第62卷 第2期 天 文 学 报 Vol.62 No.2
2021年3月 ACTA ASTRONOMICA SINICA Mar., 2021

doi: 10.15940/j.cnki.0001-5245.2021.02.009

基于深层残差网络的脉冲星候选体分类方法研究*

刘晓飞[1,2] 劳保强[1†] 安 涛[1] 徐志骏[1] 张仲莉[1]

(1 中国科学院上海天文台 上海 200030)
(2 中国科学院大学 北京 100049)

摘要 随着下一代射电天文望远镜的不断改进和发展, 脉冲星巡天观测将发现数百万个脉冲星候选体, 这给脉冲星的识别和新脉冲星的发现带来了巨大挑战, 迅速发展的人工智能技术可用于脉冲星识别. 使用Parkes望远镜的脉冲星数据集(The High Time Resolution Universe Survey, HTRUS), 设计了一个14层深的残差网络(Residual Network, ResNet)进行脉冲星候选体分类. 在HTRUS数据样本中, 存在非脉冲星候选体(负样本)的数目远远大于脉冲星候选体(正样本)数目的样本非均衡问题, 容易产生模型误判. 通过使用过采样技术对训练集中的正样本进行数据增强, 并调整正负样本的比例, 解决了正负样本非均衡问题. 训练过程中, 使用5折交叉验证来调节超参数, 最终构建出模型. 测试结果表明, 该模型能够取得较高的精确度(Precision)和召回率(Recall), 分别为98%和100%, F1分数(F1-score)能够达到99%, 每个样本检测完成只需要7 ms, 为未来脉冲星大数据分析提供了一个可行的办法.

关键词 **脉冲星**: **普通**, **数据集**: HTRUS, **方法**: **残差网络**, **方法**: **分类**

中图分类号: P145; **文献标识码**: A

1 引言

脉冲星是快速旋转、高度磁化的中子星[1]. 脉冲星为天文学的研究做出了巨大的贡献, 它既可以作为星际介质物理性质的探测器[2], 也可用于指示极端条件的天体物理现象[3], 其性质还为暗物质的存在提供了间接证据, 是研究强引力场的天体实验室[4], 尤其是对引力波辐射的检测[5–6]. 脉冲星的研究是射电天文学领域中一个有意义的课题, 因此, 发现新脉冲星、识别脉冲星候选体(可能检测到的新脉冲星)对揭示新的天文现象具有重要意义[7]. 然而, 已知脉冲星的数量很少[8], 需要寻找新的脉冲星. 自第1颗脉冲星发现以来, 脉冲星搜寻已经持续了很长一段时间[9].

在天文大数据时代, 具有大视场、快速巡天功能的平方公里阵列(Square Kilometre Array, SKA)射电望远镜[10]及500 m孔径球面射电望远镜(Five-hundred-meter Aperture

2020-07-22收到原稿, 2020-09-21收到修改稿

*国家重点研发计划项目(2018YFA0404603)和国家自然科学基金项目(U1831204)资助

†lbq@shao.ac.cn

Spherical Telescope, FAST)[11]等高灵敏度大型望远镜的出现, 使得脉冲星观测产生了大量数据, 预计会发现大量的脉冲星[12]. 在如此庞大的数据中, 不仅包含了数百万个候选脉冲星[13], 还包含了大量的人为射电频率干扰(Radio Frequency Interference, RFI)和各种非脉冲信号数据, 使得脉冲星的识别变得尤为困难. 天文学家需要寻求更加合理的筛选方法, 从而准确快速地完成海量候选体的识别.

目前, 提出了许多脉冲星候选体的筛选方法. 主要有: 手动筛选法[14–15], 其主要通过志愿者目视来对感兴趣的对象进行搜索或分类[9]; 图形工具筛选法[16–17]; 以及半自动排名法[18]. 在天文大数据时代, 目视检查大量的脉冲星候选体是不现实的, 因此, 有监督的机器学习方法应运而生[9]. 自2010年以来, 用于脉冲星候选体识别的机器学习方法迅速增多, 并且成为一种流行的方法. 文献[6, 9, 18]中的这些工作运用浅层神经网络(不超过3层)来训练数据集实现了对候选体的分类, 并且随着不断地精心设计和提取有效特征, 分类性能得到了改善. 文献[19]将人工神经网络(Artificial Neural Network, ANN)、支持向量机(Support Vector Machines, SVM)和卷积神经网络(Convolutional Neural Networks, CNN)这3种不同的监督算法组合使用, 引入图片模式识别的方法, 从诊断图中进行学习, 而不再从手动设计的特征中学习, 从而有效避免了依赖人工经验设计特征的缺点, 所设计的模型在PALFA (Pulsar Arecibo L-Band Feed Array)的观测数据中发现了6颗新的脉冲星. 此外, 统计学习方法也适用于对脉冲星候选体进行分类[7, 20], 利用统计学习的方法, 可以减小偏差. 有监督的机器学习, 特别是深度学习的最新方法[13, 21–23]也被用于处理脉冲星分类问题. 文献[13, 21]使用数据增强来解决不平衡问题, 并使用CNN的变体来训练模型, 利用提取的更深层的特征来进行分类, 使得精确度进一步提升. 文献[22]使用CNN组成的模型来训练FAST观测的脉冲星候选体数据, 在326个真实脉冲星数据中, 只有4个未正确判别. 文献[23]使用集成学习的方法对数据进行处理, 此方法是在文献[19]所设计模型的基础上进行改进的, 其利用15层残差网络(Residual Network, ResNet)代替CNN网络, 新模型可以识别出96%以上的真实脉冲星并且分类速率也进一步提升, 该模型在配置有2个GPU和24个CPU核心的计算平台上每天可完成160万个以上的候选样本分类任务(平均54 ms完成1个样本分类).

