第 35 卷 第 3 期
2017 年 8 月

天 文 学 进 展
PROGRESS IN ASTRONOMY

Vol. 35, No. 3
Aug., 2017

doi: 10.3969/j.issn.1000-8349.2017.03.03

人工智能在脉冲星候选体筛选中的应用

许余云[1]，李 茹[2,3]，刘志杰[1]，王 晨[4]，王 培[2]，张 蕾[2]，潘之辰[2]

(1. 贵州师范大学 贵州省信息与计算科学重点实验室，贵阳 550001； 2. 中国科学院 国家天文台，北京 100012； 3. 中国科学院 射电天文实验室，南京 210008； 4. 澳大利亚联邦科学与工业研究组织，堪培拉 ACT 2601，澳大利亚)

摘要： 脉冲星搜寻是对脉冲星、引力波，以及对快速射电暴 (Fast Radio Burst，简称 FRB) 等暂现源进行研究的基础。搜寻不仅可以扩大脉冲星样本，还可以发现极端性质的致密星。这有助于研究致密天体状态方程、星际介质、脉冲星导航、引力波探测等课题。目前，射电望远镜的单次巡天就可以产生百万数量级的脉冲星候选体。面对这些海量数据，仅仅依赖人工识别筛选，已不能满足数据的时效需求，更不能实现数据的实时处理。机器学习、计算机视觉应用等人工智能技术自诞生以来，其理论和技术已日益发展成熟，并已成功运用到脉冲星候选体筛选等射电天文研究领域。首先将介绍现有脉冲星搜寻的人工智能方法，再统计和分析已有脉冲星候选体筛选方法的性能，最后对 FAST 脉冲星候选体筛选工作进行展望。

关 键 词： 人工智能；脉冲星；候选体筛选

中图分类号： P162.4 **文献标识码：** A

1 引 言

自 1967 年人们发现第一颗脉冲星[1]并验证中子星假说，至今已发现 2 500 多颗脉冲星[2]。借助高灵敏度的射电望远镜，有助于探测更多的脉冲星。500 m 口径球面射电望远镜 (Five-hundred-meter Aperture radio Spherical Telescope，简称 FAST) 已于 2016 年 9 月在我国贵州落成[3−5]，其高灵敏度蕴含发现大量新脉冲星的巨大科学机遇。2015 年，Zhang 等人[6]对银河系内的脉冲星进行了模拟，结果表明银河系内可观测脉冲星超过 140 000 颗，在距太阳 1 kpc 范围内约 600 颗。在此工作基础上，又进一步模拟 FAST 采用漂移扫描球状星团中脉冲星的探测率，结果显示，在 FAST 10 次漂移扫描内，40 个球状星团中有 10 个球状星团就可能发现新脉冲星[7]。脉冲星漂移扫描观测及巡天将会是 FAST 建成后开展的重要科

收稿日期： 2016-11-25； **修回日期：** 2017-3-27

资助项目： 研究生创新基金项目 (研创 201528)；国家自然科学基金 (U1631132)；中国科学院国际合作局对外合作重点项目 (114A11KYSB20160008)；中国科学院战略性先导科技专项 (B 类)(XDB23000000)

通讯作者： 李茹，dili@nao.cas.cn；许余云，yuyunxu@outlook.com

学项目之一。

脉冲星搜索是寻找带有色散的周期性脉冲的过程。经过消色散、傅里叶变换以及周期信号搜索等过程，人们得到疑似脉冲星信号[8]，该信号是真实脉冲星信号的候选体。单次大型脉冲星巡天可以产生百万量级的脉冲星候选体，其中大多数候选体为干扰信号，如人为产生的无线电干扰，其他形式的地面信号。采用人工方法筛选出此类信号效率很低，核对速度难以超过每人每秒 1 个。目前，FAST 测试阶段接收数据速率约 100 MB/s[9]，建成后一两年内巡天数据量将增长至拍 (PetaByte，简称 PB) 量级，预计将产生千万量级的候选体，人工看图、图形工具辅助或基于统计的传统方法无法满足候选体筛选的需要。近些年，人工智能 (Artificial Intelligence，简称 AI) 在计算机领域得到广泛的应用，不仅成熟应用于机器视觉、人脸识别和图像理解等方面[10]，而且逐步应用于射电脉冲星的搜索项目中。例如，Lee 等人 (2013) 提到的 PEACE[11]，基于分数函数对候选体进行排序的机器学习 (Machine Learning，简称 ML) 方法。PEACE 被用于对阿雷西博 L 波段馈源阵列脉冲星巡天 (Pulsar Arecibo L-band Feed Array survey，简称 PALFA)[12, 14, 15, 34, 47]、绿岸北半球脉冲星巡天 (Green Bank Northern Celestial Cap pulsar survey，简称 GBNCC)[16] 和北天中银纬高时间分辨率宇宙脉冲星巡天 (North High Time Resolution Universe pulsar survey，简称 HTRU North)[17] 数据的处理，发现了 47 颗新脉冲星，其中 5 颗是毫秒脉冲星 ((Millisecond Pulsars，简称 MSPs)。AI 技术对脉冲星候选体筛选具有高效准确的巨大优势。目前，应用于脉冲星候选体筛选的 AI 技术主要有人工神经网络 (Artificial Neural Network，简称 ANN)、图像模式识别等 ML 算法[18–22]。本文将综述应用于脉冲星搜寻项目的 AI 技术。

