脉冲星候选样本分类方法综述

王元超[1,2]，郑建华[1,2]，潘之辰[3,4,5]，李明涛[1,2]

（1. 中国科学院 国家空间科学中心，北京 100190；2. 中国科学院大学，北京 100049；3. 中国科学院 国家天文台，北京 100012；
4. 中国科学院 天文大数据中心，北京 100012；5. 中国科学院FAST重点实验室，北京 100012）

摘 要： 脉冲星搜索是射电天文学中的重要前沿领域。随着现代搜索设备性能的不断提升，可以接收到更弱的信号，如何从海量信号中准确地识别出脉冲星疑似信号成为一个挑战。介绍了国内外关于脉冲星候选样本分类方法的发展历史和发展状态，归纳总结了发展过程中各个阶段的处理方法：人工识别方法和机器学习方法；最后对未来的发展趋势进行了分析。

关键词： 脉冲星；脉冲星候选样本；机器学习

中图分类号： P16 **文献标识码：** A **文章编号：** 2095-7777(2018)03-0203-09

DOI:10.15982/j.issn.2095-7777.2018.3.001

引用格式： 王元超，郑建华，潘之辰，等. 脉冲星候选样本分类方法综述[J]. 深空探测学报，2018，5（3）：203-211.

Reference format: WANG Y C，ZHENG J H，PAN Z C，et al. An overview of pulsar candidate classification methods[J]. Journal of Deep Space Exploration，2018，5（3）：203-211.

0 引 言

脉冲星是一种有强引力作用、强磁场并快速旋转的中子星，具有稳定的自转周期。脉冲星相关的发现先后两次获得诺贝尔物理学奖（第一颗脉冲星的发现[1]和脉冲星双星系统的首次发现[2]）。对脉冲星的观测研究，极大地推动了天文、天体物理、粒子物理、等离子体物理、广义相对论、引力波和导航等众多领域的发展。例如，脉冲星的射电脉冲在经过星际空间到达地球前，会受到星际介质的影响，产生色散等效应，这为星际介质的研究提供了机会[3]；作为超新星爆发的产物，脉冲星对于研究超新星爆发理论具有重要价值[4]；脉冲双星系统也为广义相对论的检验提供了机会[5]；通过分析毫秒脉冲星计时阵列的脉冲到达时间的变化，可以分析引力波信号[6]等。

自第一颗脉冲星被发现后，大量射电望远镜设备被应用到脉冲星搜索中。目前已发现2 700多颗脉冲星，其中大部分是由脉冲星巡天设备发现。例如，Parkes多波束脉冲巡天（Parkes Multi-beam Pulsar Survey，PMPS）[7]，高时间分辨率的宇宙脉冲星巡天（High Time Resolution Universe Survey，HTRU）[8]，Arecibo L波段馈源阵列脉冲星巡天（Pulsar Arecibo L-band Feed Array Survey，PALFA）[9]，低频射电（Low Frequency Array，LOFAR）阵列巡天（LOFAR Tied-Array All-sky Survey，LOTAAS）[10]，绿岸北半球脉冲星巡天（Greenbank Northern Celestial Cap Survey，GBNCC）[11]等。

随着现代脉冲星搜索设备性能的不断提升，可以接收到更弱的信号，能够探测到更多脉冲星的同时，也产生大量的候选样本，而且大部分样本是射频干扰（Radio Frequency Interferance，RFI）或噪声等。例如，1977年，投入使用的2nd Molonglo survey只接收到约2 500个样本[12]，而新一代射电望远镜500 m口径球面射电望远镜（Five Hundred Meter Aperture Spherical Telescope，FAST）[13]预计可以发现5 000颗脉冲星；平方千米阵列（Square Kilometer Array，SKA）[14]预计可以发现2万颗脉冲星。SKA按照保守估计（以HTRU数据的样本比例：脉冲星/非脉冲星为1/10 000[32]为参考计算）需要处理20亿样本。

因而如何有效地从海量数据中筛选出有价值的脉冲星疑似样本，以便进一步观测确认成为需要解决的一个重要问题。本文将阐述脉冲星候选样本分类方法的发展历史、发展现状和技术发展趋势。

