

基于残差神经网络的恒星-星系分类器^{*}

杨 阳¹⁾ 文中略²⁾ 夏俊卿^{1)†}

(1)北京师范大学天文系, 100875, 北京; (2)中国科学院国家天文台, 100101, 北京)

摘要 使用残差神经网络(residual neural network, RNN)算法对斯隆数字巡天(Sloan digital sky survey, SDSS)提供的天体伪彩色图片进行分类, 直接从图像中获得特征. 使用带有光谱信息的星系与恒星图片作为训练集和测试集. 经过训练, 在测试集上的准确率达到 98.23%, 召回率达到 98.80%. 这表明: RNN 可以实现对星系和恒星图像的精确分类, 分类器给出的恒星-星系概率是有效的, 可用于分类可靠性评估; 还可以尝试将此分类器应用到未来巡天中, 进一步测试其性能.

关键词 巡天; 深度学习; 分类; 恒星; 星系

中图分类号 P152

DOI: 10.12202/j.0476-0301.2021106

0 引言

对观测到的恒星和星系进行分类, 一直是天文学研究的一个基础问题. 早期分类研究是通过天体的形态特征和光谱信息, 现代的多波段巡天项目中, 使用复杂的颜色条件来分离星系和恒星. 例如 Scranton 等^[1] 提出了基于贝叶斯的分类器, 使用点扩散函数星等和模型星等之差定义了聚集度, 将一个天体是星系的概率和星等、聚集度、观测时视宁度等条件联系起来.

对于大数据处理, 机器学习优势显著. 机器学习算法在数据中自动学习, 根据真实标签不断优化模型, 使得模型的输出不断接近真实值. 一系列机器学习方法被应用到恒星-星系分类中, 如 k -近邻算法^[2]、决策树^[3]、支持向量机^[4]、集成学习^[5]、人工神经网络^[6]、组合分类器^[7]等. 这些分类器与传统的方法类似, 都依赖于人工的选择和构建特征. 这些特征是斯隆数字巡天(Sloan digital sky survey, SDSS)从天体图像中经过分析得到的, 例如天体的星等、大小、形状等. 手工构建的特征并非完全准确, 其中的误差会影响到恒星-星系分类的效率. 例如某些双星重叠的情况下, 给出光源的尺度就会变大. 受到前景恒星的影响, 星系的星等也会出现误差. 相比于通过天体特征, 这些特殊情况使用天体的图像更容易被区分出来.

深度学习也是机器学习的一个分支, 神经网络(neural network, NN)的层数比传统机器学习模型更

多. 这些模型中的参数不是手工构建的, 而是直接从原始数据中获得的. 深度学习模型将低层次的特征抽象, 得到更高层次的特征, 从而学习到不同层次的特征. 例如, 使用图片输入深度学习模型, 第 1 层网络可能会学习到图像中物体的边缘位置和方向等信息; 第 2 层则可能会对这些边缘进一步抽象, 组合成更复杂的图形. 后面的层也会对前面的信息进行不断抽象, 最终得到图像的分类或者其他信息. 卷积神经网络(convolutional neural network, CNN)就是一个深度学习模型, 常用于图像处理. CNN 在天文学中得到广泛应用, 例如星系的形态学分类^[8]、搜寻强引力透镜^[9]、SDSS 巡天中天体分类^[10]、估计星系测光红移^[11]等. 为了进一步提高 CNN 的层数, He 等^[12] 提出了残差神经网络(residual neural network, RNN). RNN 比 CNN 性能更好, 被应用到天文数据处理中, 例如从形态上对星系进行分类^[13-14], 通过识别星系的 Lyman- α 射线来寻找引力透镜系统^[15]. 除了天体分类外, Wu 等^[16] 还使用 RNN 对星系的金属丰度进行预测. 目前, RNN 算法还没有被用于恒星-星系分类问题上.

Pytorch 是一个基于 Python 的 NN 机器学习库, 可以自动将误差反向传播, 还提供了许多常见的 NN 模型. 使用这个库, 可以方便地构建并训练机器学习模型^[17]. 本研究修改了 RNN 18 的输出部分, 使模型适应恒星-星系分类问题. 然后, 使用 SDSS 提供的天体伪彩色图片对模型进行训练, 并且在不同星等的区间上测试了模型性能.

^{*} 国家自然科学基金-天文联合基金重点资助项目(U1931202)

[†] 通信作者: 夏俊卿(1983—), 男, 博士, 教授. 研究方向: 宇宙学的理论模型分析和实验数据处理. E-mail: xiajq@bnu.edu.cn

收稿日期: 2021-04-30

1 算法

在图像分类问题上, CNN 有着很好的表现. 相比于传统的全连接 NN, CNN 增加了卷积层、池化层. 卷积层使用一个卷积核, 通过离散卷积对图像局部的细节进行提取. 池化层是一种降采样操作, 将图像中的一个区域映射到下一层的 1 个点. 根据映射方法不同, 分为最大池化、平均池化等. 图像中的数据量很大, 信息会有冗余, 降采样可以减少后续处理的数据和计算量^[18]. CNN 中的每一个卷积层都从图像中抽取局部细节, 逐步提取更高层级、更加抽象的特征, 因此 CNN 具有很好的特征识别能力, 并且面对目标在图片中的移动和转动也能保持良好的预测性能.

