物理学报 Acta Physica Sinica
Chinese Physical Society
Institute of Physics, CAS

基于自归一化神经网络的脉冲星候选体选择

康志伟 刘拓 刘劲 马辛 陈晓

Pulsar candidate selection based on self-normalizing neural networks

Kang Zhi-Wei Liu Tuo Liu Jin Ma Xin Chen Xiao

引用信息 Citation: Acta Physica Sinica, 69, 069701 (2020) DOI: 10.7498/aps.69.20191582
在线阅读 View online: https://doi.org/10.7498/aps.69.20191582
当期内容 View table of contents: http://wulixb.iphy.ac.cn

基于自归一化神经网络的脉冲星候选体选择*

康志伟[1)†] 刘拓[1)] 刘劲[2)] 马辛[3)] 陈晓[4)]

1) (湖南大学信息科学与工程学院, 长沙 410082)

2) (武汉科技大学信息科学与工程学院, 武汉 430081)

3) (北京航空航天大学仪器科学与光电工程学院, 北京 100191)

4) (上海卫星工程研究所, 上海 200240)

(2019 年 10 月 17 日收到; 2019 年 12 月 19 日收到修改稿)

脉冲星候选体选择是脉冲星搜寻任务中的重要步骤. 为了提高脉冲星候选体选择的准确率, 提出了一种基于自归一化神经网络的候选体选择方法. 该方法采用自归一化神经网络、遗传算法、合成少数类过采样这三种技术提升对脉冲星候选体的筛选能力. 利用自归一化神经网络的自归一化性质克服了深层神经网络训练中梯度消失和爆炸的问题, 大大加快了训练速度. 为了消除样本数据的冗余性, 利用遗传算法对脉冲星候选体的样本特征进行选择, 得到了最优特征子集. 针对数据中真实脉冲星样本数极少带来的严重类不平衡性, 采用合成少数类过采样技术生成脉冲星候选体样本, 降低了类不平衡率. 以分类精度为评价指标, 在 3 个脉冲星候选体数据集上的实验结果表明, 本文提出的方法能有效提升脉冲星候选体选择的性能.

关键词: 脉冲星候选体选择, 自归一化神经网络, 特征选择, 类不平衡

PACS: 97.60.Gb, 98.52.Cf, 07.05.Mh **DOI:** 10.7498/aps.69.20191582

1 引 言

脉冲星是一种高速自转的中子星[1], 对其进行观测研究, 将极大推动星际介质研究[2]、引力波探测[3]、脉冲星导航[4–6]等众多领域的发展. 自第一颗脉冲星被发现以来[7], 在银河系、麦哲伦星云、球状星团中先后发现了 2700 多颗脉冲星[8], 其中大部分是通过现代射电望远镜探测发现的, 例如绿岸北半球脉冲星巡天[9](green bank north celestial cap survey, GBNCC)、Parkes 多波束脉冲星巡天[10](parkes multi-beam pulsar survey, PMPS)、高时间分辨率的宇宙脉冲星巡天[11](high time resolution universe survey, HTRU)、低频射电 (low frequency array, LOFAR) 阵列巡天[12] (LOFAR tied-array all-sky survey, LOTAAS), 这些都为脉冲星搜索奠定了基础.

脉冲星搜索首先需要检测出射电望远镜观测数据中的周期信号, 为便于分析, 一般要对这些具有周期性的观测数据进行统计描述, 以形成具有一定统计特征的脉冲星候选体[13]. 由于受射频或噪声等因素的干扰, 这些候选体中包含着大量的非脉冲星信号, 而脉冲星信号数量却非常少[14,15]. 为此, 需要对脉冲星候选体进行选择, 精选数据, 最后再利用射电望远镜对这些筛选后的数据进行人工分析以确定其是否为真实脉冲星[16]. 提高候选体选择的准确率能大幅减少候选体数量, 从而极大地减轻后期的人工验证工作. 因此, 提升候选体选择性能是搜索新脉冲星的一个关键步骤.

早期的脉冲星候选体选择主要依赖人工识别,

* 国家自然科学基金 (批准号: 61772187, 61873196) 资助的课题.

† 通信作者. E-mail: jt_zwkang@hnu.edu.cn

但这是一个主观耗时且易出错的过程. 一个现代脉冲星巡天项目可以产生数百万候选体, 仅依靠人工筛选效率极低且不切实际. 因此, 近几年来, 人们的研究主要集中在机器学习方法上. Eatough 等[17]提出了第一种用于解决脉冲星候选体选择问题的机器学习方法, 该方法将每个候选体简化为一个由 12 个数值特征组成的集合, 然后利用一个单隐层人工神经网络 (artificial neural networks, ANN) 从候选体中选择脉冲星. Bates 等[18]将特征增加到 22 个作为 ANN 的输入. Zhu 等[19]提出了深度神经网络图像模式识别方法——PICS (pulsar image-based classification system). PICS 将支持向量机、人工神经网络、卷积神经网络、逻辑回归等集成结合, 采用图像模式识别的方法验证候选体的真实性. Lyon 等[20]设计了 8 个特征应用到高斯-黑林格快速决策树算法. Mohamed[16]将 Lyon 等[20]设计的 8 个特征应用到模糊 k 近邻分类器上. Wang 等[21]在 Zhu 等[19]的基础上改进了 PICS 算法. 这些基于机器学习的脉冲星候选体选择方法, 有效节省了大量的人工劳动, 帮助研究人员发现了一些新的脉冲星.