有监督的机器学习是一门数据驱动的方法. 传统的机器学习方法[18]依据从数据中提取的特征或一些人为设计构造的特征来进行模型的学习与训练, 基于已知知识获得的特征可能会忽略数据的某些潜在特征. 相反, 使用深度学习方法从数据中直接学习可以获得更好的性能. 本文应用深度ResNet来构建脉冲星候选体样本分类系统, 脉冲星候选体的诊断图作为模型的输入. 由于数据中的类别不均衡, 即负样本数目远远大于正样本数目, 会导致模型误判, 本文通过对少数样本进行过采样来解决这一问题. 然后在HTRUS (The High Time Resolution Universe Survey)数据集[24]上训练与评估构建的模型, 所训练的模型在精确度(Precision)和召回率(Recall)上均获得较优的结果.

本文首先简要介绍了用于训练模型的脉冲星候选体样本数据集, 同时介绍了数据预处理过程. 其次介绍了深度残差网络和模型的结构及损失函数, 在第4节解释了模型评估指标并详细介绍模型的训练细节, 然后展示测试结果并进行对比分析. 最后, 进行了总结与展望.

2 数据集和数据预处理

下一代射电望远镜的脉冲星巡天观测能够产生大量的数据, 由于诊断图样本量太大, 人工目视的方法将耗费大量的人力和时间, 并且随着工作者目视检查时间的延长, 由于疲劳极易判断出错. 使用自动的计算机处理算法即机器学习的方法来找到脉冲星候选体是一种必然趋势. 另外, 在使用机器学习算法时, 首先需要将数据预处理成为适合模型训练学习的输入.

2.1 数据集

本文采用文献[24]公开的数据集HTRUS来训练和评估模型. 数据来自澳大利亚Parkes望远镜的多波束(13个波束)的观测, 中心频率是1352 MHz, 每个波束记录带宽为400 MHz, 实际使用了中间340 MHz带宽上的数据. 该数据集包含了1196个具有不同自旋周期、占空比和信噪比的脉冲星(正样本), 还包含了89996个非脉冲星候选体(负样本). 另外, 该数据集根据样本的属性和个数进行命名, 由此脉冲星样本表示为从pulsar0000到pulsar1195, 非脉冲星样本表示为从cand000001到cand089996. HTRUS的数据文件包含结构化数据和非结构化数据, 其中有11个数值特征, 包括信号周期和最佳值(如最佳色散测度、最佳信噪比、最佳宽度等). 非结构化数据有6个矩阵数据, 均可视作图片数据. 其中, 时间-相位图和频率-相位图将作为本文的训练集和测试集, 如图1所示. 图1是来自HTRUS数据集中脉冲星(pulsar0816)和RFI (cand036371)样本的时间-相位图和频率-相位图示例, 图中左图由pulsar0816数据产生, 右图由cand036371产生.

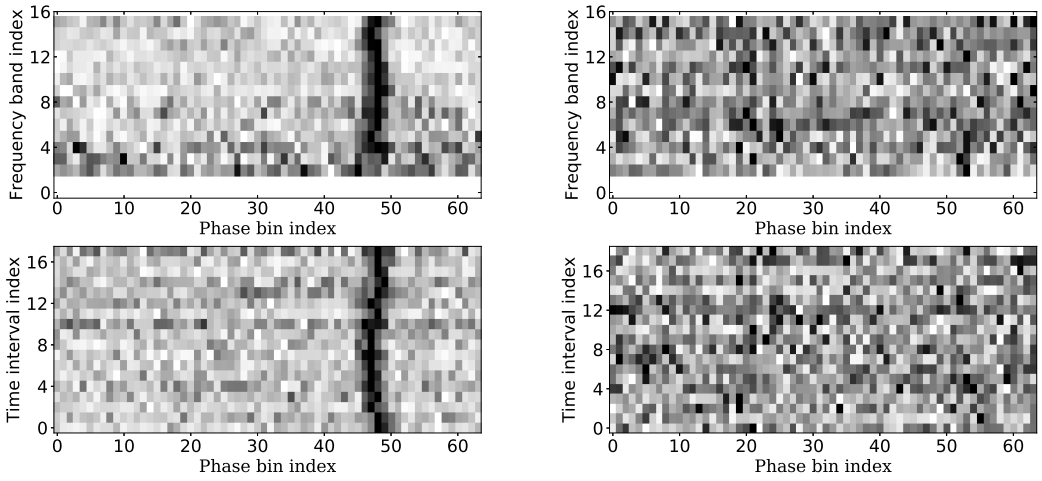

图 1 样本图像示例. 左: 脉冲星; 右: 射电频率干扰信号.

Fig. 1 Examples of sample plot. Left: a pulsar example; Right: RFI signals.

从图1可以看出脉冲星候选体和非脉冲星候选体具有明显不同的特征: (1)频率-相位图是3维(时间-频率-相位)数据在时间维度上进行求和获得的, 它反映了观测期间信号的强度. 对于理想的脉冲星信号, 将在整个观测周期内观察到该信号, 且存在一个或几个垂直条纹; (2)时间-相位图是3维(时间-频率-相位)数据在频率维度上进行求和获得的, 可以反映不同频率下的信号强度. 同样对于理想脉冲星信号, 应该存在一个或多个垂直条

纹. 通过统计我们发现HTRUS中的大多数脉冲星样本的时间-相位图和频率-相位图均存在一个或者多个垂直条纹, 特征非常明显. 非脉冲星的图像则是杂乱无章, 无规律可言. 这些从样本数据中获得的时间-相位图和频率-相位图将输入到本文的算法进行训练与测试.