本文第 2 章将详细介绍脉冲星候选体可能存在的脉冲星信号特征，并介绍传统筛选方法，然后分析影响筛选效果的因素和引入 AI 技术的重要性；第 3 章综述目前应用 AI 技术对脉冲星候选体进行筛选的几个主要方法，并分析其性能；第 4 章对脉冲星候选体筛选方法进行总结；第 5 章结合当前 AI 技术发展成果和趋势展望 FAST 筛选工作。

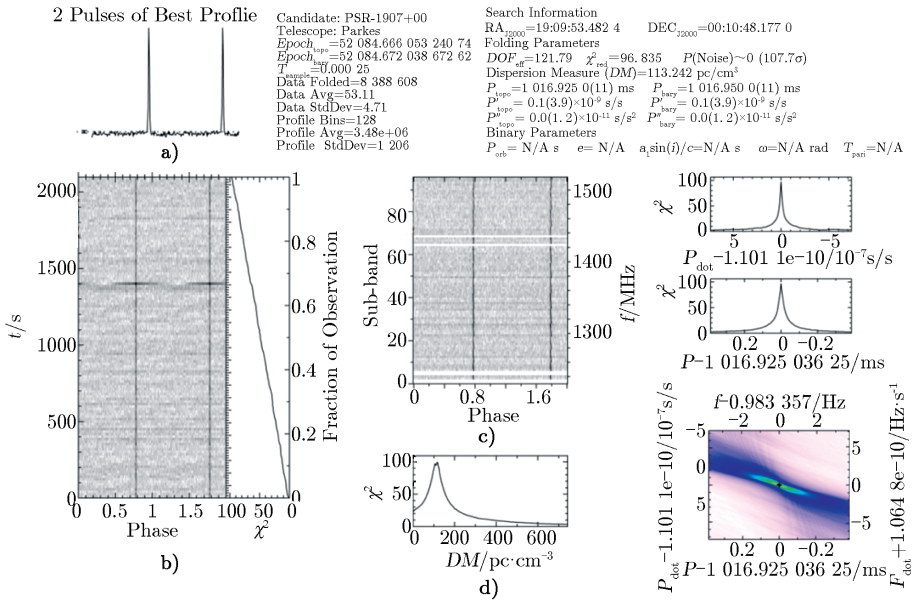
2　脉冲星候选体的特征、筛选方法以及影响筛选的因素

2.1　脉冲星候选体的主要特征

观测数据经过色散、周期折叠处理后得到可能的脉冲星旋转参数组合和统计量分布结果，称为脉冲星候选体[8, 23, 24]。已知脉冲星中，约有 2 100 颗的周期大于 100 ms，平均脉冲信号占空比约 3%，以脉冲星 B1907+00 为例[2]，用于识别的脉冲星主要特征如图 1 所示。

(1) 平均脉冲轮廓。由数据按频率和时间两个维度叠加后得到。不同脉冲星平均脉冲轮廓各异，通常由一个或多个峰组成，也有一些已知脉冲星有较宽的平均脉冲轮廓。

(2) 时间–相位图像。由数据消色散后按照周期 (相位) 排列得到。图像中垂直贯穿标识脉冲信号同时到达。如果脉冲的相位大致恒定 (即不随时间改变)，黑色区域表现为成一条竖线。而脉冲双星，则呈现出脉冲到达时间受轨道运动调制及因可能的掩食效应而出现短时脉冲消零。



注：a) 平均脉冲轮廓，b) 时间–相位图像，c) 频率–相位图像，d) DM 曲线。

图 1 脉冲星候选体诊断图

(3) 频率–相位图像。由数据按时间维折叠并消色散后得到。脉冲信号在图像垂直贯穿标识观测频率范围上接受的不同辐射频率脉冲信号同时到达。

(4) 色散量 (Dispersion Measurement，简称 DM) 曲线。每个 DM 试验值的 χ^2 分布。峰值对应物理色散值，如果洛伦兹峰的形状越明显，说明信号具有的色散特征越显著；如果峰值为零，则色散为零，即为地面干扰信号。

2.2 脉冲星候选体筛选难度和传统方法

脉冲星候选体的筛选，依据 2.1 节中描述的脉冲星候选体主要特征作为判断的标准，筛选的难度主要有：(1) 平均脉冲轮廓有单峰、双峰和多峰，形状随着观测频率的变化而变化，有些轮廓较宽；(2) 时间–相位和频率–相位图像中，竖线可能比较淡、不连贯、很短，若是双星系统，轨道运动会使信号在时间–相位图上显示倾斜的黑色曲线；(3) DM 曲线，峰值明显的只占很少的一部分，较弱的脉冲星则不明显。

依据脉冲星候选体特征筛选，数百万量级的候选体文件需要投入较大人力。为了节约人力和提高脉冲星识别效率，2006 年德州大学布朗斯威尔分校 (University of Texas at Brownsville，简称 UTB) 主导的阿雷西博远程指挥中心 (Arecibo Remote Command Center program，简称 ARCC)[①] 开发了一套允许多用户浏览和排序脉冲星候选体的 web 程序[25]，吸引高校学生参与，协同科研人员一起实时执行观察和数据分析的工作。人工判断存在主观性，在海量候选体筛选过程中，微弱的、受干扰较强的脉冲信号可能会遗漏，因此大多数巡天数