如何进一步提升脉冲星候选体选择的准确率, 是机器学习方法有意义的研究点. 考虑到自归一化神经网络 (self-normalizing neural networks, SNN)[22]可以实现深层神经网络, 且通过激活函数“缩放指数线性单元 (scaled exponential linear units, SELU)”引入了自归一化属性, 从而避免了深层网络在训练时出现的梯度消失和爆炸问题, 保持网络的稳定性与收敛性. 本文利用 SNN 构建深层网络模型以提高候选体选择的精确性. 此外, 运用遗传算法 (genetic algorithm, GA) 优化候选体的特征子集, 采用合成少数类过采样技术 (synthetic minority over-sampling technique, SMOTE) 降低不平衡率, 这些对实现高精确性的候选体选择方法具有促进作用.

2 自归一化神经网络

SNN 也是由输入层、若干隐藏层及输出层组成, 每层又由多个单一神经元构成, 其中每个神经元代表一种特定的激活函数. SNN 的关键就是通过激活函数 SELU 引进自归一化属性, 即对具有零均值与单位方差的输入变量, 通过 SELU 激活函数后其输出仍将收敛于零均值和单位方差. 为确保每层激活函数的输入为零均值与单位方差, 还需进行权重初始化. SELU 激活函数与权重初始化是实现 SNN 自归一化特性的重点.

2.1 SELU 激活函数

SELU 激活函数表达式为

$$selu\,(x)=\begin{cases}\lambda x, & x>0,\\ \lambda(\alpha \mathrm{e}^{x}-\alpha), & x\leqslant 0,\end{cases} \tag{1}$$

其中 $\alpha=1.673268362\cdots$, $\lambda=1.050700987\cdots$. 图 1 给出了 SELU 激活函数图像,可以看出该激活函数具有以下特点: 1) 有用于控制平均值的负值和正值; 2) 存在饱和区域 (导数接近零), 以减小低层出现较大的方差; 3) 部分区域斜率大于 1, 如果下层方差太小则增加方差; 4) 是连续的曲线, 确保存在一个不动点, 且在该点处的方差减幅会被方差增长所补偿[22]. 这些特点使得深层神经网络在训练中都保持着方差稳定, 从而避免了梯度爆炸与梯度消失.

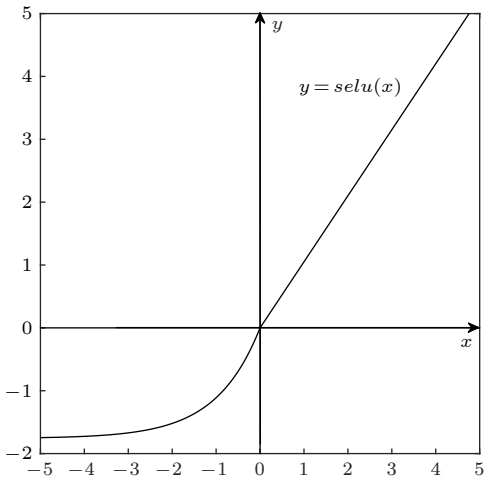

图 1 SELU 激活函数

Fig. 1. SELU activation function.

2.2 权重初始化

为确保每层激活函数的输入为零均值与单位方差, 还需进行权重初始化, 对此, 可证明如下:

考虑由一个权重矩阵 $\boldsymbol{W}$ 连接的两个连续的网络层, 下层网络的输出是上层网络的输入. 假定下层有 n 个神经元且其输出变量为 $\{z_{i,\mathrm{low}}\,|1\leqslant i\leqslant n\}$, 用 $\boldsymbol{z}_{\mathrm{low}}$ 代表其向量形式, 则上层神经元的输入 x_{up} 可以表示为

$$x_{\mathrm{up}}=\boldsymbol{z}_{\mathrm{low}}^{\mathrm{T}}w=\sum_{i=1}^{n}w_i z_{i,\mathrm{low}}, \tag{2}$$