一般来说, NN 层数越多, 网络的性能也会越好. 但是, 网络过深可能会出现梯度消失等问题, 使训练变得困难. 可以通过数据预处理和归一化“层”来缓解这些问题, 使得较深的网络也可以收敛. 然而, 随着网络深度的增加, 会出现退化问题, 训练准确率会逐渐达到饱和, 然后不断降低^[19].

为了解决深度网络训练困难的问题, He 等^[12]提出了 RNN. 这个模型不仅增加了网络的深度, 而且训练速度快, 性能比普通的 CNN 更强, 如图 1 所示. 实现这些特性主要靠残差块(residual block, RB)结构, RB 中不仅有顺序排列的卷积层, 还通过与卷积层并列的捷径连接(shortcut connection, SC)跳过了一些卷积层. SC 可以将输入 RB 的数据 x 直接传递到输出, 与通过卷积层的输出 $F(x)$ 相加; RB 的输出为 $F(x) + x$. 在训练过程中, 可以通过 SC 将误差无损反向传播, 解决了梯度消失的问题.

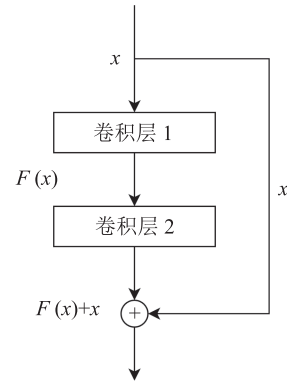


图 1 残差块的结构

RNN 网络有多种形式, 如按照其网络层数, 有 RNN 18、34、50、101、152 等模型. SDSS 天体图像较小(128 像素×128 像素), 所选天区中的数据量有限, 过深的模型可能会造成过拟合. RNN 18 可以较好地完成这个分类任务. 表 1 是本文所使用的 RNN 结构, 卷积层参数中, 卷积核的尺寸代表了局部采样的区域大小, 通道数则代表同时进行操作的卷积核数目, 每个独立的卷积核可从图像中获取不同的特征. 池化层参数中, 池化核代表了对图像降采样之后, 每个像素区域对应原始图片的大小, 步长代表 2 次相邻采样间隔的像素数. 每个卷积组由 2 个 RB 组成, 每个 RB 包含 2 个卷积层, 方括号中给出了它们的卷积核大小和通道数.

输入样本经过所有卷积层后, 模型使用带有自适应池化核的平均池化层, 将输出由 $4 \times 4 \times 512$ 的数组转化为长度 512 的向量. 为了使模型输出符合本研究的需求, 修改了原始 RNN 18 中的全连接层, 使用线性整流函数(rectified linear unit, RLU)激活.

$$f(x) = \max(0, x) = \begin{cases} x, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

这个函数不仅能对 NN 做非线性激活, 还能提高网络

表 1 本文使用的 RNN 结构

层名称	输出	层参数
输入层	128×128×3	SDSS 彩色图片
卷积层 1	64×64×64	卷积核 7×7, 通道 64
最大池化层	32×32×64	池化核 3×3, 步长 2
卷积组 2	32×32×128	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 2$
卷积组 3	16×16×128	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 2$
卷积组 4	8×8×256	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 2$
卷积组 5	4×4×512	$\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 2$
平均池化层	1×1×512	自适应池化核
全连接层 1	10	RLU 激活
全连接层 2	2	输出分类结果

的训练速度.

输出层包含 2 个神经元、输出样本与恒星和星系符合程度的评分, 即 s_1 和 s_2 . 评分越大, 说明该样本更符合这个类别的特征. $s_1 > s_2$, 模型将样本标记为恒星, 反之则为星系. 评分并不是归一化的, 即不代表样本属于此分类的概率. 对评分使用 Softmax 函数进行归一化处理:

$$P_2 = \frac{\exp s_2}{\exp s_1 + \exp s_2}.$$

给出样本是星系的归一化概率 P_2 , 则样本是恒星的概率 $P_1 = 1 - P_2$. 评分上的分类阈值可对应到概率值: 当把样本标记为星系时, 概率 $P_2 \geq 0.5$; 标记为恒星时, 则 $P_2 < 0.5$. 后面会测试这一概率值的有效性.

本研究使用了交叉熵损失函数衡量模型的训练误差:

$$L(y, \hat{y}) = -\frac{1}{N} \sum_{i=1}^N (\hat{y}_i \lg y_i + (1 - \hat{y}_i) \lg (1 - y_i)).$$

式中 N 代表输入模型的样本数目, y_i 、 \hat{y}_i 分别代表每个样本的预测值和真实值. 当某个样本标签为 1 时, 这个样本的损失为 $l_i = -1/N \cdot \lg y_i$, 即: 样本预测值越接近 0 损失越大; 相反地, 当标签为 0 时, y_i 越接近 1 损失越大, 这样就惩罚了远离真实值的预测.