2.2 数据预处理

数据预处理可以看作是一种数据挖掘技术, 可以将原始数据转换为易于处理的格式, 从而可以进一步分析. 原始数据通常存在噪声和样本不平衡等问题, 数据预处理是解决此类问题并为特征工程或下一步模型训练做好充分准备的可靠方法. 首先模型输入PNG图像大小不一, 所以需要统一大小; 另外, 在数据集中, 正负样本之间的比率约为1 : 73, 样本分布偏向负类别, 这意味着数据类别严重失衡[25]. 非平衡问题会降低模型的准确性[25–26], 并且模型最后的输出softmax函数是基于阈值的, 在数据不平衡的时候, 特定的阈值会导致模型输出倾向于类别数据多的一类.

下面介绍具体的处理步骤: 首先从数据集中获得需要的诊断因子图, 从HTRUS数据中获得的每个样本诊断图的大小是不同的. 为了便于训练模型, 先对所有的样本图片进行图片的缩放, 使其成为500 × 500像素同一大小的灰度图. 其次, 将正负样本获得的图片数据分别随机分为两部分: 训练集(80%), 测试集(20%). 然后, 处理数据非均衡问题, 处理数据不均衡问题常用的方法是对少数类别样本进行过采样, 或者对多数类别样本进行欠采样, 以此来调整数据比率达到样本均衡. 在训练数据集中, 由于少数类别(脉冲星)的数量太少, 如果对多数类别(RFI)进行欠采样就会导致训练样本数据的牺牲, 造成数据量太少, 无法进行学习, 所以采取对少数类进行过采样增加样本. 过采样包括简单的过采样[27](oversampling)和基于随机过采样算法的合成少数类过采样技术[28](Synthetic Minority Oversampling Technique, SMOTE). 由于模型输入是诊断图图片数据, 利用SMOTE方法进行处理时, 需要先对图像数据进行重塑, 然后进行过采样, 尝试这种方法后, 发现对于HTRUS数据集, SMOTE非常缓慢, 且最后分类效果不好. 分析原因可能是因为重塑图像后, 图像会丢失掉像素之间的位置关联信息, 而由于模型是基于卷积的残差网络, 图像反映的信息是使用卷积的主要原因.

综上, 选择使用简单的过采样方法处理数据. 对少数类过采样是从少数样本中随机抽取样本进行复制, 直到达到想要的数量. 由于本质并没有为模型引入更多的数据, 可能会导致过分强调少数类, 放大少数类别中噪声对模型的影响. 因此我们尝试了不同的比率取值, 将脉冲星正样本与非脉冲星负样本的比率设置在1 : 1到1 : 5的范围内, 以寻找最佳比率值. 最终发现当设置为1 : 2的时候为最佳比率, 表1中交叉验证集为过采样之后的数据量. 此处需要强调的是, 对少数样本(脉冲星正样本)进行过采样时处理的是80%的训练集, 测试集没有做处理.

数据集预处理完成后开始进行模型训练, 为了充分利用训练数据集, 采用了5折交叉验证来完成训练模型的调参, 找出效果最优的模型. 交叉验证作用是尝试利用不同的训练集/验证集划分来对模型做多组不同的训练/验证, 来应对单独测试结果过于片面以及训练数据不足的问题. 本文训练过程中所采用的5折交叉验证是指将数据集(交叉验证集)随机划分5个大小相近的互斥子集, 每次不重复地取其中1个子集做为测试集, 其余4个子集合并作为训练集. 因此, 将会产生5组不同的训练集和测试集. 计算该模型在每

组训练集上的均方误差(Mean Square Error, MSE), 将5组训练模型的MSE取平均作为最终的模型. 然后使用该模型对测试集进行预测, 从而衡量该模型的性能和分类能力. 在训练过程中, 训练集共有73272个样本用于超参数筛选, 测试集有17920个样本. 其中, 测试集中含有17680个负样本, 240个正样本. 样本数量的详细情况如表1所示. 表1中分别展示了HTRUS数据集的总样本数量, 占比80%、20%所切割出的训练集和测试集以及交叉验证集, 其中交叉验证集是对训练集中正样本进行过采样之后用来交叉验证训练模型的样本数量.

表 1 数据集的介绍
Table 1 Introduction to data set

DataSet	Pulsars	RFI	Total samples
HTRUS dataset	1196	89996	91192
Training dataset	956	72316	73272
Cross-validation dataset (oversampling)	36158	72316	108474
Testing dataset	240	17680	17920

3 机器学习模型

3.1 深度残差网络

自深层卷积神经网络从ImageNet分类挑战[29]中取得巨大突破以来获得了广泛的研究, 它在许多其他计算机视觉任务(包括对象检测、语义分割、边缘检测等)中也取得了令人惊讶的性能. 近年, 脉冲星候选体筛选使用的大多数深度学习网络模型都是基于卷积神经网络或其某种变体: 文献[19]使用5层卷积神经网络处理2维子图进行脉冲星分类; 文献[13]设计了一个11层的卷积神经网络结果以取代文献[19]中简单的5层体系结构, 并获得了更好的性能; 文献[21]使用生成对抗网络(Generative Adversarial Networks, GAN)解决样本类别不平衡问题, 然后使用卷积神经网络训练新数据集; 文献[22]描述了一个组合卷积神经网络, 来识别从FAST数据中收集的候选对象. 但深度神经网络随着层数的逐渐加深变得难以训练和优化, 而卷积神经网络模型的训练和分类相对较慢[23]. 另外, 随着网络的深入, 尽管增加了层数, 但训练误差仍大于浅层的网络, 为了解决这一问题, 残差网络随之产生. ResNet[30]赢得了ImageNet的挑战, 引起了很多关注, 因为深度残差网络能够“使深度学习变得更加深入”. 残差网络可以被视为具有残差学习框架的典型卷积神经网络. 深度神经网络随着网络深度的增加, 往往会出现梯度弥散或梯度爆炸等问题, 运用正则化可以使得网络继续训练, 但是, 随着网络层数的增加, 模型在训练集上的准确率会达到饱和甚至下降, 也就是出现了退化问题. 文献[30]提出了一种残差学习框架(图2)来解决退化问题. 如图2所示, X表示残差块的输入(诊断图图像), 模型的期望输出为$H(X)$, $F(X)$是网络学习到的映射函数, 所使用的激活函数为Relu (Rectified linear unit). 如果深层网络的后面网络层是恒等映射, 即直接把输入的X传到输出, 那么模型就会退化成为一个浅层网络, 其关系为$H(X) = F(X) + X$. 由残差块组成可知, 要学习的函数$F(X)$经过残差块转化为残差函数$F(X) = H(X) - X$. 当$F(X) = 0$, 就构成