其中 w 是 $\boldsymbol{W}$ 的一列向量. SELU 确保下层神经元输出具有零均值和单位方差, 即 $\mu = E\left(z_{i,\mathrm{low}}\right) \approx 0$, $v = \mathrm{Var}\left(z_{i,\mathrm{low}}\right) \approx 1$. 令权重初始化为

$$\mu_w = 0,\ \ \nu_w = 1/n, \tag{3}$$

用 $\tilde{\mu}$ 和 $\tilde{\nu}$ 分别代表 x_{up} 的均值和方差, 则有

$$\tilde{\mu} = E\left(x_{\mathrm{up}}\right) = E\left(\boldsymbol{z}_{\mathrm{low}}^{\mathrm{T}} w\right) = \sum_{i=1}^{n} E[z_{i,\mathrm{low}}] w_i \approx 0, \tag{4}$$

$$\tilde{\nu} = E[(x_{\mathrm{up}} - \tilde{\mu})^2] = E[{x_{\mathrm{up}}}^2]$$

$$= E\left[\left(z_{1,\mathrm{low}}^{\mathrm{T}} w_1 + \cdots z_{i,\mathrm{low}}^{\mathrm{T}} w_i + \cdots z_{n,\mathrm{low}}^{\mathrm{T}} w_n\right)^2\right], \tag{5}$$

其中

$$E[(z_{i,\mathrm{low}}^{\mathrm{T}} w_i)^2] = (\omega_i)^2 E[(z_{i,\mathrm{low}}^{\mathrm{T}})^2] = (\omega_i)^2,$$

$$E[z_{i,\mathrm{low}}^{\mathrm{T}} z_{j,\mathrm{low}}^{\mathrm{T}} \omega_i \omega_j] = \omega_i \omega_j E[z_{i,\mathrm{low}}^{\mathrm{T}}] E[z_{j,\mathrm{low}}^{\mathrm{T}}] \approx 0.$$

所以结合 (4) 式可得

$$\tilde{\nu} = \sum_{i=1}^{n} (w_i)^2 = n \cdot \nu_w \approx 1. \tag{6}$$

由此可知, 权重初始化确保了激活函数输入的归一化, 是 SELU 实现自归一化属性的一个必要条件.

3 脉冲星候选体选择方法

脉冲星候选体选择的目标就是尽可能地挑选出真实脉冲星候选体, 本文采用基于 SNN 的方法来提高候选体选择的精确性. SNN 可克服梯度消失与爆炸问题以提高训练速度, 深度神经网络结构可有效提高识别精度. GA 因其自适应性特别适合特征选择这一多目标优化任务 [23], 可用于优化特征子集. 而 SMOTE[24] 是一种不同于仅通过直接复制少数类样本的过采样技术, 因其简单有效适用于处理非平衡数据集. 因此本文提出了运用 GA 与 SMOTE 改进后的 SNN 模型 (GMO_SNN), 图 2 为 GMO_SNN 候选体选择算法流程图.

3.1 GA_特征选择算法

GMO_SNN 模型利用 GA 进行特征选择, 在原始特征空间中搜索最优特征子集. 用于特征选择的 GA 可以概括为三部分: 初始化种群、评估适应度、产生新种群.

初始化种群, 设定初始种群大小, 采用二进制进行基因编码, 长度为 L 的遗传个体编码后对应于一个 L 维的二进制基因串, 其中 L_i 为 1 表示第

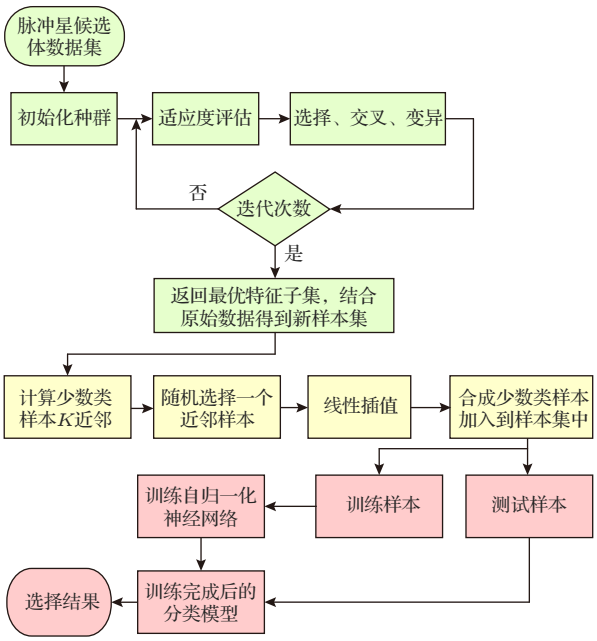

图 2 GMO_SNN 候选体选择算法流程图

Fig. 2. GMO_SNN candidate selection algorithm.

i 个特征包含于所选特征子集中, 否则 L_i 为 0. 例如: 有 6 个特征的特征集表示为⟨100100⟩, 则表示第 1 个与第 4 个特征被选中作为特征子集.

适应度函数的选择是 GA 中最关键的部分. 在特征选择问题中, 将 LightGBM 模型输出值作为遗传个体的适应值, 能直接反映不同特征组合对目标值的相关度, 适应值越高说明对应的特征组合越优良, 被选中的概率也越大.