训练就是优化模型参数, 来降低损失函数的过程. 使用动量梯度下降算法, 动量值设置为 0.97. 进行分批训练, 每个批次数据量为 64 个天体样本, 学习率设置为 10^{-4} .

2 数据准备

SDSS 巡天项目, 覆盖了 $14555(^{\circ})^2$ 的天区, 超过了整个天空的 1/3. 对 u 、 g 、 r 、 i 、 z 这 5 个波段进行了测光, 给出了超过 3 亿个恒星和星系的测光数据、 r 波段极限星等. 此外, SDSS 还对超过 300 万个恒星和星系进行了光谱观测^[20]. 这些数据可以在 SDSS 网站上通过 CasJob 公开获取^[21].

本研究选定了北银冠赤经 $150^{\circ} \sim 160^{\circ}$ 、赤纬 $0^{\circ} \sim 20^{\circ}$ 的一块区域作为训练用的天区. 此天区中, 共有恒星 12926 个, 星系 46905 个. 排除光谱红移有异常的数据, 还剩下恒星 10873 个, 星系 44371 个. 图 2 展示了天体样本所覆盖的红移范围. 其中: 恒星样本的光谱红移分布在 0 附近, 进行均匀分组; 星系的光谱红移范围较大, 使用对数间隔均匀分组. 排除了异常数据之后, 样本中恒星和星系的红移分布在合理范围内, SDSS 对于这些样本给出的类别是可靠的, 它们的类别标签被用于模型训练和测试过程中.

图 3 展示了这些恒星和星系 r 波段星等的分布. 恒星的红移和星等都围绕着 1 个峰值分布, 而星系的红移和星等分布中出现了 2 个峰值, 这是由 SDSS 拍摄光谱的选择性所造成的.

选取合适的样本之后, 还需要获取恒星和星系样本的图像. 处理天体原始图像时, 一般选用单一波段图像, 同时考虑视宁度、曝光时间等信息. 对于恒星-

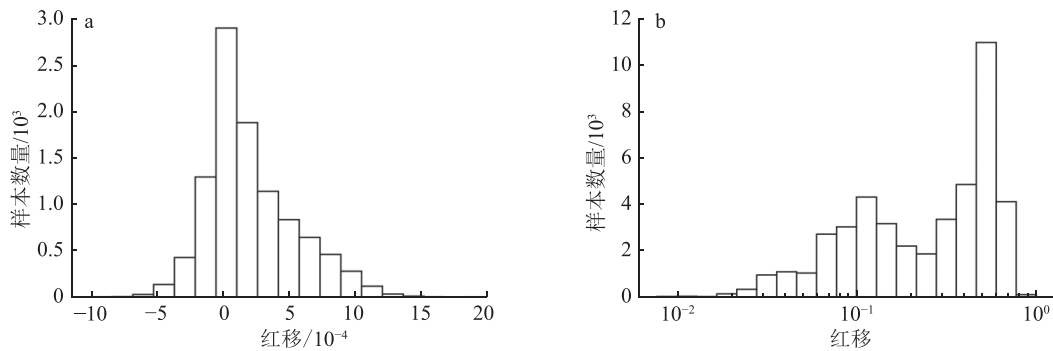


图 2 恒星 (a) 和星系 (b) 数据的红移分布

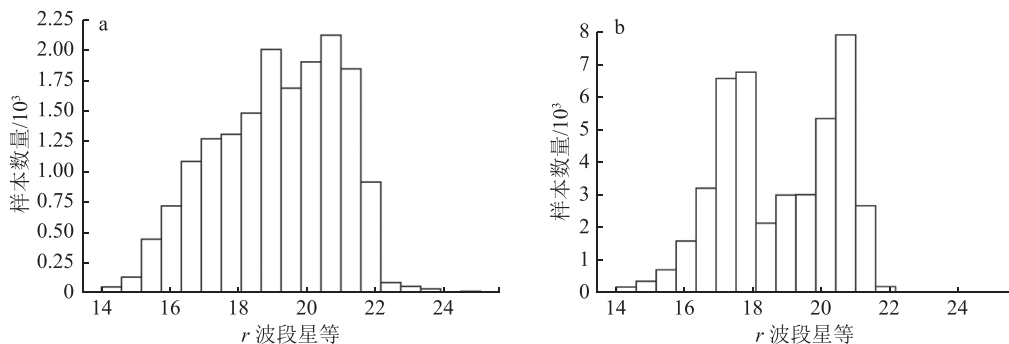


图 3 恒星 (a) 和星系 (b) 数据的 r 波段星等分布

星系分类问题, 采用多波段图像. 这样除了利用各波段亮度、轮廓等信息外, 还能获得天体的颜色信息. 本研究选取的星系红移几乎都 <0.8 , 这些星系的主要能谱特征集中在 g 、 r 、 i 这 3 个波段. SDSS 网站上提供这 3 个波段合成的伪彩色图片, 以方便研究者查看天体图像^[22]. 在研究中, 将所选样本对应的伪彩色图片输入模型训练, 这样可以使使用到天体的能谱信息, 也方便观察样本的特征.