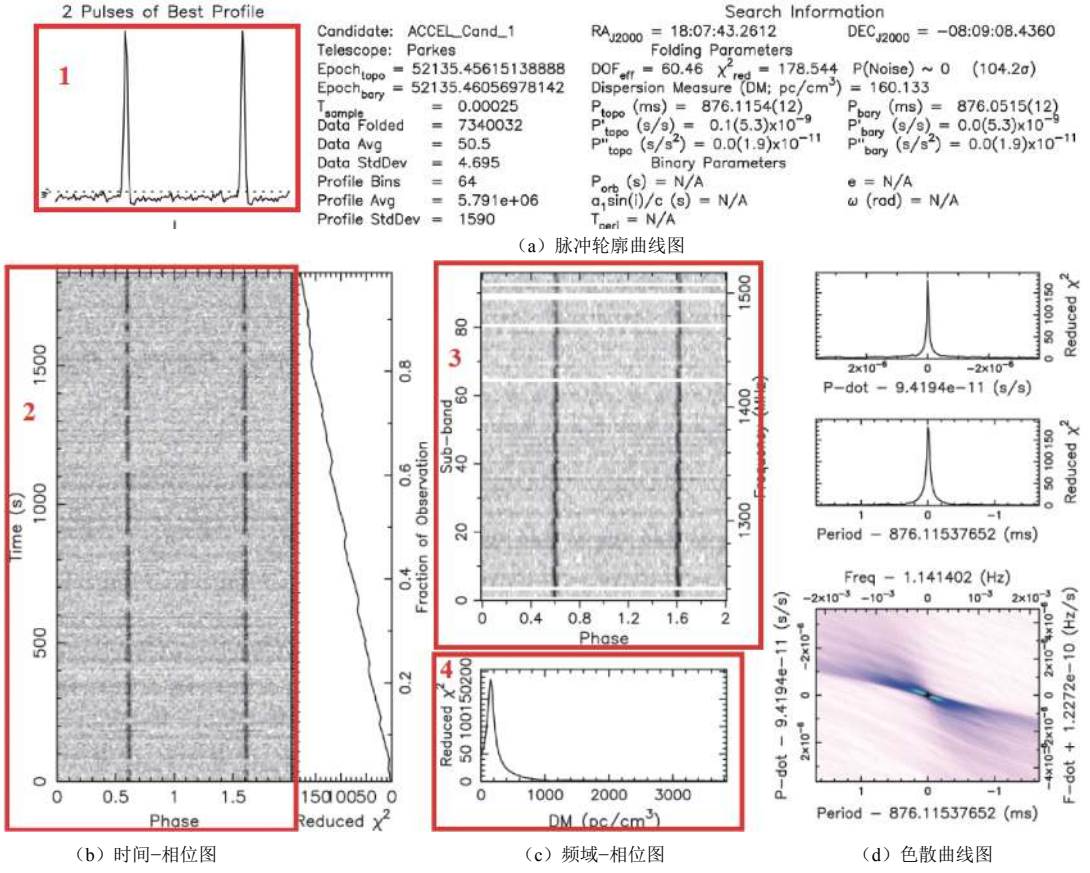
1 脉冲星候选样本

目前，脉冲星信号搜索主要借助大型射电望远镜。大部分的脉冲星信号很微弱，且信号在传播中会受到星际介质的影响，因而设备接收到周期性信号

收稿日期：2018-04-15 修回日期：2018-06-05

基金项目：中国科学院青年创新促进会优秀会员资助项目

后，会借助搜索软件（如PRESTO[15]等）进行一系列的数据处理。例如，通过剪波（clipping）处理，减少脉冲干扰[16]；进行消色散（de-dispersion）处理，消除色散延迟[17]；再借助傅立叶变换，将数据转换到频域进行分析，从而确定信号周期[18]；根据确定的信号周期，将接收到的多个周期的信号进行叠加，放大信号的信噪比，得到平均脉冲轮廓[19]。经过处理后的数据，会转换为图像形式，作为脉冲星候选样本。图 1是PRESTO处理后的一个脉冲星候选样本的图像示例（图像来自PMPS[20]）。



（a）脉冲轮廓曲线图

（b）时间–相位图　（c）频域–相位图　（d）色散曲线图

图 1　脉冲星样本图像示例，使用PRESTO软件处理得到

Fig. 1　An example figure of a pulsar candidate in PMPS，which was converted by PRESTO

标注的子图（a）～（d）依次为脉冲轮廓曲线图、时间–相位图、频域–相位图和色散曲线图。这些候选样本会被进一步分类筛选，以便选择有价值的脉冲星疑似信号进行观测确认，这个过程被称为脉冲星候选样本的分类。分类的目标是在尽可能不遗漏脉冲星信号的情况下，减少非脉冲星信号的保留（减少进一步观测的工作量）。

天文学家在判断候选样本是否是脉冲星疑似信号时，主要参考以下特征：

1）脉冲轮廓曲线图：通过折叠累加所有频域和时域信号强度得到。由于脉冲星具有稳定的自转周期，理想的脉冲星信号数据在每个周期内会形成一个或多个明显的波峰。

2）时间–相位图：通过累加信号在不同频域的数据得到，反映的是信号在观测时间内的强度。脉冲星信号具有周期性，信号会在整个观测时间内不断重复出现。在时间–相位图上，信号强度越大，颜色越深。从而对理想的脉冲星信号，在整个观测时间内，会形成与脉冲轮廓曲线图波峰位置相对应的竖直线。

3）频域–相位图：通过累计信号在观测时间内的数据得到，反映的是信号在不同频率下的强度。由于脉冲星射电辐射是宽频的，典型的脉冲信号应当出现在观测的大部分频率段上。若为脉冲星信号，对应到频域–相位图上，应当在大部分频率内，有与波峰相对应的竖直线。

4）色散曲线图：脉冲信号在经过星际介质时，会产生色散。色散曲线图反映的是使用不同色散值进行消色散时，脉冲曲线信噪比的变化情况。当使用正确的值消色散时，脉冲信噪比将最大。因而若为脉冲星信

号，曲线会在非零位置有一个峰值，曲线呈“钟形”。

在设计算法进行自动分类时，特征的设计一般也是围绕着这几点进行刻画（参见表 2、3、5～9）。但对现代设备的候选样本实现脉冲星疑似信号的有效分类，存在以下难点：①候选样本数量大；②脉冲星样本与非脉冲星样本之间、不同类型脉冲星之间的样本不均衡；③干扰信号种类多，部分RFI形似脉冲星信号；④部分脉冲星信号较弱，特征不明显，易遗漏；⑤不同设备、不同区域的RFI环境等不同，使得算法间的可移植性较差。

2　国内外研究进展

关于脉冲星候选样本的有效分类，国内外许多学者进行了大量工作。目前的方法，大致可分为人工识别方法和机器学习方法。其中，人工识别方法可分为基于信噪比信息分类方法、图像软件辅助方法、打分排序方法等；同时，将对应的机器学习方法根据特征类型分为3类：基于经验特征的方法、基于统计特征的方法和基于数据驱动的方法。表 1是对这些方法的简单比较。