产生新种群包括选择、交叉、变异, 具体采用轮盘赌算法作为选择算子, 定长基因段交叉算子, 基本位变异操作. 新的种群产生后, 通过适应度函数进行评估, 然后再选择、交叉、变异, 一直重复此步骤, 当遗传操作到达设定的最大迭代次数, 算法结束. 对末代种群中适应度值最大的个体进行解码, 就获得脉冲星候选体特征的最优子集.

3.2 SMOTE 算法

GMO_SNN 模型采用 SMOTE 算法解决脉冲星候选体的类不平衡问题. SMOTE 是一种过采样技术, 其利用 K 近邻与线性插值, 在距离较近的两个真实脉冲星候选体之间按照一定规则插入新的样本. 算法具体流程如下:

1) 对于真实脉冲星候选体中的每一个样本 r, 以欧氏距离为标准分别计算它到其他每个真实脉冲星样本的距离, 得到其 K 近邻, 一般 K 取值为 5.

2) 在每一个真实脉冲星样本 r 的 5 个近邻中随机选取一个样本, 假设选择近邻样本为 r_n.

3) 对于随机选出的近邻 r_n, 在其与 r 之间按照 (7) 式随机线性插值, 获得合成的真实脉冲星候选体样本 $\tilde{r}$:

$$\tilde{r} = r + \text{rand}(0,1) \times (r_n - r), \tag{7}$$

其中 rand(0, 1) 表示 0 到 1 之间的随机数.

3.3 GMO_SNN 候选体选择算法

首先采用 GA 进行特征选择, 找出可以分离脉冲星与非脉冲星的最优特征子集; 然后使用 SMOTE 合成新的脉冲星样本加入到数据集中; 最后将数据集分为训练集与测试集, 利用训练集对 SNN 进行训练, 训练完成后将测试集输入到神经网络中, 得到基于 GMO_SNN 模型的脉冲星候选体选择结果. 具体过程如图 2 所示.

4 实验与结果分析

在 3 个独立的脉冲星候选体数据集上进行实验, 根据 6 个典型的机器学习评价指标评估 GMO_SNN 模型性能. 在搭建自归一化神经网络时, 多次实验比较不同参数下的结果, 选择最优参数以使神经网络分类效果最佳, 并在相同网络结构下与传统 ANN 进行对比. 另外, 还分别将 GMO_SNN 与 SNN, GA-SNN (GA 特征选择后的 SNN 模型), MO-SNN (SMOTE 解决类不平衡问题后的 SNN 模型) 的候选体选择结果进行对比, 进一步证明本文方法的有效性.

实验环境为 Python3.6.4, 使用 Numpy1.14.0, Pandas0.22.0, Sklearn0.20.1 等机器学习库处理数据, 开发编译器 Spyder 调试算法; 利用 Keras 框架, 后端为 Tensorflow-GPU (NVIDIA GeForce GTX 1050) 搭建神经网络.

4.1 数据集与评价指标

3 个脉冲星候选体数据集分别为 HTRU 1[25], HTRU 2[20], LOTAAS 1[20]. 表 1 列出了 3 个数据集的非脉冲星数、脉冲星数以及总样本数. 在数据集中, 将脉冲星视为正样本, 将非脉冲星视为负样本. 3 个数据集中的候选体均采用 Bates 等[18] 提出的 22 个特征, 这些特征通过 Pulsar Feature Lab[20] 提供的工具获取. 表 2 列出了 22 个特征的具体描述, 这些特征由脉冲周期 P、脉冲宽度 W、脉冲轮廓信噪比 (signal-to-noise rate, S/N)、色散量 (dispersion measure, DM)、观测频率、观测时间等处理得到[18].

表 1 脉冲星候选体数据集

Table 1. Pulsar candidate datasets.

数据集	非脉冲星数	脉冲星数	总样本数
HTRU 1	89996	1196	91192
HTRU 2	16259	1639	17898
LOTAAS 1	4987	66	5053

在脉冲星候选体选择任务中, 使用准确率 (Accuracy)、查全率 (Recall)、查准率 (Precision)、假阳率 (false positive rate, FPR)、F1-分数 (F1-score)、G-均值 (G-mean)[26] 这 6 个评价指标对算法性能进行评估.

Accuracy 表示整体正确分类的比例, 但当测

表 2 特征描述

Table 2. Feature description.