了一个恒等映射$H(X) = X$, 拟合残差相对更容易. 残差网络由一系列残留块组成, 每个残差块包含两个分支: 恒等映射和残差分支. 恒等映射从输入到中间层以及从中间层到输出层引入了捷径连接(shortcut connections) (图2中的X identity), 缓解了梯度消失问题, 同时降低了训练难度. 残差网络容易优化, 并且随着网络深度的增加, 准确率会提高.

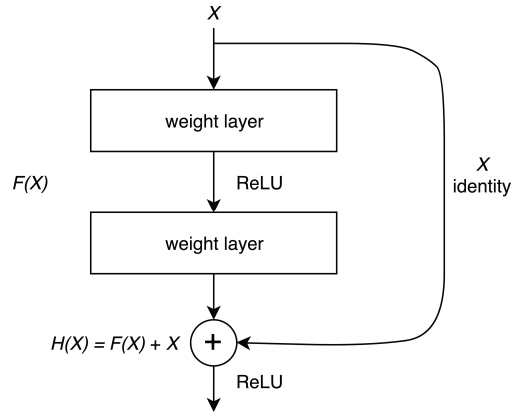

图 2 残差学习结构: 残差块[30]

Fig. 2 Residual learning structure: residual block[30]

3.2 模型架构

ResNet结构是功能强大的识别系统, 可以直接分析2维图像. 根据网络的这一特征, 直接将脉冲星数据集预处理后的2维子图(图1)作为输入. 设计的网络结构由图像大小、网络深度、特征映射的数量、卷积核大小、池化层大小等超参数确定. 为了获得最佳模型, 通过交叉验证的方法来调整模型超参数的各种组合. 首先将候选体标记数据随机打乱进行分割, 其中80%的数据样本组成训练集, 20%的样本组成测试集, 然后通过网格搜索的方式对模型进行训练, 通过在测试集上计算其F1分数(F1-score)来表征性能. 最后, 根据分类精度确定了网络结构和超参数. 用14层的深度残差网络进行实验, 发现它能够获得很高的精度. 因此, 本文设计的ResNet模型包括14层. 图3展示的是本文使用的ResNet模型网络结构, 同时显示了整个训练过程. 图3主要分为两部分, 左侧是模型的整体训练流程, 右侧的ResBlock是残差块(Residual block). Image_data是模型的输入数据, 为500 × 500像素大小的灰度图片. COV表示卷积运算, 其中COV_7 × 7_64所代表的分别是卷积核大小为7 × 7, 最后的输出通道数目为64; C1 : 112 × 112 × 64中C代表卷积层(Convolution, C), C1表示卷积层为第1层, 经过此层的卷积之后, 输出大小为112 × 112 × 64. BN表示批量标准化(Batch Normalization), ResNet引入BN是为了提高模型的训练速度, 加快收敛. Max_pooling是最大值池化, Average_pooling是平均池化, FC (Fully Connected layer)指的是全连接层, softmax是最后的输出, 输出显示了2个类别的分类得分和类别名(1表示正样本, 0表示负样本), 从而完成对样本进行预测. 所设计ResNet模型的Resblock个数分别为[2, 2, 1, 1], 加上开始的1个卷积层和最后的1个全连接层共有14层. ResBlock部分, 输入为所要分类处理的图片, 大小为$W \times H$, 输入图片的通道数目为IN, 输出图片通道数目为OUT, 如

果IN == OUT, 则经过卷积块处理之后的输出结果为$[W \times H \times \mathrm{OUT}]$; 反之, 输出结果为$\{\mathrm{int}[(W-1)/2+1], \mathrm{int}[(W-1)/2+1], \mathrm{OUT}\}$, 其中int为取整函数.

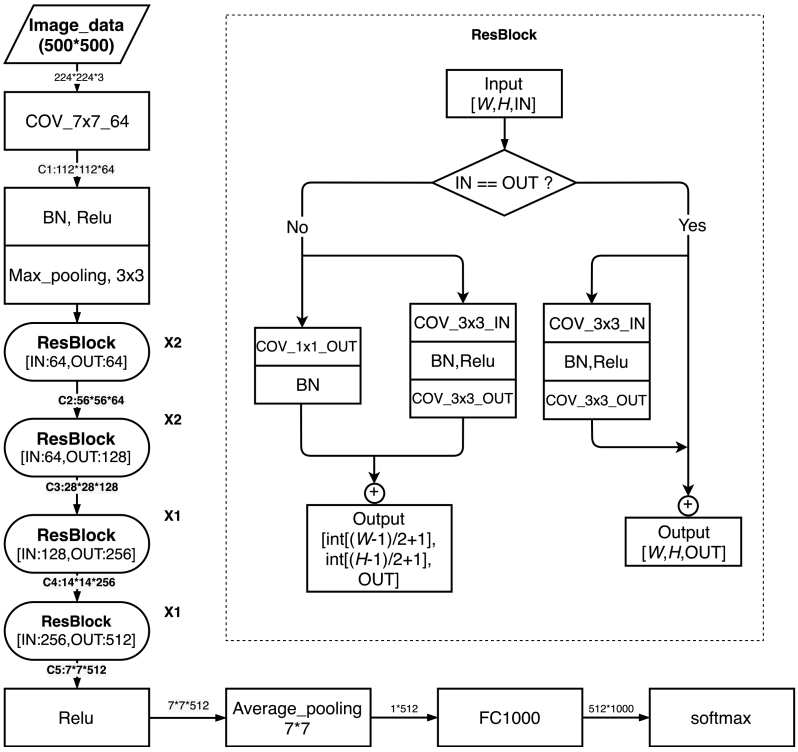

图 3 模型架构[30]