[①] http://venus.fandm.edu/pulsar/arcc/

据需要再处理。

2004 年，Faulkner 等人开发了包含图像化接口的程序 REAPER[26]：用户界面由一双可选参数的 X-Y 坐标系构成，在二维轴中每个候选体作为单个点显示，系统在脉冲周期和信噪比基础上交互式筛选候选体。REAPER 很难检测到信噪比很低的真实脉冲星，因此，2009 年 Keith 等人开发了 JREAPER[27]，使用一系列脉冲星信号属性，提出了对候选体打分和排序方法；并且提供 REAPER 功能，增加了谐波检测、着色数据点生成“3-D”图和 JREAPER 界面筛选功能，可用于不同格式的脉冲星巡天数据。JREAPER 系统评分算法，评分程序使用许多检验，计算每个脉冲星候选体的脉冲相位–观测时间图得分 (即子积分图)、DM 曲线得分和轮廓得分；最后由用户指定权重，给出每个候选体最终得分，然后进行排序。因此，JREAPER 的有效性也依赖于用户的权重设定。JPEAPER 能寻找较弱的脉冲星，其中得分函数选择具有典型的脉冲星信号属性，只能检测典型的脉冲星，用这种方法对 Parkes 多波束脉冲星巡天 (Parkes multi-beam pulsar survey，简称 PMPS)[28–30] 数据再处理，发现了 28 颗新的脉冲星。

3　脉冲星候选体自动筛选 AI 方法

从 1967 年第一颗脉冲星被发现，在之后的近 50 年里，脉冲星候选体的筛选工作经历了各种方法：从简单的基于脉冲轮廓和信噪比去筛选候选体，到图形工具辅助[31, 32] 和基于 web 候选体筛选系统[12, 25, 34]，以及近几年基于经验公式的候选体排序算法[11]、基于机器学习的分类算法[18–22]。

3.1　基于经验公式的排序算法

Lee 等人 (2013) 提出了 PEACE(Pulsar Evaluation Algorithm for Candidate Extraction) 排序算法[11]，利用了候选体的周期、DM、脉冲宽度等性质。PEACE 分析脉冲星搜寻软件生成的候选体文件，计算出诸如信噪比、脉冲轮廓宽度等评价因子，再通过对评价因子加权计算，得出每个候选体文件的评分，再排序。

PEACE 将候选体特征总结为 6 个评价因子。

(1) 信噪比 (S/N)。定义为脉冲峰值与信号均方根值之比，即读入脉冲形状数据，判断峰值所处位置，并计算信号的均方根，从而得到脉冲信噪比。

(2) 脉冲周期 ($p_{\rm top}$)。直接从脉冲星候选体文件中读取周期数据。

(3) 脉冲轮廓宽度 (w)。定义为脉冲实际宽度与信号周期之比，$w \in$[0,1]。PEACE 先用多个高斯函数对脉冲轮廓进行拟合，并用这些高斯函数半高全宽之和作为脉冲实际宽度。脉冲星的脉冲宽度一般小于 10%，而射频干扰的脉冲宽度一般较宽。

(4) 时域脉冲持续长度 (η_T)。在时域上脉冲信号长度占全部信号长度的比例。

候选体数据是一个三维数组，三个维度为时间、频率和相位。PEACE 从候选体文件中读取脉冲信号，按相位维折叠，并按式 (1) 计算得到脉冲与非脉冲的信号强度比：

$$\gamma_T = \frac{\text{脉冲信号强度平均值}}{\text{非脉冲信号强度平均值}} \quad , \tag{1}$$

脉冲信号强度定义为相位位于脉冲半高全宽范围内的数据点的辐射强度，反之亦然。定义时域脉冲持续量 η_T：

$$\eta_T = \frac{\text{满足}\gamma_T > \alpha_T\text{的积分区间数}}{\text{总积分区间数}} \quad , \tag{2}$$

α_T 是人为设定的一个阀值，在 PEACE 中的缺省值为 1，根据 η_T 的定义，应有 $\eta_T \in$[0,1]。由于来自脉冲星的脉冲信号应持续于整个观测阶段，故 η_T 接近 1。

(5) 频域脉冲持续长度 (η_F)。在频域上，与在时域上类似可定义 γ_F：

$$\gamma_F = \frac{\text{频域脉冲信号强度平均值}}{\text{频域非脉冲信号强度平均值}} \quad , \tag{3}$$

脉冲信号强度定义为相位位于脉冲半高全宽范围内的数据点的辐射强度，反之亦然。定义时域脉冲持续量 η_F：

$$\eta_F = \frac{\text{满足}\gamma_F > \alpha_T\text{的频率积分区间数}}{\text{总频率积分区间数}} \quad , \tag{4}$$

由于脉冲信号是宽带信号，所以 η_F 应接近 1。

(6) 脉冲宽度与 DM 涂污时间之比 ($\eta_{\rm DM}$)。PEACE 从候选体文件中读出质心系周期 ($p_{\rm bar}$)，频道宽度 ($\triangle f_c$)，中心频率 (f) 和 DM 值，计算出由于散射带来的涂污时间：

$$\triangle\tau = 8.3\mu{\rm s}(\frac{\triangle f_{\rm c}}{\rm MHz})(\frac{f}{\rm GHz})^{-3}(\frac{DM}{\rm cm^{-3}pc}) \quad , \tag{5}$$