编号	特征	编号	特征
1	P	12	轮廓直方图最大值/高斯拟合的最大值
2	DM	13	对轮廓求导后的直方图与轮廓直方图的偏移量
3	S/N	14	$S/N/\sqrt{(P-W)/W}$
4	W	15	拟合 $S/N/\sqrt{(P-W)/W}$
5	用sin曲线拟合脉冲轮廓的卡方值	16	DM拟合值与DM最优值取余
6	用$\sin^2$曲线拟合脉冲轮廓的卡方值	17	DM曲线拟合的卡方值
7	高斯拟合脉冲轮廓的卡方值	18	峰值处对应的所有频段值的均方根
8	高斯拟合脉冲轮廓的半高宽	19	任意两个频段线性相关度的均值
9	双高斯拟合脉冲轮廓的卡方值	20	线性相关度的和
10	双高斯拟合脉冲轮廓的平均半高宽	21	脉冲轮廓的波峰数
11	脉冲轮廓直方图对0的偏移量	22	脉冲轮廓减去均值后的面积

试集中非脉冲星占绝大多数时, 分类器可以通过将所有样本分类为负样本来获得高准确率, 因此对于非平衡数据集仅靠准确率来评价不够科学全面, 还需要其他评价指标. Recall 表示数据集中真实脉冲星候选体被正确分类的比例, 是评估脉冲星候选体选择模型一个非常重要的指标. 如果将一个真实脉冲星错误地归类为非脉冲星, 可能会漏掉脉冲星的新发现, 因此 Recall 越高, 分类器遗漏脉冲星的机率就越小. Precision 表示被归类为正样本中实际为正样本的比例, Precision 和 Recall 有时候会出现矛盾的情况, F1-score 则同时兼顾了这两者, 定义为 Precision 和 Recall 的调和平均, 是评价分类器分类少数类的综合指标. FPR 是非脉冲星被归类为真实脉冲星的比例, 当候选体选择完成之后, 会对被分类为真实脉冲星的候选体进行最终验证, 如果 FPR 太高, 会带来许多不必要的工作量. G-mean 是正负样本准确率的比值, 衡量在非平衡数据集下模型的综合性能.

4.2 参数设置

GA 中种群规模为 20, 种群最大遗传次数为 10 次, 适应度函数中使用的 LightGBM 模型使用默认参数; 自归一化网络结构采用“conic layers”设定隐藏单元数: 即从第一层中给定的隐藏单元数开始, 根据几何级数将隐藏单元的数目减小到输出层的大小[22]; 每个数据集使用 75% 的样本作为训练集, 余下作为测试集; 优化算法为“Adam”, 损失函数采用“交叉熵损失函数”. 通过实验分析, 神经网络相关参数设置如下.

1) 网络层数: 选择最佳结果 8 层.

2) 批次大小: 取 32 最佳.

3) 学习速率: 取 0.001 最佳.

4.3 结果分析

4.3.1 网络参数的最优选择

脉冲星候选体选择更加关注真实脉冲星候选体 (即少数类样本) 的分类准确率, 由于 F1-score 是评价分类器分类少数类的综合指标, 因此根据 3 个数据集上的平均 F1-score 值来确定参数, F1-score 值越高, 神经网络分类效果越好.

1) 网络层数的最优选择

深层次的网络结构通常会获得更好的分类效果, 但随着网络层数的增大, 网络结构也越复杂. 本文分别对隐藏层数为 2, 4, 8, 9 的网络进行实验, 表 3 列出了不同隐藏层数下的平均 F1-score 值. 由表 3 可知, 当隐藏层数为 8 层时效果最佳.

表 3 不同隐藏层数下的分类效果

Table 3. Classification results with the different hidden layers.

隐藏层数	F1-score/%
2	82.48
4	89.58
8	94.56
9	94.20

2) 批次大小的最优选择

为了提高神经网络的训练效率, 将训练样本分批次输入. 批次大小会对模型优化程度和训练速度产生影响. 若批训练量过小, 会增加网络训练时间; 如果批训练过大, 其分类效果会变差. 本文分别对批次大小为 16, 32, 64, 128 的模型进行训练, 表 4 列出了不同批次大小下的平均 F1-score 值及运行时间. 由表 4 可知, 随着批次减小, F1-score 值在逐步上升, 但运行时间也有明显的增加. 当批次大小为 16 时, 其 F1-score 值对比批次为 32 时只上升了 0.0031, 但其运行时间却增加了一倍. 因此综合考虑分类效果与算法运行时间, 本文神经网络的批次大小取 32.

表 4 不同批次大小下的分类效果

Table 4. Classification results with the different batch size.

批次大小	F1-score/%	运行时间/s
16	94.87	74
32	94.56	43
64	93.90	23
128	91.05	11

3) 学习速率的最优选择

学习速率是影响网络性能的一个重要参数. 过大导致损失函数振荡, 神经网络无法收敛; 过小会导致收敛速度过慢, 可能会陷入局部最优. 本文分别对学习速率为 0.1, 0.01, 0.001, 0.0001 时的模型进行训练, 表 5 列出了迭代 10 次后不同学习速率下的平均 F1-score 值. 由表 5 可知, 在相同的迭代次数下, 当学习速率减小时, F1-score 值会降低, 模型分类效果变差. 当学习速率增大到 0.1, 此时算法无法优化, 因此学习速率取值 0.001 最佳.