表 1　典型脉冲星候选样本分类方法对比

Table 1　The comparison of pulsar candidate classification methods

类型	典型方法	文献（年份）	测试数据	测试结果	
				查全率/%	假正率/%
人工识别方法	S/N等	/	/	/	/
	REAPER	Faulkner等（2004）	/	/	/
	JPEAPER	Keith等（2009）	/	/	/
	PEACE	Lee等（2013）	GBNCC	95	0.34
	ANN	Eatough等（2010）	PMPS	93	1
				92	0.5
	ANN	Bates等（2012）	HTRU	85	1
	ANN（SPINN）	Morello等（2014）	HTRU 1	100	0.64
				99	0.11
				95	0.01
机器学习方法	集成CNN + ANN + SVM + LR（PICS）	Zhu等（2014）	PALFA	92	1
			GBNCC	100	3.8
			GBNCC	68	0.16
			GBNCC	100	1.1
	GH-VFDT	Lyon等（2016）	HTRU 1	92.8	0.5
			HTRU 2	82.9	0.8
			LOTAAS 1	78.9	0.1
	fuzzy knn	Mohamed（2017）	HTRU 2	94.2	1.8
	DCGAN-SVM	Guo等（2017）	HTRU 1	96.6	0.05
			HTRU 1	96.3	0.05
			PMPS-26K	89.5	0.5
			PMPS-26K	89.1	0.5
	集成学习	Tan等（2018）	LOTAAS 2	98.7	1.1

2.1　人工识别方法

脉冲星搜索的前期，由于设备性能等原因，接收到的数据样本有限，研究人员可以借助以往的经验，直接对接收的信号进行人工筛选。

2.1.1　基于信噪比信息分类方法

在早期的识别中信噪比作为重要的判别特征使用。一些简单的筛选软件（例如MSP Find[21]）应用到相关的搜索设备上，只接受一定信噪比范围内的信号，辅助减少样本数量。比如，在Arecibo Phase II survey上，Stokes等通过只保留信噪比大于8σ的信号，得到5 000多个候选样本再进行人工进一步识别处理[22]。在Parkes 20 cm survey上，Johnston等使用同样的策略，筛选出约15万个候选样本[23]。人工初步筛选处理

速度慢，存在较大的主观性。同时，仅根据信噪比等信息筛选，分类误差较大，会遗漏信号较弱的脉冲星。后续研究人员加入更多信息（比如周期等），在一定程度上提升了准确度，但效果有限。

2.1.2 图像软件辅助方法

利用信噪比、脉冲周期等数据信息进行分类，直观性不强，不利于分析判断，处理速度较慢。因而基于统计特征的图像分类软件被开发用于辅助脉冲星疑似信号的分类操作。例如，2004年，Faulkner等设计了图像分类软件REAPER[24]。它可以根据基本特征（周期、脉冲宽度等），直观地把不同样本展示在二维图像中，将明显的噪声信号与脉冲星疑似信号区分开，减少候选样本的数量。借助REAPER，在对PMPS数据进行再次处理中，新发现了128颗脉冲星。2009年，Keith等对REAPER进行了改进，设计了JREAPER软件[25]。在JPEAPER的帮助下，在PMPS数据中又发现了之前被错分遗漏的28颗脉冲星。

另外，也出现了一些基于网络的图像样本查看评分系统。比如，Pulsar Search Collaboratory[26]，通过培训后的高中生，对类似图 1所示的样本图像的多个特征进行在线的评分，从而进行样本的分类[27]。该项目开始于2008年，目前已发现了7颗新脉冲星[28]。

基于统计特征的图像分类软件可以有效地筛除一部分明显的干扰信号，减少进一步人工观测的工作量，提升分类速度。但使用基于一定的经验和假设，依赖于研究人员的认知水平与经验模式，手动调整，存在很强的主观性。

2.1.3 打分排序方法

为实现更智能的分类，研究人员尝试对样本进行打分排序。Keith等在JREAPER软件[25]中，设置了经验式的评分标准，对样本进行排序，筛除低分的候选样本（见表2）。2013年，Lee等通过分析大量的脉冲星数据，设计了PEACE系统[29]，通过6个特征（见表 3）来刻画脉冲星信号，利用函数分别进行评分，并将分数线性组合，根据最终的评分进行排序。在GBNCC数据测试集上，实现了查全率95%时，假正率为0.34%，并从PALFA、GBNCC和HTRU数据集中发现了47颗脉冲星。PEACE提升了分类识别的效率和准确度，但需要人工预先设定评分函数并调节，对人类经验依赖程度很高，只是“半自动化”的分类方法。

随着样本数量的不断增加，人工识别的方法越来越无法满足脉冲星候选样本分类的需求。因而如何在算法中减少主观性，实现自动化，进一步提升准确度和处理速度，成为需解决的一个问题。

表 2 Keith等（2009）使用的特征[28]

Table 2 The features used in Keith et al.（2009）[28]

编号	特征描述
1	脉冲周期
2	脉冲轮廓的信噪比
3	色散值
4	高于阈值的脉冲占比

表 3 Lee等（2013）使用的特征[29]

Table 3 The features used in Lee et al.（2013）[29]

编号	特征描述
1	脉冲轮廓的信噪比
2	脉冲周期
3	脉冲轮廓的宽度
4	信号在时域的持续度
5	信号在频域的持续度
6	脉冲宽度与色散值的比值

2.2 机器学习方法

为应对数据量不断增大的挑战，随着机器学习的发展，相关的算法也被引入脉冲星候选样本分类任务中。由于样本极度不均衡并且研究人员更关注脉冲星的分类准确度，因而一般使用查全率（Recall）、查准率（Precision）、假正率（False Positive Rate，FPR）来反映算法的性能。其中Recall刻画的是正样本（脉冲星信号）被正确分类的比例；Precision反映的是分类器认定为正类的样本中实际正样本的比例；FPR计算的是负样本（非脉冲星信号）中被分类器错认为正类的比例。Recall越高，脉冲星样本被正确分类的越多；Precision越高或FPR越低，非脉冲星信号被错分的越少。