Fig. 3 Model architecture[30]

与文献[23]相比, 网络模型的整体结构相似, 均是利用ResNet提取特征与分类. 本文的网络模型是在传统的18层ResNet网络的基础上, 以得到较优的网络性能为目标, 经过多次网络结构修改和参数调整确定的. 不同之处有以下两点: 首先是网络的层数不同. 本文的ResNet层数是14层, 文献[23]的ResNet层数为15层. 所设计ResNet模型如图3所示, 包含1个输入的卷积层、1个全连接输出层、中间是12个卷积层, 共14层. 12个卷积层由4组残差块组成, 每组残差块包含的残差块的个数分别为[2, 2, 1, 1], 每个残差块均包含2个卷积层. 文献[23]的网络模型包含1个输入卷积层、2个输出层、中间是12个卷积层, 共15层. 12个卷积层由3组残差块组成, 每组有2个残差块, 每个残差块均有2个卷积层; 其次, 卷积层的步长和输出通道大小均不同. 本文采用的步长分别是[2, 1, 2, 2, 2], 因此最后一层卷积输出的图像的宽和高为输入图像的1/16. 文献[23]使用的卷积层的步长均是1, 因此最后一层卷积输出的图像大小和输入图像大小相同. 本文的卷积层的输出通道大小分别是[64, 64, 128, 256, 512], 第1个是输入卷积层的输出通道大小, 其余的是4组残差块的输出通道大小. 文献[23]卷积层的输出通道大小分别是[16, 16, 32, 64], 第1个是输入卷积层的输出通道大小, 其余3个是3组残差块的输出通道大小.

除了模型本身已有的批量标准化处理, 在模型训练过程中为了防止过度拟合, 使用批量归一化处理和$L2$正则化, 具体定义见3.3节. 用深层训练神经网络具有挑战性, 因为

它们可能对初始随机权重和学习算法的配置敏感. 批量归一化处理[31]是一种用于训练非常深的神经网络的技术, 该技术将每个小批次的输入标准化, 使模型学习过程稳定, 并显著减少训练深度网络所需的训练时间. 另外, 还使用Adam优化器[32]在特定批次大小的数据中训练网络, 以加快训练过程.

3.3 损失函数

脉冲星候选体识别任务是二分类问题, 因此将交叉熵损失用作损失函数来优化模型. 交叉熵损失$L(\omega)$的定义如下:

$$L(\omega) = -\frac{1}{N}\sum_{n=1}^{N}[y_n \lg \hat{y}_n + (1-y_n)\lg(1-\hat{y}_n)]\,, \tag{1}$$

其中, ω表示权重向量, N表示样本数, y_n和$\hat{y}_n$分别表示实际值和预测输出值.

为了使模型更具有鲁棒性, 添加了$L2$正则化作为权重衰减. 此时损失函数的计算公式重新定义为:

$$\hat{L}(\omega) = -\frac{1}{N}\sum_{n=1}^{N}[y_n \lg \hat{y}_n + (1-y_n)\lg(1-\hat{y}_n)] + \frac{\lambda}{2}\omega^2\,, \tag{2}$$

其中λ是惩罚系数.

$L2$正则化可以减轻模型对样本的敏感, 此外还使用批量归一化处理技术和Adam优化器来优化模型.

4 训练结果与分析

4.1 评估指标

由于HTRUS数据集类别存在高度不平衡问题, 无法直接根据分类准确度推断模型性能. 因此, 需要建立新的评估指标来评估分类器的有效性, 引入了二分类的混淆矩阵, 如下表2所示:

表 2 混淆矩阵

Table 2 Confusion matrix

Actual category \ Predicted results	Positive	Negative
Pulsar (True)	TP	TN
RFI (False)	FP	FN

TP是模型将正样本正确识别为正样本的数量, 即脉冲星被正确识别的数量. 这意味着观测样本属于正类, 并且被分类器正确地标记为1;

TN是模型将正样本错误识别为负样本的数量, 即脉冲星被误判为非脉冲星的数量;

FP是模型将负样本错误识别为正样本的数量, 即非脉冲星被误判为脉冲星的数量;

FN是模型将非脉冲星正确识别为负样本的数量. 这意味着观测样本属于负类, 并且被分类器正确地标记为0.

总之, TP和FN表示可以正确识别的对象, 而TN和FP表示错误识别的对象. 基于这些值来定义指标以评估本文算法的性能, 以下是对本文所使用评估指标的解释.

召回率(Recall): 又称查全率, 表示正确识别脉冲星的比例.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}\,; \tag{3}$$

精确度(Precision): 精确度表示预测为正样本的数据中脉冲星数量所占的比例.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}\,; \tag{4}$$

准确率(Accuracy): 准确率是一般分类场景中最常用的评分指标, 它需要确保样本平衡. 在HTRUS失衡数据集中, 这个评估分数几乎没有什么意义.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}\,; \tag{5}$$

F1分数(F1-score): F1-score是精确度和召回率的调和平均值, 它提供了另一种准确性度量. 它可以用作精确度和召回率之间的折衷, 可以确保训练的模型在不会忽略某些脉冲星信号的同时也有很高的分类准确率, 值越接近1代表分类效果越好.