表 5 不同学习速率的分类效果

Table 5. Classification results with the different learning rates.

隐藏层数	F1-score/%
0.1	无法收敛
0.01	94.29
0.001	94.55
0.0001	84.10

4.3.2 不同方法的比较

为证明 SNN 的有效性, 本文对 SNN 与传统 ANN 在 HTRU 2 数据集上进行对比实验, 图 3 给出了 8 层神经网络训练过程中的损失函数曲线对比图, 迭代次数为 100 次. 损失函数是用来衡量模型预测值与真实值的不一致程度, 损失函数越小, 模型鲁棒性就越好. 由图 3 可知 SNN 模型比传统 ANN 具有更低的误差, 且其收敛速度明显大于 ANN, 证明了 SNN 在深层网络中的有效性.

表 6 分别列出了 3 个数据集上 SNN, GA_SNN, MO_SNN, GMO_SNN 的脉冲星候选体选择结果, 最优结果加粗表示.

利用 GA 进行特征选择, 从候选体样本的 22 个特征中筛选出 8 个作为最优特征子集, 数据集缩减率达到 63%. 以 HTRU 1 数据集为例, 对比表 6 中 GA_SNN 与 SNN 的选择结果可知, 利用最优特征子集训练分类模型, 其结果均表现出不同程度的优化, 其余两个数据集除少数几个评价指标外, 也达到了类似的效果. 表明该特征选择算法可以在压缩特征空间的同时又不丢失原有信息, 提升模型性能.

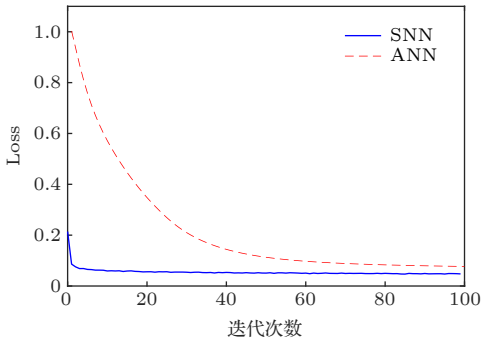

图 3 SNN 与 ANN 损失函数的对比

Fig. 3. Comparison of the loss function between SNN and ANN.

由表 6 中 SNN 与 MO_SNN 的评价指标可知, 利用 SMOTE 处理类不平衡问题后, Recall 值在 HTRU 1 与 HTRU 2 数据集上分别提高了 1.79 和 4.44 个百分点, 其中 LOTAAS 1 数据集上 Recall 值达到 100%, 说明该方法使分类器对非平衡学习问题具有较强的鲁棒性, 防止了分类器在训练时向丰富的非脉冲星类倾斜.

由表 6 可知, 在 3 个数据集上, 本文提出的 GMO_SNN 模型在 Recall, Precision, F1_score, FPR 以及 G_mean 上均优于其他模型. 例如 HTRU 1 数据集, 其 Recall 值为 95.53, FPR 仅有 0.03, 说明该方法既能有效避免脉冲星的遗漏, 又能减少需要人工再次验证的非脉冲星候选体, 进一步证明了本文方法的有效性.

表 6 不同方法在 3 个数据集上的分类效果

Table 6. Classification results with different methods on three datasets.

数据集	模型	Accuracy/%	Recall/%	Precison/%	F1-score/%	FPR/%	G-mean/%
HTRU 1	SNN	99.82	92.44	93.45	92.94	0.08	96.11
	GA_SNN	99.85	92.45	95.19	93.80	0.06	96.12
	MO_SNN	99.81	94.23	97.94	96.05	0.05	97.05
	GMO_SNNNNNNNNN	**99.85**	**95.32**	**98.51**	**96.89**	**0.04**	**97.61**
HTRU 2	SNN	98.30	87.73	93.93	90.73	0.59	93.38
	GA_SNN	**98.30**	88.91	92.86	90.84	0.71	93.96
	MO_SNN	97.89	92.17	95.08	93.60	0.95	95.54
	GMO_SNNNNNNNNN	98.03	**92.53**	**95.58**	**94.03**	**0.08**	**95.78**
LOTAAS 1	SNN	99.92	93.75	100.00	96.77	0.08	96.79
	GA_SNN	99.92	100.00	93.33	96.55	0	100.00
	MO_SNN	99.69	100.00	87.10	93.10	0.31	99.84
	GMO_SNN	**100.00**	**100.00**	**100.00**	**100.00**	**0**	**100.00**

位于中国贵州省的 500 米口径球面射电望远镜 (five-hundred-meter aperture spherical radio telescope, FAST) 是目前世界上最大、最灵敏的射电天文望远镜, 其主要科学目标之一就是开展脉冲星的搜寻[27]. FAST 采用 19 波束接收机进行巡天, 可产生上亿量级的脉冲星候选体[13]. 本文的候选体选择模型运用机器学习方法提高了筛选速度, 使用单个 GPU 每秒可以识别约 2 万个候选体, 同时得到高精度的选择结果. 这种速度和效率的提高能促进对 FAST 巡天产生的脉冲星候选体数据的实时处理, 可减小大数据量带来的筛选难度.