需要指出的是，由于缺少公共数据集，且大部分算法是针对不同的脉冲星搜索设备的数据进行的设计，因而多数算法是采用各不相同的数据集进行的性能测试。由于样本数量、样本分布、样本比例、样本质量等因素的不同，算法间不能直接定量比较。为方便对照，将部分数据集样本数量信息汇总于表 4。

根据分类特征的类型，将目前的关于脉冲星分类的机器学习方法，大致分为：基于经验特征的方法、基于统计特征的方法和基于数据驱动的方法。

2.2.1 基于经验特征的方法

基于经验特征的方法，参照人工分类时的判别方式，引入启发式特征，实现自动评分分类。例如使用信噪比特征、使用sin函数/高斯函数对脉冲曲线进行拟合等。

2010年，Eatough等对启发式评分方法进行了改

进，引入机器学习方法，不再进行人工评分[30]。基于射电天文学的专业知识，挑选了信噪比、脉冲宽度等12个特征（见表5）作为三层人工神经网络（Artificial Neural Network，ANN）的输入、输出对应的评分（见表5）。在1.3万个PMPS数据测试集上，实现了93%的查全率，1%的假正率（只使用前8个特征时，查全率为92%，假正率为0.5%）。在对部分PMPS数据进行再处理时，从中发现一颗新的脉冲星。作者对测试数据分析发现：由于毫秒脉冲星与普通脉冲星的不同，以及训练样本的不均衡等原因，使得约50%的脉冲周期小于10 ms的脉冲星被错分；60%的信噪比超过400的脉冲星被错分。

2012年，Bates等[31]将特征增加到22个（表 6），借助人工神经网络，在HTRU测试集上实现了85%的查全率、1%的假正率。从部分HTRU Medlat数据中发现了75颗脉冲星。相比于Eatough等[30]的处理结果，在脉冲周期小于10 ms的脉冲星和长周期的脉冲星分类性能上得到了一定的提升，但也增加了模型的复杂度。

表 4　部分数据集样本数量比较

Table 4　The comparison of some datasets

数据集	文献（年份）	样本数量	脉冲星	非脉冲星	备注
PMPS	Eatough等（2010）[30]	～15 385	760	～14 625	训练集259 + 1 625
HTRU	Bates等（2012）[31]	>780	580	>200	训练集70 + 200
GBNCC	Lee等（2013）[29]	～100 000	70	～100 000	不需训练集
PALFA	Zhu等（2014）[36]	3 756	1 659	2 097	训练:测试 = 3:2
GBNCC	Zhu等（2014）[36]	90 008	277	89 731	全部用于测试
HTRU 1	Morello（2014）[32]	91 191	1 196	89 995	训练:测试 = 4:1
	Lyon（2016）[33]				训练集200 + 200
	Guo等（2017）[37]				训练:验证:测试 = 3:3:4
HTRU 2	Lyon等（2016）[33]	17 898	1 639	16 259	训练集200 + 200
	Mohamed（2017）[34]				训练:测试 = 4:1
LOTAAS 1	Lyon等（2016）[33]	5 053	66	4 987	训练集33 + 200
PMPS-26K	Guo等（2017）[37]	26 000	2 000	24 000	训练:验证:测试 = 3:3:4
LOTAAS 2	Tan等（2018）[35]	2 403	986	1 417	训练集同时作为测试集

表 5　Eatough等（2010）使用的特征[30]

Table 5　The features used in Eatough et al.（2010）[30]

编号	特征描述
1	信噪比
2	脉冲轮廓宽度
3	对DM-S/N曲线拟合的卡方值
4	S/N > 10的DM值的数量
5	对优化后的DM-S/N曲线拟合的卡方值
6	对acceleration-S/N曲线拟合的卡方值
7	S/N > 10的加速度值的数量
8	对优化后的acceleration-S/N曲线拟合的卡方值
9	子频域最大值的均方根离散度
10	频域间的线性相关度
11	子时域最大值的均方根离散度
12	时域间的线性相关度

2014年，Morello等对人工神经网络方法进行了进一步的优化，设计了SPINN（Straightforward Pulsar Identification using Neutral Networks）分类器[32]。选取了6个特征（表 7）作为人工神经网络的输入。在91 192个不均衡样本构成的HTRU Medlat测试集（简称HTRU 1）上，调节阈值参数，可以在达到100%查全率时，假正率为0.64%；99%查全率时，假正率为0.11%；95%查全率时，假正率为0.01%。并对434万个样本再次处理后，筛选出2 400个疑似样本，经过进一步观测确认，发现4颗新的脉冲星。需要指出的是，SPINN“100%查全率时，假正率为0.64%”是根据所有的脉冲星得分中最低分作为分类阈值时，推算得出的。在挑选特征时，考虑了对弱信号的兼顾、对噪声干扰的稳定性以及减少特征间相关度，降低模型的复杂度的同时提升了算法的性能。但对一些形似脉冲星信号的RFI，SPINN并不能很好地分类。因而建议，对RFI的特征进行更好地刻画；同时增加脉冲星数据，降低不均衡度。

基于经验特征的人工神经网络方法的应用极大地提高了脉冲星候选样本分类的准确度和处理速度。对于特征的选取，Eatough等[30]、Bates等[31]、Morello等[32]

学者进行了不断的优化。但他们是基于一定的经验和假设，特征对数据集依赖性较强[32]，同时，根据人工处理的思路设计的特征有可能使得算法“模仿”人工处理的错误[33]。例如，反复出现的信噪比，会使得算法倾向于信噪比高的“强”信号，而更多的较弱的信号会被忽略。为进一步提高性能，研究人员考虑使用不同的机器学习方法和不同的特征选取策略。

表 6 Bates等（2012）使用的特征[31]

Table 6 The features used in Bates et al.（2012）[31]