定义 $\eta_{\rm DM}$：

$$\eta_{\rm DM} = \frac{p_{\rm bar}w}{\triangle\tau} \quad , \tag{6}$$

由于实测脉冲宽度总大于一个频段内的 DM 涂污宽度，故对于真实脉冲信号应有 $\eta_{\rm DM}$ >1。

每个候选体的评分是上述评价因子的线性组合：

$$\begin{aligned} S = {} & \beta_{S/N}S_{S/N}(S/N) + \beta_{\rm p_{top}}S_{\rm p_{top}}(p_{\rm top}) + \beta_w S_w(w) + \\ & \beta_{\eta_T}S_{\eta_T}(\eta_T) + \beta_{\eta_F}S_{\eta_F}(\eta_F) + \beta_{\eta_{\rm DM}}S_{\eta_{\rm DM}}(\eta_{\rm DM}) \quad , \end{aligned} \tag{7}$$

其中，权系数 $\beta_{S/N}$, $\beta_{p_{\rm top}}$, β_w, β_{η_T}, β_{η_F}, $\beta_{\eta_{\rm DM}}$ 为常数，缺省值为 1。$S_{S/N}$, S_w, S_{η_T}, S_{η_F}, $S_{\eta_{\rm DM}}$ 的函数如下所示：

$$S_{S/N}(S/N) = \begin{cases} -(S/N-5)^2 & S/N \le 5 \\ 0 & S/N > 5 \end{cases} \quad , \tag{8}$$

$$S_w(w) = \begin{cases} -280.7w^2 + 11.4w + 1.6 & w < 0.125 \\ -37.9w^2 - 4.1w - 4.0 & 0.125 \le w < 0.6 \\ -20 & 0.6 \le w \end{cases} \quad , \tag{9}$$

$$S_{\eta_T}(\eta_T) = -9(\eta_T - 1)^2 \quad , \tag{10}$$

$$S_{\eta_F}(\eta_F) = -9(\eta_F - 1)^2 \quad , \tag{11}$$

$$S_{\eta_{\rm DM}}(\eta_{\rm DM})=\begin{cases}-10.2 & \eta_{\rm DM}<0.4\\ -4(\eta_{\rm DM}-2)^2 & 0.4\le\eta_{\rm DM}<2\\ 0 & 2\le\eta_{\rm DM}\end{cases}. \tag{12}$$

$S_{p_{\rm top}}$ 函数需要根据当地的射频干扰环境确定，PEACE 提供了一个工具——bulidSP，可以用一列候选体的周期来生成 $S_{p_{\rm top}}$ 函数。对比 ARCC，结果表明 PEACE 排序靠前的 0.17% 候选体包含已识别的 68% 脉冲星，靠前的 0.34% 包含 95%，靠前 3.7% 包含 100%。然而，PEACE 依赖于一些假设，如脉冲轮廓为高斯形状，实际中很多都不能很好拟合，会导致脉冲形状中细节被掩盖，所以 PEACE 可能导致一些有特殊形状脉冲，如宽脉冲、偏 DM 曲线或者低流量的脉冲星被遗漏。

3.2　基于机器学习的算法

近几年，监督学习算法，特别是基于神经网络的算法也被用于脉冲星自动筛选。Eatough 等人 (2010) 使用人工神经网络算法自动识别脉冲星候选体[18]，并且运用该方法在 PMPS[8] 数据中发现了一颗新脉冲星 J1926+0739。这种方法使用三层神经网络，前后运用了 8:8:2 和 12:12:2 (每个数字代表相应层的神经元数目) 两种结构训练。输入到人工神经网络具有特征的输入向量是由各种候选参数形成的，输入向量 (如表 1 所示) 和隐藏层有同等数量的元素，随后输出层包括两个元素，代表候选体的概率是一个真正的脉冲星 (0 − 1) 和非脉冲星 (1 − 0)。

表 1　2 种人工神经网络的参数[18]

编号	分数表述
1	Pulse profile SNR
2	Pulse profile width
3	Chi-square of fit to theoretical DM-SNR curve
4	No· of DM trials with $S/N>10$
5	Chi-square of fit to optimized theoretical DM-SNR curve
6	Chi-square of fit to theoretical acceleration-SNR curve
7	No. of acceleration trials with $S/N>10$
8	Chi-square of fit to optimized theoretical acceleration-SNR curve
9	RMS scatter in subband maxima ♣
10	Linear correlation across subbands ♣
11	RMS scatter in subintegration maxima ♣
12	Linear correlation across subintegrations ♣

注：编号 1 到 8 用于 8:8:2 网络，9 到 12 带 ♣ 只添加进 12:12:2 网络

该扩展评分方法，使用斯图加特神经网络模拟器 (Stuttgart Neural Network Simulator，简称 SNNS)[35] 实现，构建的是弹性反向传播 (Resilient backPROPagation，简称 RPROP)[36] 人工神经网络，训练过程中输入向量 x=x_1,x_2,x_3,....,x_l 及其权值，形成一个

线性组合 S_j^y 如下：

$$net_k = x_0 w_{k0} + x_1 w_{k1} + ... + x_m w_{km} = \sum_{i=0}^{m} x_i w_{ki} = X \cdot W = S_{\mathrm{j}}^y \quad . \tag{13}$$