观察下载的图片样本, 包含了各种颜色和星等的恒星, 样本示例如图 4 所示. 其中, 恒星 4 是双星. 双星因为光源尺度较普通恒星更大, 在没有光谱的数据集中可能被误分类成星系. 使用这些多样化颜色和形态的样本训练模型, 其可靠性和泛化能力更好.

图 5 是星系图片示例, 星系样本数目明显多于恒星. 为了使模型不偏向于某个类别, 训练中使用的 2 种样本数目尽量平衡. 按照恒星图片的个数对星系图片进行随机抽样, 用于训练和测试过程, 恒星则使用全部样本. 然后对取得的恒星、星系样本再次进行随机取样, 80% 的天体图片加入训练集, 20% 图片加入测试集. 经过抽样, 训练集中共有 17396 个天体样本, 测试集中共有 4350 个样本. 这里测试集仍按恒星与星系 1:1 取样, 测试结果反映出模型对一个未知天体分类的性能, 而不假设其更有可能是哪种天体.

3 训练与分析

3.1 衡量分类模型的指标 训练好一个机器学习模型之后, 还要测试模型的性能. 除了损失函数之外, 还需要对模型进行更详细地评估. 对于二分类问题,

可将 2 种类别记为正类和负类. 针对输入输出的类别, 共划分为 4 种情况.

真阳性 (true positive, TP): 样本被分类器判断为正类, 实际为正类. 真阳性数目记为 N_{TP} .

假阳性 (false positive, FP): 样本被分类器判断为正类, 实际为负类. 假阳性数目记为 N_{FP} .

真阴性 (true negative, TN): 样本被分类器判断为负类, 实际为负类. 真阴性数目记为 N_{TN} .

假阴性 (false negative, FN): 样本被分类器判断为负类, 实际为正类. 假阴性数目记为 N_{FN} .

由此, 可以定义以下性能评价指标^[23]:

准确率 (accuracy, Y_a): 指模型预测正确的样本所占的比例, 计算式为

$$Y_a = \frac{N_{TP} + N_{TN}}{N_{TP} + N_{TN} + N_{FP} + N_{FN}}.$$

精确率 (precision, R_p): 衡量被识别为正类的样本中真正为正类的比例, 计算式为

$$R_p = \frac{N_{TP}}{N_{TP} + N_{FP}}.$$

召回率 (recall, E_r): 衡量在所有真实值为正类的样本中被正确识别为正类的比例, 计算式为

$$E_r = \frac{N_{TP}}{N_{TP} + N_{FN}}.$$

F_1 是 R_p 和 E_r 的调和平均数, 可以兼顾精确率和召回率, 即

$$F_1 = \frac{2}{R_p^{-1} + E_r^{-1}} = \frac{2N_{TP}}{2N_{TP} + N_{FN} + N_{FP}}.$$

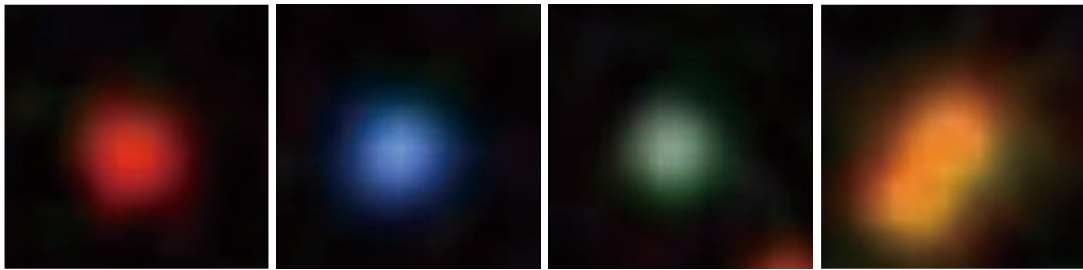


图 4 恒星样本示例

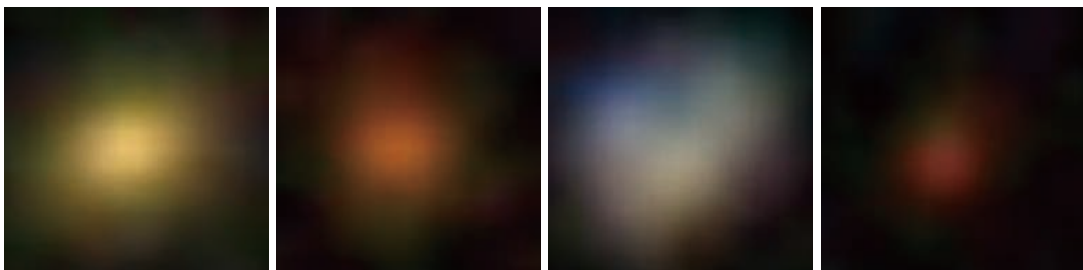


图 5 星系样本示例

C_{au} 定义为接收者操作特征曲线(receiver operating characteristic curve, ROC 曲线)下的面积^[24]. ROC 曲线是一种显示分类模型在所有分类阈值下的性能图表. x 轴为假阳性率 (R_{FP}), y 轴为真阳性率 (R_{TP}), 其中 $R_{TP} = \frac{N_{TP}}{N_{TP} + N_{FN}}$, $R_{FP} = \frac{N_{FP}}{N_{FP} + N_{TN}}$. 使用不同分类阈值对样本进行划分, 就得到一系列 R_{TP} 和 R_{FP} , 绘制出 ROC 曲线. 因为 R_{TP} 、 R_{FP} 均为比率, 量纲为一, 因此 C_{au} 也是一个量纲为一的数, 且 $C_{au} \in [0, 1]$.