编号	特征描述
1	周期
2	色散值
3	信噪比
4	脉冲宽度
5	用Sin曲线拟合脉冲轮廓的卡方值
6	用$\sin^2$曲线拟合脉冲轮廓的卡方值
7	高斯拟合脉冲轮廓的卡方值
8	高斯拟合的半高宽
9	双高斯拟合脉冲轮廓的卡方值
10	双高斯拟合脉冲轮廓的平均半高宽
11	轮廓直方图对0的偏移量
12	轮廓直方图最大值/高斯拟合的最大值
13	d（profile)/dx直方图与轮廓直方图的偏移量
14	$S/N_{\mathrm{data}}/\sqrt{(P-W)/W}$
15	$S/N_{\mathrm{fit}}/\sqrt{(P-W)/W}$
16	mod $(DM_{\mathrm{fit}}-DM_{\mathrm{best}})$
17	DM曲线拟合的卡方值
18	峰值处对应的所有频段值的均方根
19	任意两个频段线性相关度的均值
20	线性相关度的和
21	脉冲轮廓的波峰数
22	脉冲轮廓减去均值后的面积

表 7 Morell等（2014）使用的特征[32]

Table 7 The features used in Morello et al.（2014）[32]

编号	特征描述
1	信噪比（log值）
2	脉冲周期
3	周期和色散值的比值（log值）
4	色散值
5	信号在时域的连续度
6	脉冲轮廓与子时域的均方根

2.2.2 基于统计特征的方法

2016年，Lyon等针对SKA实时处理样本的需求，同时为避免特征对数据集的依赖性和倾向性，设计了新的特征和算法[33]。从脉冲轮廓曲线和DM曲线中提取均值、方差、峰度、偏度共8个无偏向性的统计特征（表 8），具有较好的区分度；考虑到实时接收数据时可能存在的数据样本不均衡、不同区域观测可能产生的样本分布漂移等问题，设计了针对不均衡数据流的Gaussian Hellinger快速决策树算法（Gaussian Hellinger Very Fast Decision Tree，GH-VFDT），实现在线处理不均衡的数据。GH-VFDT处理速度快，每秒可以处理7万张样本（单个2.2 GHz，Intel i7-2720QM处理器），但也牺牲了一定的分类准确度。在HTRU1、HTRU2、LOTAAS1数据集上测试时，对应的查全率和假正率依次是：92.8%（0.5%）、82.9%（0.8%）、78.9%（0.1%）。

表 8 Lyon等（2016）使用的特征[33]

Table 8 The features used in Lyon et al.（2016）[33]

编号	特征
1	脉冲轮廓的均值
2	脉冲轮廓的标准差
3	脉冲轮廓的超额峰度
4	脉冲轮廓的偏度
5	DM-S/N曲线的均值
6	DM-S/N曲线的标准差
7	DM-S/N曲线的超额峰度
8	DM-S/N曲线的偏度

另外，Lyon等新设计的8个特征具有较好的区分性，被之后一些研究人员所借鉴使用。2017年，Mohamed将这些特征（表 8）应用到模糊k近邻分类器（Fuzzy K Nearest Neighbors，Fuzzy KNN）算法上[34]，在HTRU2数据集上测试提升了一定的查全率，实现了94.2%的查全率、1.8%的假正率，进一步证明了统计特征的有效性。

针对Lyon等[33]使用的特征缺少时间–相位图和频域–相位图信息，并在实际分类处理中对宽脉冲脉冲星容易错分的情况，2018年，Tan等[35]进行了改进，通过计算时间–相位图或频域–相位图与脉冲轮廓曲线的相关系数，增加了对应的8个统计特征（表 9）。同时将形似脉冲星信号的RFI单独分类，由2分类（脉冲星、非脉冲星）变为3分类（脉冲星、噪声、RFI）。通过利用不同波束探测到的脉冲星信号数据，构造多个决策树，集成提升性能。算法可以较好地识别宽脉冲的脉冲星，在新的LOTAAS数据测试集（为与之前的数据集区别，代称LOTAAS 2）上，相比较于Lyon等[34]算法的结果，查全率提升2.5%，为98.7%；假正率FPR则从

2.5%降到了1.1%。该算法被应用于LOTAAS搜索系统中。需要指出的是，由于样本数量有限，在作者的测试实验中测试集包含训练集，因而实际性能可能会稍低一些。

表 9　Tan等（2018）新添加的特征[35]

Table 9　The features added in Tan et al.（2018）[35]

编号	特征
1	Sub-bands与脉冲轮廓的相关系数的均值
2	Sub-bands与脉冲轮廓的相关系数的标准差
3	Sub-bands与脉冲轮廓的相关系数的超额峰度
4	Sub-bands与脉冲轮廓的相关系数的偏度的绝对值
5	Sub-ints与脉冲轮廓的相关系数的均值
6	Sub-ints与脉冲轮廓的相关系数的标准差
7	Sub-ints与脉冲轮廓的相关系数的超额峰度
8	Sub-ints与脉冲轮廓的相关系数的偏度的绝对值

2.2.3　基于数据驱动特征的方法

卷积神经网络可以实现自动提取特征，避免人工设计特征的缺陷，因而基于卷积神经网络的数据驱动的机器学习方法在脉冲星疑似样本的分类领域具有重要价值。

2014年，Zhu等从图像模式识别的角度分析，设计了PICS（Pulsar Image-based Classification System）系统[36]。PICS将支持向量机（Support Vector Machines，SVM）、ANN、卷积神经网络（Convolutional Neural Networks，CNN）、逻辑回归（Logistic Regression，LR）等集成结合。直接使用PRESTO软件输出的4幅子图（图 1已标出）作为SVM、ANN或CNN的输入，并将多个分类器的输出使用LR进行整合，最终输出对应的评分。PICS完全由数据驱动，避免了人工设计特征可能存在的不足（倾向性、数据集的依赖性等），在PALFA数据集上，实现92%的查全率、1%的假正率。并且基于PALFA数据训练的模型，直接在90 008个样本的GBNCC数据上测试，可达到100%的查全率、3.8%的假正率。PICS系统提升了识别的准确度，并具有一定的泛化性能，但模型也相对更复杂。