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}\,. \tag{6}$$

4.2 模型训练

在本小节中, 描述了14层残差网络(ResNet)的训练和测试过程, 所使用的数据集为HTRUS数据集. 本文采用的实验平台是中国SKA区域数据中心原型机[33], 该原型机已经成功应用于SKA连续谱成像数据处理[34]和SKA脉冲星数据处理[35]. 本文的训练和测试任务均是在该原型机的GPU集群上完成, 该集群共有3个GPU节点, 共配置了16个GPU卡. 本文的工作主要使用了1个GPU卡, 型号是Nvidia Tesla V100 SXM2, 这个GPU卡的理论浮点运算能力约为15.7 TFLOP (单精度).

训练采用了交叉验证方法和模型超参数的设置与筛选. 在数据预处理完成之后, 采用5折交叉验证的方法对交叉验证集(表1)进行模型训练, 详细过程见2.2节. 本文的网络训练一共进行了35轮(epoch)的迭代, 如图4所示, 画出了损失函数随着迭代次数增加的变化曲线. 训练的目的是尽可能减小损失并确保模型能够收敛, 从而得到较优的模型. 观察曲线可看出, 训练过程中损失函数数值随着训练轮数的增加在不断地降低, 最终到达一个较低的数值后保持平稳, 说明模型能够收敛. 在进行实验时也尝试过用14层以上的ResNet进行模型训练, 由于层数太深, 出现了过拟合, 最终获得的损失函数曲线随着训练轮数的增加一直震荡, 无法收敛.

根据损失函数曲线来诊断模型, 一般而言, 完美拟合是模型调整的目标, 它在损失函数曲线上的特点是训练误差(training loss)和测试误差(validation loss)两者都能够收敛, 并且两者之间相差很小. 从图4可以看出模型大概在10轮之后两个损失函数曲线都开始收敛, 在35轮之后, 两条曲线没有肉眼可见的明显差距, 证明最终构建的模型可以很好地拟合数据.

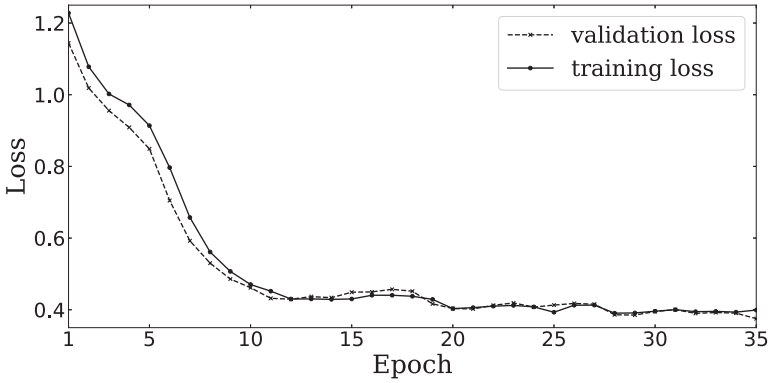

图 4 损失函数曲线

Fig. 4 Loss function curve

关于模型超参数的设置与筛选, 在所设计模型训练过程中为了防止过拟合采取了很多措施. 在损失函数求解过程中加入$L2$正则化, 在3.3节已经给出介绍. 还使用批量归一化对数据进行处理, 利用Adam优化器优化模型. 其中批处理的大小(batch size)为512, 在每轮中, 每次从训练集中随机筛选512个候选体样本来训练模型和更新参数. 重复此步骤, 直到使用完所有训练数据为止. 一共训练了35轮, 最终的损失函数值为0.38. Adam优化器的超参数初始值的设置在表3中详细给出, 其中batch_size代表批处理大小, epoch表示训练轮数. CONV_WEIGHT_DECAY为卷积层权重衰减率, 用于权重衰减($L2$正则化), CONV_WEIGHT_STDDEV为卷积层权重标准差, 可以避免模型过拟合问题. BN_DECAY为BN层衰减率, BN_EPSILON为BN层初始化参数, 会影响模型的训练速率和准确率.

表 3 模型超参数数值

Table 3 Hyperparameter values of model

Hyperparameters	Value
batch_size	512
epoch	35
CONV_WEIGHT_DECAY	0.00004
CONV_WEIGHT_STDDEV	0.1
BN_DECAY	0.9997
BN_EPSILON	0.001

4.3 测试结果与分析

模型训练完成后, 用测试集进行模型性能的测试. 本小节展示了最终的模型测试结果, 并与利用人工智能方法构建的其他模型在HTURS数据集上的性能(表现)进行比较, 然后对误判样本进行了成图分析.

测试结果如表4所示, 展示的是在测试集上所得出的混淆矩阵. 表4中, 已知测试集中含有17680个负样本, 240个正样本, 测试结果中全部的脉冲星都被判别正确, RFI则

有6个样本被误判为脉冲星. 通过进一步计算可以得出, 对于HTRUS数据集, 在测试数据上, ResNet模型可以实现100%的召回率和98%的精确度, F1-score能够达到99%. 另外, 每个样本检测完成所消耗的时间约为7 ms, 运行速度相当高.