图 6 中, 曲线上的任何一点, 代表在某种分类阈值下, 模型可以达到的 R_{TP} 和 R_{FP} . 图中虚线是一个随机猜测的分类器, $C_{au}=0.5$, 模型没有预测价值; 实线代表一个有效的分类器, $C_{au}=0.79$, 阴影部分是 C_{au} 代表的面积, C_{au} 越大, 模型分类的性能就越好. 例如图中实线代表的模型可以达到比虚线更高的真阳性率和更低的假阳性率. $C_{au}=1$ 代表完美的分类器, 此时存在合适的分类阈值, 使得 $R_{TP}=1$, $R_{FP}=0$, ROC 曲线可取到左上角的点. 上面提到的各种指标, 都衡量在特定分类阈值下模型的性能, C_{au} 则兼顾了各种分类阈值的性能.

3.2 训练和模型选取 使用表 1 中定义的 RNN 模型进行训练, 训练的循环次数为 20 次. 由图 7 可以看出, 损失函数值呈先快速后缓慢下降趋势. 损失函数给出的是一个量纲为一的距离, 用来衡量模型的表现. 为了防止过拟合, 在选择模型时主要参考模型在测试集上的表现. C_{au} 综合考虑了各种可能阈值下的 Y_a 和 E_r , 使用这个指标来衡量模型是比较全面的. 图 8 是训练的循环次数、训练集和测试集 C_{au} 的变化情况.

可以看到, 在前 3 个循环中, 测试集 C_{au} 比训练集要高, 这是因为训练的过程是分批进行的. 在每一个小批次 ($N=64$) 中, 模型都进行参数更新. 训练集 C_{au} 是对训练过程 C_{au} 的平均, 而测试是在每个训练循环之后进行, 得到的 C_{au} 是对每个循环最终模型的评价. 从第 4 个循环之后, 训练集的 C_{au} 就超过了测试集, 并且持续增加. 测试集上的 C_{au} 整体趋势也逐渐增加, 在第 17 个循环时取得最大值. 选取该循环输出的模型作为恒星-星系分类器, 因为这个模型在测试集上的表现是最好的. 使用上章提到的性能指标衡量模型, 在训练集和测试集上的表现如表 2 所示.

在测试集上, Y_a 达 98.28%, E_r 达 98.80%, 整体上分类准确. 图 9 中实线为模型的 ROC 曲线, 曲线下面积 $C_{au}=0.9979$. Kim 等^[10] 基于 CNN 的恒星-星系分类器, 在 SDSS 数据集上测试时 $C_{au}=0.9948$. 相比之下, RNN 模型的 C_{au} 分数略高.

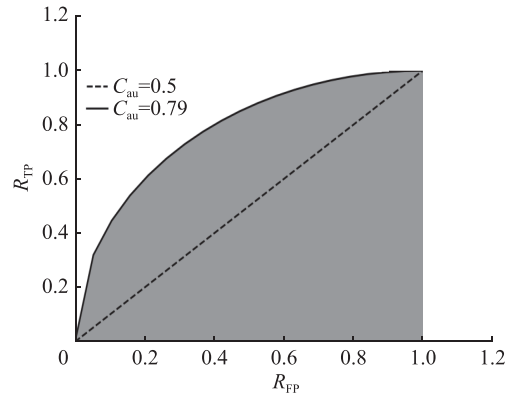


图 6 ROC 曲线和曲线下面积 (C_{au}) 示意

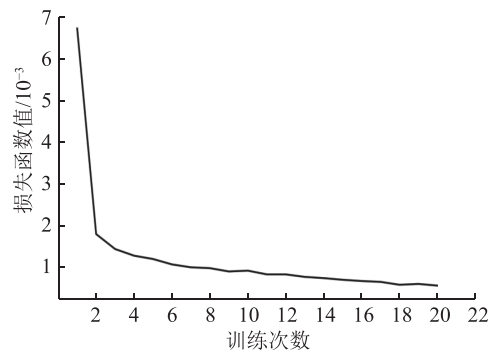


图 7 损失函数值随着训练循环次数的变化曲线

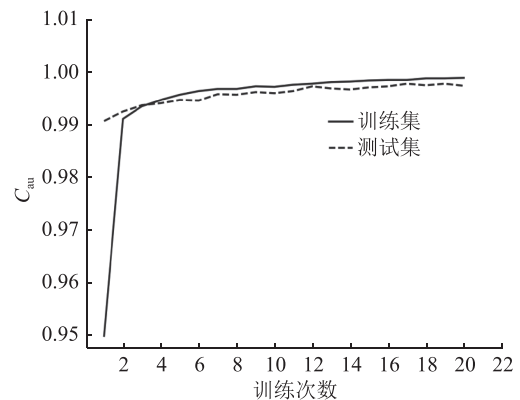


图 8 训练集和测试集 C_{au} 随着训练循环次数的变化曲线

表 2 所选模型的性能指标

指标	训练集	测试集
$Y_a/\%$	98.71	98.23
$R_p/\%$	98.48	97.68
$E_r/\%$	98.94	98.80
F_1	0.9871	0.9824
C_{au}	0.9986	0.9979