激活函数为对数 S 型函数，定义如下：

$$y_j(S_j^y) = 1/[1 + \exp(-S_j^y)] \quad . \tag{14}$$

再将 y 层节点的每个计算结果结合其 y_j 上的权重 w_{ij} 的线性组合 S_k^z 传递给 z 层的激活函数 $z_k(S_k^z)$，S_k^z 定义如下：

$$S_k^z = \sum_{j=1}^{m} w_{jk} y_j \quad . \tag{15}$$

其中，RPROP 中成本函数定义：

$$E = \sum (z_k - d_k)^2 + 10^{-\alpha} \sum w_{ij}^2 \quad . \tag{16}$$

从已知的脉冲星里形成 259 个输入向量的训练集，时长按照 SNNS 应该在 120 次循环前结束，否则将过度训练。输出向量 z 由 z_1 和 z_2 两个元素组成，当 z_1 >0.5 和 z_2 <0.5 时，候选体被选出，否则排除掉。该方法使用 ANN 对 PMPS[8] 数据重复分析，取 250 万个脉冲星的候选体 (包含 501 颗脉冲星，其中 51 颗属于训练集)。在 8:8:2 的 ANN 中，对 250 万个脉冲星候选体的分析总共花了 50 个 CPU 时间 (CPU 全速工作时完成该进程所花费的时间)，任务分配计算机集群中并行化处理，结果显示符合 z_1 >0.5 和 z_2 <0.5 的占 0.5%，对应约 13 万候选体，并且 92% 的脉冲星已经被找到，即每 30 个候选体中就有一个是已知脉冲星；而在标准的图形分析中，每 4 900 个候选体中才可能有 1 个真正的脉冲星。在 12:12:2 的 ANN 中添加 4 个分数，会提高到 93%，但计算时间将增加一倍。该扩展方法对 MSPs 训练不佳，并且由于训练集等原因使一部分遗漏。

Bates 等人 (2012) 应用 ANN 分类候选体[19]，在高时间分辨率巡天项目 (High Time Resolution Universe pulsar survey，简称 HTRU)[36] 数据处理中发现了 75 颗脉冲星。Bates 使用的方法借鉴了 Keith 等人 (2009)[27] 和 Eatough 等人 (2010)[18] 的研究，使用脉冲周期、DM、检测的信噪比、脉冲宽度和来自 sin 函数拟合脉冲轮廓的 χ^2 值等 22 个候选体参数，在 ANN 中使用对数 S 型激活函数 (如式 (14) 所示)，成本函数定义如下：

$$E = \sum_k E_k = \sum_k \frac{1}{2} \sum (z_k - d_k)^2 \quad . \tag{17}$$

其中，ANN 的训练和生成使用 SNNS[35]，训练集为 70 颗脉冲星候选体，验证集为 200 个非脉冲星候选体，采用 22:22:2 结构的神经网络体系 (输入和隐藏层均 22 个，输出层 2 个)，输出层格式为 “X Y”，其中 X,Y = [01]，ANN 输出筛除 $X < 0.5$ 和 $Y > 0.5$ 的候选体。随着 ANN 训练，当验证误差 (公式 (17)) 达到最小值点时最佳。对 LTO-4 的数据检测显示，ANN 能够筛除约 99.7% 的候选体。

HTRU 两年的巡天数据获得了 580 颗已知脉冲星的候选体，剔除掉用于训练集的 70 颗，剩余的 510 颗使用 ANN 能够检测 85% 。Bates 构建的 ANN 能够检测所有周期信号的脉冲星, 但明显不擅长识别较宽的脉冲 (即占空比大于等于 20%)，以及信噪比较低的和毫秒周期的脉冲星，在脉冲周期大于 100 ms 时能检测 86.2%，低于 100 ms 时检测 71% 。ANN 检测中纬度数据集, 约 15 颗正常脉冲星没有检测到。

Zhu 等人 (2014) 提出了深度神经网络图像模式识别——PICS (Pulsar Image-based Classification System)[20]，忽略候选体的周期、DM 和噪声等特征，采用图像模式识别的方法辨别候选体图像是否是真实脉冲星，该方法运用于 PALFA[12, 37, 38]数据，发现 6 颗新的脉冲星。PICS 通过 ML 的方法使计算机得到脉冲星图像模式，训练集采用的脉冲星具有弱、宽、多峰等脉冲形状特点，保证 PICS 能够正确识别有特殊脉冲形状的脉冲星。

PICS 处理过程可分为两步：(1) 分析候选体图像中 4 个子图 (见图 1)，其中一维曲线 (脉冲轮廓，DM 曲线) 采用 ANN 和支持向量机 (Support Vector Machine，简称 SVM) 结合，二维图像 (时间 - 相位图像，频率 - 相位图像) 采用卷积神经网络和 SVM 结合，为每个子图打出一个从 0 (非脉冲星) 到 1 (脉冲星) 的评分，得到 8 个评分；(2) 采用卷积神经网络学习评分器对第一步中得到的 8 个评分采用逻辑回归算法，分配各自的权重 w_i，根据逻辑方程将加权求和值转化为一个概率值。逻辑方程表示为：

$$P = \frac{1}{1 + \mathrm{e}^{-\sum_i w_i x_i}} \quad . \tag{18}$$

处理时间相位和频率相位图像的卷积神经网络分五层，共有 8 820 个神经元。第一层系统将图像下采样为 48×48 像素矩阵，并用 20 个不同的 16×16 图像核进行卷积，产生 20 个 33×33 的特征映射。第二层是最大池层，用 3×3 的块取块中每个特征映射的最大值形成 11×11 大小矩阵；并在第三层中与 50 个 8×8 的图像核取卷积，得到 50 个 4×4 的特征映射。第四层是另一个最大池层，它用 2×2 的块再将特征映射转换成 50×4 个数字来表征原图像中相应块的特征。最后一层是全连接的神经网络层，通过 500 个隐式变量来计算这 200 个数的最终评分。

