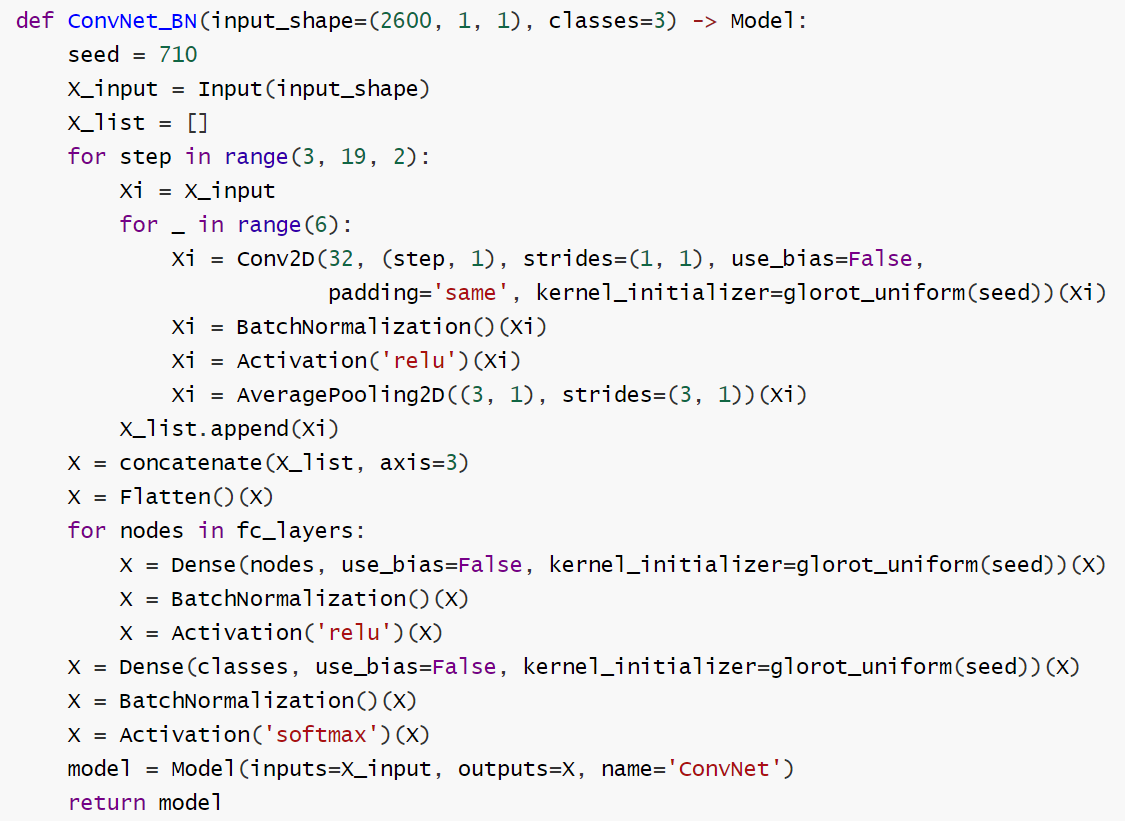
最初的模型采用GBDT梯度提升决策树，使用的是LightGBM开源实现库，它可以更加高效地并行计算。将前述的PCA降维得到的750维特征和特征工程构建的416维特征进行拼接，供给LGB进行训练，最终线下验证集Macro-F1 score可以达到0.9756。