将测试集按照 r 波段星等分成 8 个区间(14~22). SDSS 巡天的极限星等约为 22 等, r 波段星等 > 22 的数据很少. 数据量太小, 无法衡量 Y_a 和 E_r , 因此舍弃

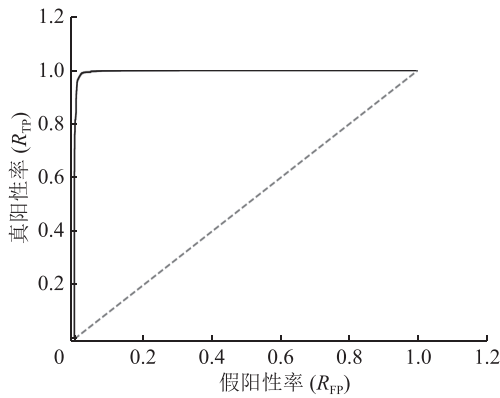


图 9 所选模型的 ROC 曲线 (测试集)

掉这些数据. 衡量这些区间上的 Y_a 和 E_r , 如图 10 所示. 可以看出: 在 14~20 区间段, 比较亮的天体, 模型 Y_a 、 E_r 均 $>90\%$, 甚至接近 100% , 性能很好; 对于星等 >20 的暗天体, 模型性能下降.

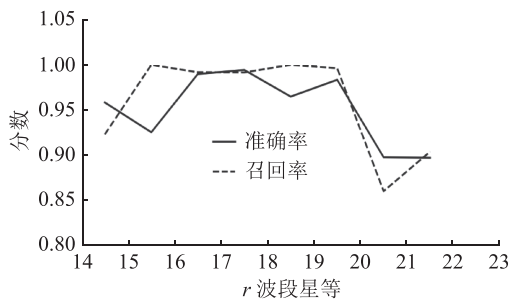


图 10 不同星等区间的模型性能变化 (测试集)

遥远的星系较暗, 更接近于点源, 与暗的恒星难以区分. 对于使用机器学习算法构建的分类器, SDSS 数据集中星等在 21 等附近的天体分类, 一直是个疑难问题. Vasconcellos 等^[3] 使用 SDSS 有光谱数据集, 测试了一系列决策树算法对恒星-星系的分类性能. 其中在最暗天体区间 ($20.5 < r < 21.0$), 这些算法给出的星系 $E_r \geq 42.29\%$, 表现最好的功能树 (FT) 算法为 74.04% . 李超等^[25] 使用集成学习算法, 也用相同的数据测试了分类性能, 其中极端提升树模型的 $E_r = 80.75\%$, 堆叠算法的 $E_r = 84.42\%$. 作为对比, RNN 模型在该星等区间的星系 $E_r = 83.26\%$, 接近最佳集成学习算法的 E_r .

第 1 章提到, RNN 输出值为样本是星系的概率. 若概率接近 0 或者 1, 代表样本非常有可能是恒星或星系; 若概率接近 0.5, 则代表模型对于样本的类别不确定. 图 11 绘制出了测试集样本对应的输出概率分布. 其中, 绝大部分样本都在 0 或 1 附近. 为了显示出中间区域的少量样本, 图中 y 轴设置为对数尺度. 在每个概率区间, 柱的整体高度代表此区间的所有样本, 其中灰色部分代表被分类正确的样本, 白色部分代表被分类错误的样本. 可以看出: 在 0 和 1 附近, 几

乎全部分类正确; 在中间区域, 分类的错误率较高. 这个结果符合预期, 说明了输出概率的有效性. 在模型的实际应用中, 可以使用输出概率来评估模型对分类的确信程度.

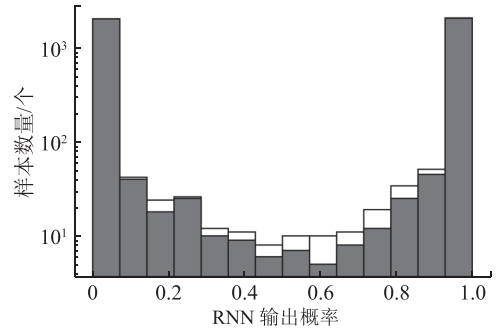


图 11 RNN 输出概率的分布 (测试集)

3.3 模型的应用 为了测试模型在 SDSS 测光数据中的性能, 使用 hyper supprime-cam (HSC) 数据作为天体类别标签. HSC 项目是利用 8.2 m 口径的昴星团望远镜 (subaru telescope) 进行的测光巡天项目. 第 2 次公开数据发布包括了 174 个夜晚观测的数据. 在 5 个波段 (g, r, i, z, y) 进行长时间曝光, 其中 g, r 波段曝光 10 min, i, z, y 波段曝光 20 min. 这为巡天提供了高品质的天体图像, 将天体准确分为点源和面源^[26].