在 PICS 的检验过程中，检验数据中 10% 的脉冲星和 94% 的脉冲星谐频排在了前 1%，排除了 99%，大大降低了工作量。PICS 在测试阶段完成了复杂的计算，尽管网络较为复杂，但是大多数评价器只是进行简单的点积运算，而候选体评分的确有着较高的效率。在测试过程中，PICS 用一个 2.7 GHz 主频 24 核计算节点在约 45 min 内完成了 90 008 颗脉冲星候选体的评分。

Morello 等人 (2014) 提出了新的脉冲星候选体识别方法——SPINN(Straightforward Pulsar Identification using Neural Networks)[21]。通过使用 SPINN 处理脉冲星候选体，已在澳大利亚 Parkes 射电望远镜上开展的南天中银纬高时间分辨率巡天项目 (HTRU)[36] 中发现了 4 颗新脉冲星 (其中 3 颗为 MSPs，$P < 5$ ms)。SPINN 使用监督式学习方式的 ANN 算法，构造二元分类器，激活函数是式 (14) 所示的 S 型非线性函数，成本函数选择类似式 (16)。为了超过先前自动化候选体筛选工具的性能，特别是期望正确标记脉冲星的成功率能达到 100%，SPINN 使用了一个更大的训练数据集，其中包括 1 196 颗脉冲星中 542 颗显著的脉冲

星，写了一个自定义 ANN (网络由 8 个隐藏单元，1 个输出单元构成) 用于增加控制训练过程，设计了新特征去描述一个脉冲星候选体的性质。

SPINN 在分析脉冲星搜寻软件生成的候选体文件时，综合脉冲星信号的多方面特征，主要采用 6 个评价因子：信噪比、占空比、lg(脉冲周期/DM)、与最小 DM 的偏离值、时域脉冲持续强度以及积分区间脉冲轮廓与平均脉冲轮廓偏差的均方根，对每个候选体进行评分。

为了增强对暗弱、低占空比脉冲星信号鉴别的能力，SPINN 不会直接选取信噪比的评价较高的信号，而是首先评价除了信噪比之外的其它 4 个因子。当出现此 4 个因子相似，而信噪比不同的信号时，高信噪比与低信噪比的信号做平等处理，避免遗漏较暗弱、占空比较小的脉冲星信号，特别是 MSPs。通过与 PSRCAT① 中已知脉冲星数据对比，SPINN 对候选体的处理中大概能找到 95% 的已知脉冲星。SPINN 在低分处仍存在不能有效识别的脉冲星，需要进一步人工进行确认。

此外，为适应大型射电望远镜观测数据的快速增长，以及观测数据特征不断变化的需求，Lyon 等人 (2015) 提出一种自适应的在线分类算法[22]。它使用高斯 - 黑林格快速决策树 (Gaussian Hellinger Very Fast Decision Tree，简称 GH-VFDT)[39] 筛选有可能的候选体，为了避免“维数灾难”相关问题导致分类的性能降低，采用了 8 个新的特征 (4 个为统计折叠后的脉冲轮廓得到，另外 4 个类似上述方法，从 DM-SNR 曲线中得到) 去表述典型脉冲星候选体，使用 HTRU 1[21]、HTRU 2[40] 和 LOTAAS 1[41, 42] 这 3 个独立的数据集来测试识别的性能。GH-VFDT 使用单个 2.2 GHz Quad Core mobile CPU (Intel Core i7-2720QM 处理器) 可以每秒识别约 7 万个候选体，以及高水准的脉冲星识别率 (90% 以上)。这种方法已经使用在 LOTAAS 中，在观测获得的数据中发现了 20 颗新的脉冲星。

4 相关讨论

脉冲星按周期分为毫秒脉冲星[43–45]、长周期脉冲星，现有观测的周期范围在 1.39 ms ∼ 11.79 s 之间[2]；按类别分为单脉冲星、脉冲双星和双脉冲星等；按脉冲流量分为暗弱脉冲星和强脉冲星等。由于地面干扰信号、天空背景噪声和系统噪声的影响会直接导致对应的脉冲星候选体四个子图异常复杂，并且还不能准确构建候选体中的统计量与真实脉冲星之间的内在关系，导致一些特征不易提取。上述几种 AI 方法，在一定程度上成功提取了具有典型的脉冲星特征，运用机器学习的方法发现人工筛选很难识别的脉冲星，且提高了筛选速度，对 FAST 即将开展的候选体筛选工作具有一定的借鉴意义，FAST 可根据这些脉冲星的特点 (周期、DM、脉冲轮廓、脉冲峰值和图像特征等) 进行特征提取，并开展机器学习方法。现从测试数据、准确度和构建方法等几个方面，总结已有的脉冲星候选体筛选方法，如表 2 所示。