5 结 论

基于自归一化神经网络的脉冲星候选体选择是一种能高准确率识别真实脉冲星的有效方法. 利用 GA 进行特征选择, 能在压缩特征空间的同时又不丢失原有信息, 提升模型性能; 使用 SMOTE 处理非平衡数据集, 可降低数据集的不平衡率, 提高了分类器对少数类样本的识别能力; 采用自归一化神经网络比传统人工神经网络在深层结构中具有更高的准确率以及更快的收敛速度. 在 3 个脉冲星候选体数据集上的实验结果表明, 该方法既能有效避免真实脉冲星的遗漏, 又能减少非脉冲星的保留, 从而提高脉冲星搜寻的工作效率.

参考文献

[1] Sun H F, Xie K, Li X P, Fang H Y, Liu X P, Fu L Z, Sun H J, Xue M F 2013 *Acta Phys. Sin.* **62** 109701 (in Chinese) [孙海峰, 谢楷, 李小平, 方海燕, 刘秀平, 傅灵忠, 孙海建, 薛梦凡 2013 物理学报 **62** 109701]

[2] Heiles C, Li D, Mcclure-Griffiths N, Qian L, Liu S 2019 *Res. Astron. Astrophys.* **19** 5

[3] Yi S X, Zhang S N 2016 *Sci. China, Phys. Mech. Astron.* **59** 689511

[4] Liu J, Ning X L, Ma X, Fang J C 2019 *IEEE Trans. Aerosp. Electron. Syst.* **55** 2556

[5] Kang Z W, Wu C Y, Liu J, Ma X, Gui M Z 2018 *Acta Phys. Sin.* **67** 099701 (in Chinese) [康志伟, 吴春艳, 刘劲, 马辛, 桂明臻 2018 物理学报 **67** 099701]

[6] Fang J C, Ning X L, Liu J 2017 *Principles and Methods of Spacecraft Celestial Navigation* (2nd Ed.) (Beijing: National Defense Industry Press) p8 (in Chinese) [房建成, 宁晓琳, 刘劲 2017 航天器自主天文导航原理与技术 (第二版) (北京: 国防工业出版社) 第8页]

[7] Hewish A, Bell S J, Pilkington J D H, Scott P F, Collins R A 1968 *Nature* **217** 709

[8] Thornton D 2013 *Ph. D. Dissertation* (Manchester: University of Manchester)

[9] Stovall K, Lynch R S, Ransom S M, et al. 2014 *Astrophys. J.* **791** 67

[10] Manchester R N, Lyne A G, Camilo F, Bell J F, Kaspi V M, D'Amico N, McKay N P F, Crawford F, Stairs I H, Possenti A, Kramer M, Sheppard D C 2001 *Mon. Not. R. Astron. Soc.* **328** 17

[11] Keith M, Jameson A, Van Straten W, Bailes M, Johnston S, Kramer M, Possenti A, Bates S, Bhat N, Burgay M 2010 *Mon. Not. R. Astron. Soc.* **409** 619

[12] van Leeuwen J, Stappers B W 2010 *Astron. Astrophys.* **509** A7

[13] Xu Y Y, Li D, Liu Z J, Wang C, Wang P, Zhang L, Pan Z C 2017 *Prog. Astron.* **35** 304 (in Chinese) [许余云, 李菂, 刘志杰, 王晨, 王培, 张蕾, 潘之辰 2017 天文学进展 **35** 304]

[14] Wang Y C, Zheng J H, Pan Z C, Li M T 2018 *J. Deep Space Explor.* **5** 203 (in Chinese) [王元超, 郑建华, 潘之辰, 李明涛 2018 深空探测学报 **5** 203]

[15] Lee K J, Stovall K, Jenet F A, Martinez J, Dartez L P, Mata A, Lunsford G, Cohen S, Biwer C M, Rohr M D 2013 *Mon. Not. R. Astron. Soc.* **433** 688

[16] Mohamed T M 2018 *Futur. Comput. Inf. J.* **3** 1

[17] Eatough R P, Molkenthin N, Kramer M, Noutsos A, Keith M J, Stappers B W, Lyne A G 2010 *Mon. Not. R. Astron. Soc.* **407** 2443

[18] Bates S D, Bailes M, Barsdell B R, Bhat N D R, Burgay M, Burke-Spolaor S, Champion D J, Coster P, D'Amico N, Jameson A, Johnston S, Keith M J, Kramer M, Levin L, Lyne A, Milia S, Ng C, Nietner C, Possenti A, Stappers B, Thornton D, van Straten W 2012 *Mon. Not. R. Astron. Soc.* **427** 1052