将 HSC 深场数据和 SDSS 测光天体数据进行交叉匹配, 得到应用集. 其中 SDSS 数据提供天体的图像 (与训练集、测试集保持一致), HSC 深场数据提供天体的类别. 从应用集中, 随机抽取 986 个面源、985 个点源对模型进行测试.

使用 SDSS 测光巡天的应用集来测试模型, 模型 $Y_a = 86.93\%$, $R_p = 84.60\%$, $E_r = 90.26\%$; $C_{au} = 0.9342$, $F_1 = 0.8734$. 对比测试集, 各项指标均有所下降. 这个结果一方面是因为天体的星等分布不同. 相比于测试集中天体的星等分布 (图 3), 应用集样本天体星等分布 (图 12) 更加集中在暗天体, 这就导致了在模型评分中暗天体所占的权重更大. 图 10 中不同星等下模型性能测试的

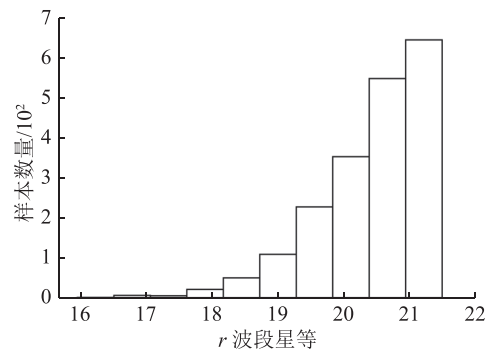


图 12 应用集天体的星等分布

结果说明, RNN 在暗天体上的表现更差, 因此在应用集模型的整体评分较低。

另一方面, HSC 给出的分类只有点源和面源。点源为恒星, 面源大部分为星系。除了星系外, 面源还包括其他天体, 例如行星状星云、原恒星盘、超新星遗迹、彗星等。这些天体的亮度、轮廓、颜色等特征与星系有差异, 也会导致模型在应用集上性能下降。对于应用集的测试, 给出了模型在测光数据上的表现。若将模型应用到测光样本上, 仍需进一步提升模型在暗天体上的分类性能, 并且增加类别以排除恒星和星系之外的样本。

4 结束语

本研究使用了 RNN 算法, 输入 SDSS 的星系和恒星图像进行训练, 构建了恒星-星系分类器。对比已有的 CNN 模型, 取得了更高的 C_{au} 分数。使用测试集还验证了输出概率的有效性, 可以使用输出概率来衡量模型分类的可靠程度。此外, 还测试了不同星等区间的分类性能: 模型对星等 < 20 的亮天体分类足够准确; 对星等 > 20 的暗天体, 分类性能有所下降。RNN 在 SDSS 暗天体上的 E_r 超过了各种决策树算法, 接近较好集成学习模型的表现。

将训练好的模型应用在 SDSS 测光数据集上, $Y_a=86.93\%$, $E_r=90.26\%$ 。应用集的样本整体更暗, 相比测试集, 分类性能有所下降。要将 RNN 应用到巡天数据处理工作中, 仍需提高模型在暗天体上的表现。未来的巡天项目将会带来更多的、更高精度的恒星和星系数据, 而且在暗星等范围上会有更多的样本。应用到未来的巡天数据中, 预期 RNN 分类器的性能相比于使用 SDSS 数据集会得到提高。