① http://www.atnf.csiro.au/research/pulsar/psrcat/download.html

表 2　6 种 AI 方法筛选脉冲星候选体的总结

方法类型	处理的巡天数据	找到的新脉冲星数	准确度	构建方法	参考文献
ANN	PMPS	1	92% 93%	8:8:2 ANN 12:12:2 ANN	[18]
ANN	HTRU Medlat	75	85%	22:22:2 ANN	[19]
PEACE	PALFA GBNCC HTRU North	47	0.17%(68%) 0.34%(95%) 3.7%(100%)	6 个评价因子	[11]
PICS	PALFA GBNCC GBNCC GBNCC	6	1%(92%) 3.8%(100%) 1.1%(100%) 0.16%(68%)	ANN+SVM+ 卷积神经网路	[20]
SPINN	HTRU Medlat	4	0.11%(95%) 0.01%(95%)	6 个评价因子	[21]
GH-VFDT	LOTAAS	20	$>90\%$	8 个新的特征	[22]

5　探讨与展望

Parkes 64 m 射电望远镜的 PKSMB[8, 27, 29, 30, 46–49] 巡天是目前为止最成功的一次脉冲星巡天，探测到 1 122 颗脉冲星，发现新脉冲星 834 颗[6]，其中，Manchester 等人 (2001) 在 parkes 脉冲星巡天中使用 13 波束接收机，巡天数据产生了 800 万个脉冲星候选体，发现了 600 多颗脉冲星[8]。500 m 口径 FAST 望远镜将采用 19 波束接收机进行脉冲星巡天，依据 parkes 观测结果推算，FAST 巡天数据将产生千万甚至上亿脉冲星候选体。另外，世界最大综合孔径射电望远镜平方公里阵 (Square Kilometre Array，简称 SKA) 已进入建设准备阶段，一旦建成，其相当于 100 面 100 m 口径天线组成的射电望远镜阵列，灵敏度比 FAST 提高一个量级[50]，需要处理的脉冲星候选体将会达到数十亿甚至更多。应对 FAST 和 SKA 即将产生的脉冲星候选体，人工智能是唯一能对脉冲星候选体进行实时、准确筛选的方案。

第 3 章介绍的 AI 方法对长周期脉冲星以及 MSPs 均有效。然而，宇宙中不排除存在一些奇异类型的脉冲星，对这些可能存在类型的候选体特征，利用 AI 也将无能为力，未来 FAST 可尝试构建各种脉冲星理论模型，模拟脉冲信号特征，生成模拟数据，用于 AI 训练集，搜索特定类型的候选体。其次，改进 FAST 脉冲星搜索过程，构建搜索数据库，在多个候选体参数空间上设置阈值，快速挑选出比较好的结果。另外，利用 FAST 数据构造 ML 的训练集时，可分为窄脉冲、宽脉冲以及潜在的双星系统等。最后，可将 FAST 产生的候选体进行分类，如 MSPs、长周期脉冲星、噪声信号，甚至是常见的 RFI 信号，此四类作为训练 AI 算法的数据集，在数据处理中可以分别获得我们想要的部分巡天数据，以及剔除掉不想要的部分，既快速检测出真实脉冲星的候选体，又减少噪声和常见 RFI 的候选体数量。

近几年 AI 快速发展，各种新技术和新成果不断涌现，如神经网络之后的又一突破——深度学习，其框架将特征和分类器结合到一个框架中，用数据去学习特征，在使用中减少了手