2017年，Guo等提供了一个新的思路，利用深度卷积对抗生成网络（Deep Convolution Generative Adversarial Network，DCGAN）来进行特征的提取[37]。DCGAN是一种生成模型，将时间–相位图或频率–相位图分别作为输入，利用DCGAN来生成更多的样本；同时DCGAN自动学习对应的特征，作为SVM分类器的输入。该算法在解决样本不均衡问题时，也避免了人工特征的设计提取。在HTRU1 数据集上，使用时间–相位图作为输入时，查全率为96.6%、精度为96.1%，假正率约为0.05%；

使用频率–相位图作为输入时，查全率为96.3%、精度为96.5%、假正率约为0.05%。在PMPS-26K数据集上，使用时间–相位图作为输入效果更好，查全率为89.5%、精度为88.5%、假正率约0.5%。但是DCGAN模型复杂且训练不稳定；仅考虑了二维子图，会缺失部分特征信息，影响算法性能，对复杂情况的分类效果有待进一步提升。表 10中对这3类机器学习算法进行了简单的优缺点总结。

表 10　3类机器学习算法的优缺点总结

Table 10　The advantages and disadvantages of three kinds of machine learning methods

类型	主要代表	优点	缺点
基于经验特征的方法	Eatough等[30]、Bates等[31]、Morello等[32]	算法简单，复杂度低	基于一定的经验和假设，特征对数据集依赖性较强
基于统计特征的方法	Lyon等[33]、Tan等[35]	特征无偏，计算简单，复杂度较低，性能较高	需要人工设计特征
基于数据驱动的方法	Zhu等[36]、Guo等[37]	数据驱动，避免了人工设计特征，准确度高	需要大量数据，模型复杂度较高，训练困难

为实现更好的机器学习算法的效果，针对样本不均衡问题，许多学者做了一些有益的尝试。Morello等[32]在对人工神经网络方法进行优化时，使用过采样方法，使得训练集正负样本比例为4:1。Lyon等测试了Hoeffding树分类器处理不平衡数据流的效果[38]，并进一步设计了针对不均衡数据流的GH-VFDT算法[39]。2016年，Yao等在目标函数中设置不同的权重，通过集成支持向量机算法提升分类效果[40]。2017年，Ford利用Lyon等提出的特征[33]，分别测试了支持向量机、神经网络、决策树、Bagging集成等算法，在上采样、下采样、ROSE采样、SMOTE采样下的性能，实验表明性能均有提升[41]。2017年，Guo等[37]使用DCGAN来进行特征提取的时候，同时生成高质量的新数据，从而缓解样本不均衡对算法的影响。

3　未来的发展趋势

1）在分类器算法设计方面，传统的机器学习算法目前依然占据主流；在特征设计提取上，已经由传统

的手工设计特征发展到自动抽取特征。深度学习在特征提取学习方面具有优势，PICS和DCGAN-SVM已经做出了深度学习在脉冲星样本分类上的有益尝试。深度学习、对抗生成网络等技术将会发挥更大的作用。

2）目前，数据多为二分类（脉冲星、非脉冲星），或简单的三分类（脉冲星、RFI、噪声）。而Tan等[35]也提到，不同类型的脉冲星之间、不同类型的RFI之间也会有很多的差异。因而在数据量允许的情况下，结合数据分布的特点，进行更合适的、更细致的样本分类，可以使得特征提取更加有效，进一步提升分类算法的性能。

3）大部分机器学习算法都是作为线下处理使用，针对在线数据处理的较少。Lyon[42]提到，随着设备性能的不断提升，数据量将会持续增加，需要设计更好的数据管理工具、文件格式、数据标准，同时需要更好地实时在线自动处理数据。因而针对数据流设计在线不均衡数据的处理算法，具有重要的价值。

4）目前的算法均为根据已知的脉冲星特征来构建有监督学习，对数据集有很强的依赖性。如果出现未知的新现象、新样本，将会被当成干扰而筛除。因而考虑半监督学习或无监督学习，对离群点进行深入分析，有助于充分利用数据。

5）不同设备间数据分布等存在差异，使得需要分别设计或调整算法。因而提升算法的泛化能力，使之适应不同数据，具有重要的意义。

4 结束语

本文从脉冲星识别的意义、历史发展及现状、未来趋势等角度阐述了脉冲星疑似信号分类识别的问题。基于机器学习算法设计有效的分类器将有助于脉冲星候选样本的准确识别分类，促进脉冲星天文学的发展。

参 考 文 献

[1] HEWISH A，BELL S J，PILKINGTON J D H，et al. Observation of a rapidly pulsating radio source[J]. Nature，1968，217(5130)：709-713.

[2] HULSE R A，TAYLOR J H. Discovery of a pulsar in a binary system[J]. Annals of the New York Academy of Sciences，1975，262(1)：490-492.

[3] CORDES J M，KRAMER M，LAZIO T J W，et al. Pulsars as tools for fundamental physics and astrophysics[J]. New Astronomy Reviews，2005，48(11)：1413-1438.

[4] LORIMER D R，LYNE A G，CAMILO F. A search for pulsars in supernova remnants[J]. Astronomy and Astrophysics，1998，331(3)：1002-1010.

[5] LYNE A G，BURGAY M，KRAMER M，et al. A double-pulsar system：a rare laboratory for relativistic gravity and plasma physics[J]. Science，2004，303(5661)：1153-1157.

[6] HOBBS G B，BAILES M，BHAT N D R，et al. Gravitational-wave detection using pulsars：status of the parkes pulsar timing array project[J]. Publications of the Astronomical Society of Australia，2009，26(2)：103-109.