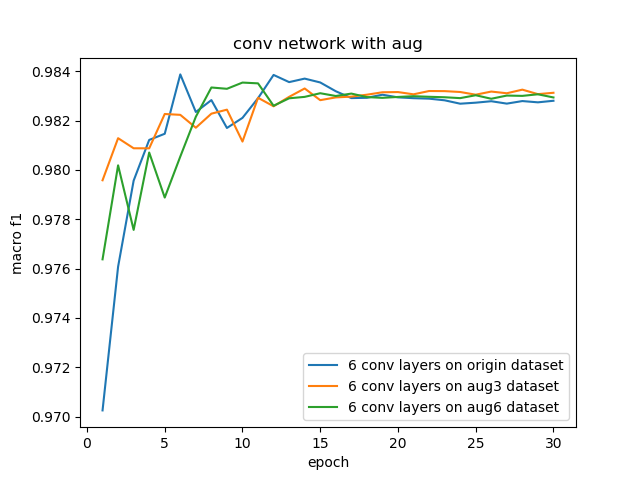
**（二）神经网络探索**

我们对光谱数据归一化之后进行简单可视化发现，不同类别的光谱数据在趋势上具有较明显的区别。考虑光谱数据具有的局部性，采用一维卷积神经网络模型，取得了较好的效果。为提取光谱数据中具有的不同尺度的特征，模型采用[3, 5, 7, 9, 11, 13, 15]不同大小的卷积核进行卷积，并拼接为一维向量输入全连接层。

考虑过拟合问题，我们先后采用了DropOut和Batch Normalization，实验DropOut层时发现验证集表现依赖于DropOut层中dropout rate该参数，为避免过多超参数难以寻找最佳组合，我们改用Batch Normalization，模型收敛速度明显加快，效果得到有效提升。

查阅卷积神经网络有关文献发现，选择不同抽象层次的特征将显著影响模型的特征。我们尝试将模型中的三层卷积调整为六层卷积之后效率略有提升。猜想不同抽象层次的特征之间可能也存在相互关系，而简单的卷积神经网络不具备使用不同层次的特征的能力，可能采用ResNet或者DenseNet会取得更好的效果。

在验证集上取得最佳效果的网络结构如下：

该网络模型在原数据集上和数据增强后的数据集上训练后，验证集上macro f1随epoch变化如图：

六、最终采用的方案

经过对GBDT和神经网络的分类效果进行线下验证和实验，尝试了不同的模型融合策略后，我们最终舍弃了GBDT，选择从模型结构、训练增强等维度训练不同的神经网络，进行模型融合，融合方法是对各模型的预测结果softmax score求平均。我们组对卷积层数、Dropout / Batch-Normalization 层，全连接层数，batch size等参数的设置进行了大量实验，并选择其中七个表现较好的模型进行融合，macro f1 score最高为0.98448。

七、结果

最终我们的方案在决赛测试集上达到0.98448的Macro-F1 score，在480支参赛队伍中排名第7名。

八、经验分享

我们根据天文光谱数据的特点，设计相应的预处理方案和网络结构，并尝试了不同的方案组合，采用了基于传统方案的GBDT以及基于深度神经网络的方案，都取得了不错的效果。通过对原始数据进行不同程度噪声和不同类别比例的增强，我们训练出来的模型具有较强的鲁棒性，且表现出来较好的性能，线下和线上的得分基本保持一致，线下结果略好于线上结果。

我们的方案也仍存在改进的地方，可以设计Macro-F1 score 的近似可微函数，直接应用到神经网络中，网络结构也可以继续调优，采取更加深层的网络结构。数据预分析方面做的也还不够，可以加强数据清洗和缺失值的处理。

参考文献

1. 潘儒扬,李乡儒.基于深度学习技术的恒星大气物理参数自动估计[J]. 天文学报, 2016, 57(4)

2. 韩帅,李悦.基于BP神经网络（自编码）的恒星大气物理参数估计[J]. 自动化与仪器仪表, 2016(9):230-231

3. 刘真祥,荣容,许婷婷,等.基于深度信念网络的天体光谱自动分类研究[J]. 云南民族大学学报（自然科学版）,2017(2)

4.天文数据挖掘大赛\_银河护卫队\_梯度提升决策树部分 <https://tianchi.aliyun.com/forum/postDetail?spm=5176.12586969.1002.18.7e032c85kbYOQi&postId=5066>

5. 王奇勋,赵刚,范舟.基于DenseNet的天体光谱分类方法

6. 李乡儒,刘中田,胡占义,吴福朝,赵永恒.巡天光谱分类前的预处理 ———流量标准化