工设计特征的巨大工作量[51]。因此，FAST 可考虑采用深度学习的方式去筛选候选体。

参考文献：

[1] Hewish A, Bell S J, Pilkington J D H, et al. Naure, 1968, 217: 709
[2] ATNF Pulsar Catalogue, http://www.atnf.csiro.au/people/pulsar/psrcat/
[3] 南仁东. 中国科学, 2005, 35: 449
[4] 南仁东. 天文学报, 2016, 57(6): 623
[5] Nan R D, et al. IJMPD, 2011, 20: 989
[6] Zhang L, Wang P, Li D, et al. Progress In Astronomy, 2015, 33: 1000
[7] Zhang L, Hobbs G, Li D, et al. 2016, RAA, 16(10): 11
[8] Manchester R N, Lyne A G, D Amico N, et al. MNRAS, 2001, 279: 1235
[9] Smits R, Lorimer D R, Kramer M, et al. 2009, A&A, 505: 919
[10] http://www.erogol.com/brief-history-machine-learning/
[11] Lee K J, Stovall K, Jenet F A, et al. MNRAS, 2013, 433: 688
[12] Cordes J M, Frelre P C C, Lorimer D R, et al. ApJ, 2006, 637: 446
[13] Lorimer D R, Stairs I H, Freire P C, et al. ApJ, 2006, 640: 428
[14] Hessels J W T, Nice D J, Gaensler B M, et al. ApJ 2008, 682: L41
[15] Nice D J, Altiere E, Bogdanov S, et al. ApJ, 2013, 772: 50
[16] Stovall K, Lynch R S, Ransom S, et al. ApJ, 2014, 791: 67
[17] Barr E. AIPC, 2011, 1357: 52
[18] Eatough R P, Molkenthin N, Kramer M, et al. MNRAS, 2010, 407: 2443
[19] Bates S D, Bailes M, Barsdell B R, et al. MNRAS, 2012, 427: 1052
[20] Zhu W W, Berndsen A, Madsen E C, et al. ApJ, 2014, 781: 2
[21] Morello V, Barr E D, Bailes M, et al. MNRAS, 2014, 443: 1651
[22] Lyon R J, Stapper B W, Cooper S, et al. MNRAS, 2016, 459: 1104
[23] Lyne A G, Graham-Smith F. Pulsar astronomy, 3rd ed., Cambridge: Cambridge University Press, 2006
[24] Lorimer D R, Kramer M. Handbook of pulsar astronomy, Cambridge: Cambridge University Press, 2005
[25] Miller A, Rodriguez-Zermeno A, Jenet F. AAS, 2006, 38: 993
[26] Faulkner A J, Stairs I H, Kramer M, et al. MNRAS, 2004, 355: 147
[27] Keith M J, Eatough R P, Lyne A G, et al. MNRAS, 2009, 395: 837
[28] Manchester R N, Lyne A G, Camilo F, et al. MNRAS, 2001, 328: 17
[29] Morris D J, Hobbs G, Lyne A G, et al. MNRAS, 2002, 335: 275
[30] Hobbs G, Faulkner A, Stairs I H, et al. MNRAS, 2004, 352: 1439
[31] Edwards R T, Bailes M, van Straten W, et al. MNRAS, 2001, 326: 358
[32] Burgay M, Rea N, Israel G L, et al. MNRAS, 2006, 372: 410
[33] Rodriguez M, Stovall K, Banaszak S A, et al. AAS, 2015, 225: 346
[34] Deneva J S, Cordes, J M, Mc Laughlin M A, et al. Ap, 2009, 703: 2259
[35] http://www.ra.cs.uni-tuebingen.de/SNNS/, 2017
[36] Keith M J, Jameson A, van Straten W, et al. MNRAS, 2010, 409: 619
[37] Kaspi V M. BAAS, 2012, 219: 237
[38] Lazarus P. IAU Symp, 2013, 291: 35
[39] Haykin S. Neural Networks and Learning Machines, 3rd edition, Upper Saddle River: Prentice Hall, 2009
[40] Thornton D. PhD thesis, Univ. Manchester, 2003
[41] LOFAR Pulsar Working Group. 2013, presentation at LOFAR Status Meeting, Dwingeloo, The Netherlands. https://www.astron.nl/lofarwiki/lib/exe/fetch.php?media=public:lsm%20new:2013%2003%2006%20hesself.pdf
[42] Cooper S. 2014, presentation at LOFAR Science, Amsterdam, The Netherlands. http://www.astron.nl/lofarscience2014/Documents/Tuesday/Session%20III/Cooper.pdf

[43] Bailes M, Lorimer D. ASP Conference Series, 1995, 72: 17
[44] Vasisht G, Gotthelf E V. ApJ, 1997, 486: 129
[45] Hessels J W T, Ransom S M, Stairs I H, et al. Science, 2006, 311: 1901
[46] Kramer M, Bell J F, Manchester R N, et al. MNRAS, 2003, 342: 1299
[47] Lorimer D R, Faulkner A J, Lyne A G, et al. MNRAS, 2006, 372: 777
[48] Mickaliger M B, Lorimer D R, Boyles J, et al. ApJ, 2012, 759: 127
[49] Knispel B, Eatough R P, Kim H, et al. ApJ, 2013, 774: 93
[50] Peng B, Jin C J, Du B, et al. Sci Sin-Phys Mech Astron, 2012, 42: 129
[51] David S, Aja H, Chris J M, et al. Nature, 2016, 529(7587): 484

Application of Artificial Intelligence in the Selection of Pulsar Candidate

XU Yu-yun[1], LI Di[2,3], LIU Zhi-jie[1], WANG Chen[4], WANG Pei[2], ZHANG lei[2], PAN Zhi-chen[2]

(1. Guizhou Normal University, Key Laboratory of Information and Computing Science Guizhou Province, Guizhou 550001, China; 2. National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100012, China; 3. Key Laboratory of Radio Astronomy, Chinese Academy of Sciences, Nanjing 2100008, China; 4. Commonwealth Scientific and Industrial Research Organisation, Canberra ACT 2601, Australia)

Abstract: Pulsar searching, aiming at discovering pulsars and transient sources such as Rotation RAdio Transients and Fast Radio Bursts(FRB), is the first step of sciences based on pulsars and transient sources. As more pulsars are discovered in pulsar searches, some special kinds of pulsars, including those having very short spin periods and very eccentric orbits, will also be illuminated. These special pulsars provide extreme physical conditions for us to examine our current understands, which is imperative to the research of the dense matter state equation, interplanetary navigation, the interstellar medium, and the measurement of gravitational waves.

Nowadays, one typical pulsar search/survey in radio frequencies (e.g. 1.4 GHz) will bring us millions of candidates to confirm. To check the quality of so many candidates, manual operations are too slow and inefficient. Machine learning and artificial intelligence have greatly developed in recent years, helping us identify high quality pulsar candidates from pulsar search results. This paper will discuss the use of aritificial intelligence in pulsar candidates checking, statistics and analysis in pulsar candidates identification, and the efficiency expectations of using similar methods in identifying candidates from Five-hundred-meter Aperture Spherical radio Telescope.

Key words: Artificial Intelligence; pulsar; candidate selection