表 4 测试结果的混淆矩阵
Table 4 Confusion matrix of testing results

Actual category \ Predicted Results	Positive	Negative
Pulsar (True)	240 (TP)	0 (TN)
RFI (False)	6 (FP)	17674 (FN)

表5是在HTRUS数据集上不同模型的测试结果, 其中SPINN[24] (Straightforward Pulsar Identification using Neural Networks)是通过提取输入数据6个数值特征作为网络输入, 训练神经网络ANN获得模型; GH-VFDT[7] (the Gaussian-Hellinger Very Fast Decision Tree)是利用从输入数据提取的8个数值特征对树网络进行训练获得模型; DCNN-S[13]是先用SMOTE方法对原始数据进行数据增强, 然后训练深度卷积神经网络(Deep Convolutional Neural Network, DCNN)得到最终模型; CGAN-L2-SVM-1[21]是时间-相位图作为网络输入, 通过训练由深度卷积生成的对抗网络(Deep Convolution Generative Adversarial Network, DCGAN), 分类层采用线性SVM (L2-SVM)来进行分类, 从而获得最终模型. 根据结果可以看出本文的训练模型获得的分类性能较好, Recall能够达到100%, 相比于其他模型, 精确度也提高到98%, 即预测为正样本的数据中脉冲星数量所占的比例是98%, 因为在测试结果中仅有6个样本被判断错误(表4). 虽然精确度不能达到100%, 但是由于分类任务的关键是识别所有的脉冲星样本, Recall的结果更为重要, 已经达到了100%, 意味着模型在HTRUS测试集上能够识别全部的脉冲星样本. 综上, 实验结果表明本文设计的网络获得的模型具有较优秀的分类性能.

表 5 HTRUS数据集上不同分类器分类性能的比较
Table 5 Comparison of classification performance for different classifiers on the HTRUS data set

Reference	Method	Recall	Precision	F1-score
Bates et al.[6]	ANN	85%	/	/
Morello et al.[24]	SPINN	100%	/	/
Lyon et al.[7]	GH-VFDT	93%	96%	94%
Wang et al.[13]	DCNN-S	96%	96%	96%
Guo et al.[21]	DCGAN-L2-SVM-1	97%	96%	96%
Our method	ResNet	100%	98%	99%

另外为了分析模型的不足之处, 我们对判断错误的RFI样本进行了分析. 分析发现导致RFI误判的原因主要有2个: 第1个原因是训练数据中正样本的问题, 在用于训练的

数据样本中, 存在一部分流量密度小, 且噪声水平较高的脉冲星正样本, 它们的数据并不存在图1所展示的明显条纹, 其时间-相位图和频率-相位图中信号强度的分布是杂乱无章的. 而在进行模型设计时, 除了要保证高精确度外, 还要保持极高的脉冲星召回率, 使其不错过任何一个候选体样本, 因为这个原因可能会导致训练出的模型存在偏置. 第2个原因是预测数据中的负样本(RFI)数据的问题, 这些数据产生的时间-相位图和频率-相位图与脉冲星图像很相似. 图5给出了测试集中一个被误判为脉冲星的RFI样本(cand072775)示例. 在分析过程中, 为了防止人为观察不够准确引入错误, 加入了色散度量(Dispersion Measure, DM)-信噪比(Signal to Noise Ratio, SNR)曲线对样本的分析, 即图5下方子图. 图5上方的两个子图是模型用来训练的输入子图(时间-相位图和频率-相位图). DM-SNR曲线记录了SNR与DM的关系, 当使用不同的色散值进行消色散时, 色散曲线显示脉冲曲线的相应SNR. 如果是脉冲信号, 则曲线将在非零位置出现一个峰值, 而非脉冲信号的曲线则没有明显峰值或峰值在零处. 观察此图像的DM-SNR曲线没有发现明显的峰值, 这类信号大概率就是大气层内人造射电源的信号, 并非来自脉冲星的脉冲信号.

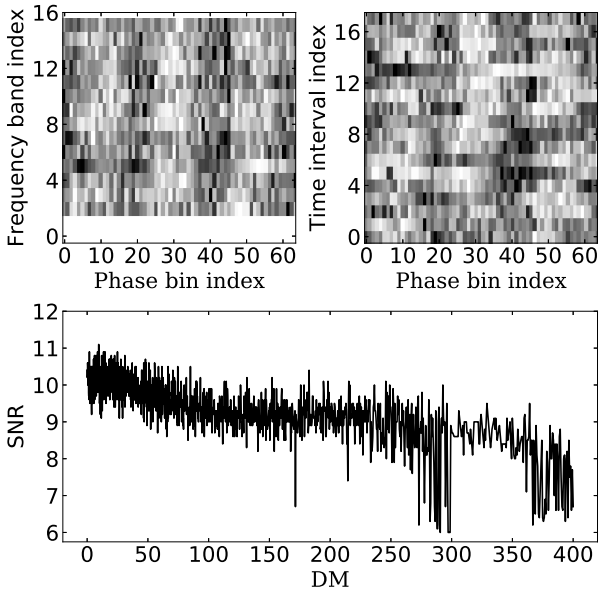

图 5 RFI误判样本示例

Fig. 5 Examples of misjudged RFI

5 总结和展望

为了解决脉冲星候选体识别问题, 提出了一种基于ResNet的网络结构. 通过直接使用脉冲星候选体的2维子图作为输入, 它避免了数据中人为设计提取特征的繁琐, 并且还可防止人为因素对数据的干扰, 更符合深度学习数据驱动的性质. 在此项工作中, 为了解决类别不平衡问题, 使用了过采样技术对少数类别样本进行处理. 为了应对样本数量不足的问题采取交叉验证的方法来训练数据集, 使得所构建的模型能更充分地学习, 更

好地拟合数据. 所设计的14层ResNet模型最终达到了100%的Recall, 能够识别所有的脉冲星样本, 具有出色的分类能力.