[7] MANCHESTER R N，LYNE A G，CAMILO F，et al. The parkes multibeam pulsar survey - i. observing and data analysis systems，discovery and timing of 100 pulsar[J]. Monthly Notices of the Royal Astronomical Society，2001，328(1)：17-35.

[8] KEITH M J，JAMESON A，STRATEN VAN W，et al. The high time resolution universe pulsar survey - i. system configuration and initial discoveries[J]. Monthly Notices of the Royal Astronomical Society，2010，409(2)：619-627.

[9] DENEVA J，CORDES J M，MCLAUGHLIN M A，et al. Arecibo pulsar survey using alfa. iii. probing radio pulsar intermittency and transients[J]. Astrophysical Journal，2008，703(2)：2259-2274.

[10] COENEN T，LEEUWEN J VAN，HESSELS J W T，et al. The LOFAR pilot surveys for pulsars and fast radio transients[J]. Astronomy & Astrophysics，2014，570(1)：1-18.

[11] STOVALL K，LYNCH R S，RANSOM S M，et al. The green bank northern celestial cap pulsar survey - i：survey description，data analysis，and initial results[J]. Astrophysical Journal，2014，791(791)：2837-2854.

[12] MANCHESTER R N，LYNE A G，TAYLOR J H，et al. The second Molonglo pulsar survey – discovery of 155 pulsars[J]. Monthly Notices of the Royal Astronomical Society，1978，185(2)：409-421.

[13] NAN R，LI D，JIN C J，et al. The five-hundred-meter aperture spherical radio telescope project[J]. International Journal of Modern Physics D，2011，20(06)：989-1024.

[14] SMITS R，KRAMER M，STAPPERS B，et al. Pulsar searches and timing with the square kilometre array[J]. Astronomy & Astrophysics，2009，493(3)：1161-1170.

[15] RANSOM S. Presto[EB/OL]. [2018-7-5]. https://www.cv.nrao.edu/~sransom/presto/.

[16] HOGDEN J，VANDER WIEL S，BOWER G C，et al. Comparison of radio-frequency interference mitigation strategies for dispersed pulse detection[J]. Astrophysical Journal，2012，747(2)：141.

[17] LORIMER D R. Radio pulsar statistics[J]. Astrophysics and Space Science Library，2009(357)：161-172.

[18] LORIMER D R，KRAMER M. Handbook of pulsar astronomy[M]. Longdon：Cambridge University Press，2005.

[19] LYON R J. Why are pulsars hard to find?[D]. Manchester：University of Manchester，2016.

[20] ATNF. Parkes multibeam pulsar survey[EB/OL]. [2018-7-6]. http://www.atnf.csiro.au/people/pulsar/pmsurv/.

[21] JOHNSTON S，LYNE A G，MANCHESTER R N，et al. A high-frequency survey of the southern galactic plane for pulsars[J]. Monthly Notices of the Royal Astronomical Society，1992，255(3)：401-411.

[22] STOKES G H，SEGELSTEIN D J，TAYLOR J H，et al. Results of two surveys for fast pulsars[J]. Astrophysical Journal，1986，311(311)：694-700.

[23] JOHNSTON S，LYNE A G，MANCHESTER R N，et al. A high-frequency survey of the southern galactic plane for pulsars[J]. Monthly Notices of the Royal Astronomical Society，1992，255(3)：401-411.

[24] FAULKNER A J，STAIRS H I，KRAMER M，et al. The parkes multibeam pulsar survey：v. finding binary and millisecond pulsars[J]. Monthly Notices of the Royal Astronomical Society，2004，355(1)：147-158.

[25] KEITH M，EATOUGH R，LYNE A，et al. Discovery of 28 pulsars using new techniques for sorting pulsar candidates[J]. Monthly Notices of the Royal Astronomical Society，2009，395(2)：837-846.

[26] ROSEN R，HEATHERLY S A，MAURA A，et al. Pulsar search collaboratory[EB/OL]. (2010-5-6)[2018-7-5]. http://pulsarsearch collaboratory.com/.

[27] ROSEN R，HEATHERLY S，MCLAUGHLIN M A，et al. The pulsar search collaboratory[J]. Astronomy Education Review，2010，9(1)：010106.

[28] WILLIAMSON K，MCLAUGHLIN M，HEATHERLY S A，et al. The pulsar search collaboratory：expanding nationwide[J]. Instrumentation and Methods for Astrophysics，2018，arXiv：1807.06059.

[29] LEE K J，STOVALL K，JENET F A，et al. Peace：pulsar evaluation algorithm for candidate extraction – a software package for post-analysis processing of pulsar survey candidates[J]. Monthly Notices of the Royal Astronomical Society，2013，433(1)：688-694.

[30] EATOUGH R P，MOLKENTHIN N，KRAMER M，et al. Selection of radio pulsar candidates using artificial neural networks[J]. Monthly Notices of the Royal Astronomical Society，2010，407(4)：2443-2450.

[31] BATES S D，BAILES M，BARSDELL B R，et al. The high time resolution universe pulsar survey VI：An artificial neural network and timing of 75 pulsars[J]. Monthly Notices of the Royal Astronomical Society，2012，427(2)：1052-1065.

[32] MORELLO V，BARR E D，BAILES M，et al. SPINN：a straightforward machine learning solution to the pulsar candidate selection problem[J]. Monthly Notices of the Royal Astronomical Society，2014，443(2)：1651-1662.

[33] LYON R J，STAPPERS B W，COOPER S，et al. Fifty years of pulsar candidate selection：from simple filters to a new principled real-time classification approach[J]. Monthly Notices of the Royal Astronomical Society，2016，459(1)：1104-1123.

[34] MOHAMED T M. Pulsar selection using fuzzy KNN classifier[J]. Future Computing and Informatics Journal，2017，3(1)：1-6.

[35] TAN C M，LYON R J，STAPPERS B W，et al. Ensemble candidate classification for the LOTAAS pulsar survey[J]. Monthly Notices of the Royal Astronomical Society，2018，474(4)：4571-4583.