[19] Zhu W W, Berndsen A, Madsen E C, et al. 2014 *Astrophys. J.* **781** 117

[20] Lyon R J, Stappers B W, Cooper S, Brooke J M, Knowles J D 2016 *Mon. Not. R. Astron. Soc.* **459** 1104

[21] Wang H F, Zhu W W, Guo P, Li D, Feng S B, Yin Q, Miao C C, Tao Z Z, Pan Z C, Wang P, Zheng X, Deng X D, Liu Z J, Xie X Y, Yu X H, You S P, Zhang H 2019 *Sci. China, Phys. Mech. Astron.* **62** 959507

[22] Klambauer G, Unterthiner T, Mayr A, Hochreiter S 2017 *Advances in Neural Information Processing Systems*, Long Beach, USA, December 4–9, 2017 p971

[23] Oh I S, Lee J S, Moon B R 2004 *IEEE Trans. Pattern Anal. Mach. Intell.* **26** 1424

[24] Chawla N V, Bowyer K W, Hall L O, Kegelmeyer W P 2002 *J. Artif. Intell. Res.* **16** 321

[25] Morello V, Barr E D, Bailes M, Flynn C M, Keane E F, van Straten W 2014 *Mon. Not. R. Astron. Soc.* **443** 1651

[26] Yao Y, Xin X, Guo P 2016 *12th International Conference on Computational Intelligence and Security*, Wuxi, China, December 16–19, 2016 p120

[27] Nan R D, Li D, Jin C J, Wang Q M, Zhu L C, Zhu W B, Zhang H Y, Yue Y L, Qian L 2011 *Int. J. Mod. Phys. D.* **20** 989

Pulsar candidate selection based on self-normalizing neural networks*

Kang Zhi-Wei [1)†] Liu Tuo [1)] Liu Jin [2)] Ma Xin [3)] Chen Xiao [4)]

1) (*College of Computer Science and Electronic Engineering, Hunan University, Changsha* 410082, *China*)

2) (*College of Information Science and Engineering, Wuhan University of Science and Technology, Wuhan* 430081, *China*)

3) (*College of Instrument Science and Opto Electronic Engineering, Beihang University, Beijing* 100191, *China*)

4) (*Shanghai Institution of Satellite Engineering, Shanghai* 200240, *China*)

(Received 17 October 2019; revised manuscript received 19 December 2019)

Abstract

Pulsar candidate selection is an important step in the search task of pulsars. The traditional candidate selection is heavily dependent on human inspection. However, the human inspection is a subjective, time consuming, and error-prone process. A modern radio telescopes pulsar survey project can produce totally millions of candidates, so the manual selection becomes extremely difficult and inefficient due to a large number of candidates. Therefore, this study focuses on machine learning developed in recent years. In order to improve the efficiency of pulsar candidate selection, we propose a candidate selection method based on self-normalizing neural networks. This method uses three techniques: self-normalizing neural networks, genetic algorithm and synthetic minority over-sampling technique. The self-normalizing neural networks can improve the identification accuracy by applying deep neural networks to pulsar candidate selection. At the same time, it solves the problem of gradient disappearance and explosion in the training process of deep neural networks by using its self-normalizing property, which greatly accelerates the training process. In addition, in order to eliminate the redundancy of the sample data, we use genetic algorithm to choose sample features of pulsar candidates. The genetic algorithm for feature selection can be summarized into three steps: initializing population, assessing population fitness, and generating new populations. Decoding the individual with the largest fitness value in the last generation population, we can obtain the best subset of features. Due to radio frequency interference or noise, there are a large number of non-pulsar signals in candidates, and only a few real pulsar signals exist there. Aiming at solving the severe class imbalance problem, we use the synthetic minority over-sampling technique to increase the pulsar candidates (minority class) and reduce the imbalance degree of data. By using k-nearest neighbor and linear interpolation to insert a new sample between two minority classes of samples that are close to each other according to certain rules, we can prevent the classifier from becoming biased towards the abundant non-pulsar class (majority class). Experimental results on three pulsar candidate datasets show that the self-normalizing neural network has higher accuracy and faster convergence speed than the traditional artificial neural network in the deep structure, By using the genetic algorithm and synthetic minority over-sampling technique, the selection performance of pulsar candidates can be effectively improved.

Keywords: pulsar candidate selection, self-normalizing neural networks, feature selection, class imbalance

PACS: 97.60.Gb, 98.52.Cf, 07.05.Mh **DOI:** 10.7498/aps.69.20191582

* Project supported by the National Natural Science Foundation of China (Grant Nos. 61772187, 61873196).

† Corresponding author. E-mail: jt_zwkang@hnu.edu.cn