ResNet是一种先进的网络, 在脉冲星搜索领域具有良好的应用前景. 为了提高泛化性能和分类精度, 可以考虑其他的图像处理方法, 不仅仅是应用过采样的方法对数据进行处理, 还可以对脉冲星正样本加入噪声, 这样模型就能够更精细地学习到样本数据之间的区别, 非脉冲星样本被误判的可能性可能会被降低; 其次, 还可以考虑数据的其他诊断图特征(总轮廓直方图, DM-SNR曲线等). 未来的工作将考虑DM-SNR曲线作为模型的输入, 并尝试使用其他模型拟合曲线以确定是否值得使用其他特征作为输入. 最后, 最重要的是需要收集新的观测数据样本, 尤其是脉冲星正样本的数量, 样本数据问题越少(比如样本不均衡问题), 数据量越大, 模型就越能发现数据所存在的内在规律, 其分类精度就会越高.

参 考 文 献

[1] Lorimer D R, Kramer M. Handbook of Pulsar Astronomy. Cambridge: Cambridge University Press, 2005

[2] Armstrong J W, Rickett B J, Spangler S R. ApJ, 1995, 443: 209

[3] Lattimer J M, Prakash M. Science, 2004, 304: 536

[4] Antoniadis J, Freire P C C, Wex N, et al. Science, 2013, 340: 1233232

[5] Taylor J H, Weisberg J M. ApJ, 1982, 253: 908

[6] Bates S D, Bailes M, Barsdell B R, et al. MNRAS, 2012, 427: 1052

[7] Lyon R J, Stappers B W, Cooper S, et al. MNRAS, 2016, 459: 1104

[8] Hobbs G, Manchester R, Teoh A, et al. Proceedings of the International Astronomical Union, 2004, 218: 139

[9] Eatough R P, Molkenthin N, Kramer M, et al. MNRAS, 2010, 407: 2443

[10] Smits R, Kramer M, Stappers B, et al. A&A, 2009, 493: 1161

[11] Nan R D, Li D, Jin C J, et al. IJMPD, 2011, 20: 989

[12] An T. SCPMA, 2019, 62: 989531

[13] Wang Y C, Li M T, Pan Z C, et al. RAA, 2019, 19: 133

[14] Stokes G H, Segelstein D J, Taylor J H, et al. ApJ, 1986, 311: 694

[15] Johnston S, Lyne A G, Manchester R N, et al. MNRAS, 1992, 255: 401

[16] Faulkner A J, Stairs I H, Kramer M, et al. MNRAS, 2004, 355: 147

[17] Keith M J, Eatough R P, Lyne A G, et al. MNRAS, 2009, 395: 837

[18] Lee K J, Stovall K, Jenet F A, et al. MNRAS, 2013, 433: 688

[19] Zhu W W, Berndsen A, Madsen E C, et al. ApJ, 2014, 781: 117

[20] Tan C M, Lyon R J, Stappers B W, et al. MNRAS, 2018, 474: 4571

[21] Guo P, Duan F Q, Wang P, et al. MNRAS, 2019, 490: 5424

[22] Zeng Q G, Li X R, Lin H T, et al. MNRAS, 2020, 494: 3110

[23] Wang H F, Zhu W W, Guo P, et al. SCPMA, 2019, 62: 959507

[24] Morello V, Barr E D, Bailes M, et al. MNRAS, 2014, 443: 1651

[25] He H B, Garcia E A. IEEE Transactions on Knowledge and Data Engineering, 2009, 21: 1263

[26] Buda M, Maki A, Mazurowski M A, et al. Neural Networks, 2018, 106: 249

[27] Kalker A A C M. ElL, 1992, 28: 567

[28] Chawla N V, Bowyer K W, Hall L O, et al. Journal of Artificial Intelligence Research, 2002, 16: 321

[29] Krizhevsky A, Sutskever I, Hinton G E, et al. Proceedings of the 25th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2012, 1: 1097

[30] He K M, Zhang X Y, Ren S Q, et al. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 770

[31] Ioffe S, Szegedy C. Proceedings of the 32nd International Conference on Machine Learning. 2015, 37: 448

[32] Kingma D P, Ba J. Adam: A Method for Stochastic Optimization. Proceedings of the 3rd International Conference on Learning Representations. Ithaca, 2015

[33] An T, Wu X P, Hong X Y. NatAs, 2019, 3: 1030

[34] Lao B Q, An T. SSPMA, 2020, 50: 059501

[35] Gong H Y, Zhang Z L, Xue M Y, et al. SSPMA, 2020, 50: 109501

Research on Pulsar Candidate Identification Method Based on Deep Residual Neural Network

LIU Xiao-fei[1,2] LAO Bao-qiang[1] AN Tao[1] XU Zhi-jun[1] ZHANG Zhong-li[1]

(1 Shanghai Astronomical Observatory, Chinese Academy of Sciences, Shanghai 200030)
(2 University of Chinese Academy of Sciences, Beijing 100049)

Abstract As the next generation of radio astronomical telescopes continue to improve and develop, the pulsar survey will produce millions of pulsar candidates, which pose considerable challenges for pulsar identification and classification. The rapidly evolving artificial intelligence (AI) techniques are being used for pulsar identification and discovery of new pulsars. Using the pulsar data set observed with the Parkes telescope, the High Time Resolution Universe Survey (HTRUS), a 14-layer deep residual network was designed (called Residual Network, ResNet) for pulsar candidate classification. In the HTRUS sample, the number of non-pulsar candidates (i.e., negative sample) is much larger than that of pulsar candidates (i.e., positive sample). The imbalance of positive and negative samples is prone to result in model misinterpretation. By using the oversampling technique to enhance the data of positive samples in the training set and adjust the ratio of positive and negative samples, we solve this Non-equilibrium problem. During training, the hyperparameters were adjusted using 5-fold cross-validation to construct the model. The test results show that the model can achieve high precision (98%) and recall (100%), respectively, and the F1-score is able to reach 99% for each sample tested. It takes only 7 ms to complete each candidate classification, providing a viable approach to future pulsar big data analysis.

Key words pulsar: general, data set: HTRUS, methods: ResNet, methods: classification