5 参考文献

- [1] SCRANTON R, JOHNSTON D, DODELSON S, et al. Analysis of systematic effects and statistical uncertainties in angular clustering of galaxies from early Sloan digital sky survey data[J]. *Astrophysical Journal*, 2015, 579(1): 48
- [2] BAQUI P O, MARRA V, CASARINI L, et al. The miniJPAS survey: star-galaxy classification using machine learning[J]. *Astronomy and Astrophysics*, 2021, 645(A87): 1
- [3] VASCONCELLOS E C, CARVALHO R R, GAL R R, et al. Decision tree classifiers for star/galaxy separation[J]. *The Astronomical Journal*, 2011, 141(6): 189
- [4] FADELY R, HOGG D W, WILLMAN B. Star-galaxy classification in multi-band optical imaging[J]. *Astrophysical Journal*, 2012, 760(1): 15
- [5] MORICE-ATKINSON X, HOYLE B, BACON D. Learning from the machine: interpreting machine learning algorithms for point- and extended-source classification[J]. *Monthly Notices of the Royal Astronomical Society*, 2017, 481(3): 4194
- [6] ODEWAHN S C, STOCKWELL E B, PENNINGTON R L, et al. Automated star/galaxy discrimination with neural networks[J]. *Astronomical Journal*, 1992, 103(1): 318
- [7] KIM E J, BRUNNER R J, KIND M C. A hybrid ensemble learning approach to star-galaxy classification[J]. *Monthly Notices of the Royal Astronomical Society*, 2015, 453(1): 507
- [8] HUERTAS-COMPANY M, GRAVET R, CABRERA-VIVES G, et al. A catalog of visual-like morphologies in the 5 CANDELS fields using deep-learning[J]. *The Astrophysical Journal Supplement Series*, 2015, 221(1): 1
- [9] LANUSSE F, MA Q, LI N, et al. CMU DeepLens: deep learning for automatic image-based galaxy-galaxy strong lens finding[J]. *Monthly Notices of the Royal Astronomical Society*, 2018, 473(3): 3895
- [10] KIM E J, BRUNNER R J. Star-galaxy classification using deep convolutional neural networks[J]. *Monthly Notices of the Royal Astronomical Society*, 2017, 464(4): 4463
- [11] PASQUET J, BERTIN E, TREYER M, et al. Photometric redshifts from SDSS images using a convolutional neural network[J]. *Astronomy and Astrophysics*, 2018, 621(A26): 1
- [12] HE K M, ZHANG X, REN S Q, et al. Deep residual learning for image recognition[C]//Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, IEEE Computer Society, Las Vegas, Nevada: IEEE, 2016: 770-778
- [13] 戴加明, 佟继周. 基于深度残差网络的星系形态分类[J]. *天文学进展*, 2018, 36(4): 384
- [14] ZHU X, DAI J, BIAN C, et al. Galaxy morphology classification with deep convolutional neural networks[J]. *Astrophysics and Space Science*, 2019, 364(4): 1
- [15] LI R, SHU Y, SU J, et al. Using deep residual networks to search for galaxy-Ly α emitter lens candidates based on spectroscopic selection[J]. *Monthly Notices of the Royal Astronomical Society*, 2019, 482(1): 313
- [16] WU J F, BOADA S. Using convolutional neural networks to predict galaxy metallicity from three-colour images[J]. *Monthly Notices of the Royal Astronomical Society*, 2019, 484(4): 4683
- [17] PASZKE A, GROSS S, MASSA F, et al. PyTorch: an imperative style, high-performance deep learning library[C]//Advances in Neural Information Processing Systems 32 (NeurIPS 2019), December 8-14, 2019, Vancouver Convention Center, Vancouver, Canada: Neural IPS Foundation, 2019: 8024
- [18] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. *Proceedings*

- of the IEEE, 1998, 86(11): 2278
- [19] GLOROT X, BENGIO Y. Understanding the difficulty of training deep feed forward neural networks[C]// Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, May 13-15, 2010, Chia Laguna Resort, Sardinia, Italy: PMLR, 2010:249
- [20] ABOLFATHI B, AGUADO D S, AGUILAR G, et al. The fourteenth data release of the Sloan digital sky survey: first spectroscopic data from the extended baryon oscillation spectroscopic survey and from the second phase of the apache point observatory galactic evolution experiment[J]. *Astrophysical Journal Supplement*, 2017, 235(2): 1
- [21] LI N, THAKAR A. Casjobs and myDB: a batch query workbench[J]. *Computing in Science & Engineering*, 2008, 10(1): 18
- [22] GRAY J, SZALAY A, NIETO-SANTISTEBAN M, et al. SDSS DR14 image list tool[EB/OL]. (2004-02)[2021-04-25]. <http://skyserver.sdss.org/dr14/en/tools/chart/listinfo.aspx>
- [23] BETHAPUDI S, DESAI S. Separation of pulsar signals from noise with supervised machine learning algorithms[J]. *Astronomy and Computing*, 2018, 23: 15
- [24] SWETS J A, DAVES R M, MONAHAN J. Better decisions through science[J]. *Scientific American*, 2000, 283(4): 82
- [25] 李超, 张文辉, 李然, 等. 基于Stacking集成学习的恒星/星系分类研究[J]. *天文学报*, 2020, 61(2): 102
- [26] AIHARA H, ALSAYYAD Y, ANDO M, et al. Second data release of the hyper supprime-cam subaru strategic program[J]. *Publications-Astronomical Society of Japan*, 2019, 71(6): 1

Star-galaxy separation by the residual neural network algorithm

YANG Yang¹⁾ WEN Zhonglue²⁾ XIA Junqing^{1)†}

(1)Department of Astronomy, Beijing Normal University, 100875, Beijing, China;

2)National Astronomical Observatories, Chinese Academy of Sciences, 100101, Beijing, China)

Abstract In this paper, the residual neural network (RNN) algorithm was used to classify pseudo-color images of stars and galaxies from Sloan digital sky survey (SDSS), with features obtained directly from images. Images of galaxies and stars with spectral information were used as training and test sets. After training, accuracy rate on the test set can reach 98.23%, and recall rate 98.80%, indicating that the RNN algorithm can accurately classify images of galaxies and stars. The probability of being a star or galaxy given by the classifier is verified, and the probability can be used to evaluate the reliability of classification. This classifier can be applied to future sky surveys to further test its performance.

Keywords sky-survey; deep learning; classification; star; galaxy

【责任编辑: 陆有忠】