[36] ZHU W W，BERNDSEN A，MADSEN E C，et al. Searching for pulsars using image pattern recognition[J]. The Astrophysical Journal，2014，781(2)：117-128.

[37] GUO P，DUAN F. WANG P，et al. Pulsar candidate identification with artificial intelligence techniques[J]. 2017，arXiv preprint arXiv：1711.10339v1.

[38] LYON R J，BROOKE J M，KNOWLES J D，et al. A study on classification in imbalanced and partially-labelled data streams[C]//2013 IEEE International Conference on Systems，Man，and Cybernetics. [S.l]：IEEE，2013：1506-1511.

[39] LYON R J，BROOKE J M，KNOWLES J D，et al. Hellinger distance trees for imbalanced streams[C]//ICPR'14 Proceedings of the 2014 22nd International Conference on Pattern Recognition. Washington，DC，USA：IEEE Computer Society，2014：1969-1974.

[40] YAO Y，XIN X，GUO P，et al. Pulsar candidate selection by assembling positive sample emphasized classifiers[C]//12th International Conference on Computational Intelligence and Security. [S.l]：IEEE，2016：120-124.

[41] FORD J M. Pulsar search using supervised machine learning[D]. Florida：Nova Southeastern University，2017.

[42] LYON R J. Fifty years of candidate pulsar selection - what next？[J]. Proceedings of the International Astronomical Union，2017：337-340.

作者简介：

王元超(1991–)，男，博士研究生，主要研究方向：机器学习。

通信地址：北京市海淀区中关村南二条一号(100190)

电话：(010)62586383

E-mail：wangyuancaho15@mails.ucas.edu.cn

郑建华(1966-)，女，研究员，博士生导师，主要研究方向：飞行器动力学控制与仿真、飞行器自主导航与控制。

通信地址：北京市怀柔区京密路北二街国家空间科学中心(100091)

电话：(010)62586374

E-mail：zhengjianhua@nssc.ac.cn

潘之辰(1988–)，男，助理研究员，主要研究方向：脉冲星搜索。

通信地址：北京市朝阳区大屯路甲20号(100012)

E-mail：panzc@bao.ac.cn

李明涛(1982–)，男，研究员，主要研究方向：航天动力学与控制、空间任务分析与设计。

通信地址：北京市海淀区中关村南二条一号(100190)

电话：(010)61611037

E-mail：limingtao@nssc.ac.cn

(下转第218页)

Celestial Doppler Difference/Pulsar for Formation Flying and Its Integrated Navigation

YU Ziyuan[1]，LIU Jin[1]，NING Xiaolin[2]，MA Xin[2]，GUI Mingzhen[2]，KANG Zhiwei[3]

（1. College of Information Science and Technology，Wuhan University of Science and Technology，Wuhan 430081，China；
2. School of Instrumentation Science &Opto-Electronics Engineering，Beihang University，Beijing 100191，China；
3. College of Computer Science and Electronic Engineering，Hunan University，Changsha 410082，China）

Abstract：In order to improve the spacecraft capability of autonomous celestial navigation，a celestial Doppler difference/pulsar for formation flying and its integrated navigation method is proposed. The Sun light is strong，and the accuracy of the Sun Doppler difference navigation is high，but it is difficult to provide multi-directional velocity information. Star light is weak，and the accuracy of star Doppler difference navigation is low，but it can provide multi-directional velocity information. The Sun Doppler difference navigation and the star Doppler difference navigation are complementary，but which cannot be fully observable. Using three or more pulsar navigation is completely observable，but the filtering period is longer，and it is difficult to obtain continuous navigation information. The three navigation methods are complementary and can be used for integrated navigation. The extended Kalman filter is used as a navigation filter to fuse the difference and arrival time of the astronomical Doppler，and can provide absolute and relative navigation information for formation flying. Simulation results show that the integrated navigation method for formation flight can provide absolute and relative highly-accurate navigation information.

Key words：formation flight；celestial Doppler difference navigation；Kalman filter；pulsar

High lights：

- In this paper, the Doppler difference method can be used to solve Solar spectral line drift problem caused by the Sun's surface activity.
- Combining the star Doppler difference navigation method with the Sun Doppler difference navigation can provide multi-directional navigation information.
- The celestial Doppler difference navigation system is not fully observable. In this paper, we combine it with the pulsar navigation method to provide highly-accurate navigation information for formation flight.

[责任编辑：高莎，英文审校：任树芳]

(上接第211页)

An Overview of Pulsar Candidate Classification Methods

WANG Yuanchao[1,2]，ZHENG Jianhua[1,2]，PAN Zhichen[3,4,5]，LI Mingtao[1,2]

（1. National Space Science Center，Chinese Academy of Sciences，Beijing 100190；2. University of Chinese Academy of Sciences，Beijing 100049；3. National Astronomical Observatories，Chinese Academy of Sciences，Beijing 100012；4. Center for Astronomical Mega-Science，Chinese Academy of Sciences，Beijing 100012；5. CAS Key Laboratory of FAST，NAOC，Chinese Academy of Sciences，Beijing 100012）

Abstract：Pulsar searching is an important frontier in radio astronomy. Weaker signals can be received as the performance of search facilities continualy improved. However，how to accurately identify the suspected pulsar signal from massive candidates has become a challenge. The pulsar candidate classification methods about development history and current situation at home and abroad. The classification methods in each stage include：manual selection methods and machine learning methods. At last，the future development trends are analyzed.

Key words：pulsar；pulsar candidate；machine learning

High lights：

- The development history of pulsar candidate classification methods is introduced.
- The pulsar candidate classification methods based on manual selection and machine learning are summarized and compared.
- The future trends of pulsar candidate classification are analyzed.

[责任编辑：杨晓燕，英文审校：朱恬]