doi: 10.15940/j.cnki.0001-5245.2023.05.006

面向单脉冲信号分类的集成特征选择与评价*

张金区[1†] 凌 毓[1] 杜 平[2] 李乡儒[1‡] 李 慧[1]

(1 华南师范大学计算机学院 广州 510631)
(2 广东建设职业技术学院建筑信息学院 清远 511500)

摘要 受大量射频干扰信号影响, 快速从海量观测数据中准确识别出单脉冲信号已成为天文数据处理的一项重要任务, 而设计和提取有效数据特征, 是利用机器学习进行单脉冲信号高效识别的决定因素. 针对如何选择最优特征, 进而提升单脉冲信号的分类精度这一关键问题, 设计了面向单脉冲信号分类的集成特征选择方法. 方法首先混合单脉冲信号的参数特征、统计特征和抽象特征, 然后分别利用5种单一特征选择方法选出各自的最优特征集, 最后利用贪心策略对5种单一方法获取的最优特征集进行集成筛选, 获取最优集成特征集. 实验表明, 最优特征集合既包含统计特征也包含抽象特征. 在相同特征数量下, 利用集成特征选择比单一特征选择能获得更高的模型精度, 可使F1值最高提升1.8%. 在海量数据背景下, 集成特征选择对减少特征数量、提升分类性能和加快数据处理速度具有重要作用.

关键词 脉冲信号, 射电脉冲星, 作用变量, 方法: 数据分析

中图分类号: P161; **文献标识码**: A

1 引言

单脉冲信号是指由宇宙天体发出的没有固定周期的脉冲辐射信号, 主要分为自转型暂现射电源(Rotating Radio Transients, RRATs)和快速射电暴(Fast Radio Bursts, FRBs)两类[1–3]. 随着科技的不断发展和天文观测设备灵敏度的不断提升, 观测接收的脉冲信号中夹杂着越来越多的干扰信号, 受飞机、雷达、电离层等影响的干扰信号呈指数增长, 如何从海量观测数据中准确识别出属于天体的单脉冲信号已成为天文数据处理的一项重要任务. 为此, 国内外学者进行了大量的研究工作. 目前, 基于机器学习的方法已经成为单脉冲信号挖掘的主要方法, 而如何设计和提取脉冲信号特征是影响机器学习性能的关键因素[4]. 通过筛选有效特征, 不但能够去除冗余特征, 在一定程度上降低了数据处理的计算量, 而且能够提升识别准确度. 这在高速大规模巡天背景下, 有助于提升单脉冲信号搜索的效率.

根据特征的来源和计算方式, 脉冲信号的特征主要分为3类, 分别是参数特征、统计特征和抽象特征. 参数特征是指在接收脉冲信号时由信号接收器、空间环境和数据处理管线等决定的一些特征. 例如, 色散(Dispersion Measure, DM)是宇宙天体和地球之间沿信号传播方向上的自由电子积分柱密度, 单位为$\mathrm{pc\cdot cm^{-3}}$, 它由空间环境决定, 但是对脉冲信号的分类识别有重要影响, 是典型的参数特征. 同样, 信噪比(S/N)是射电天文望远镜接收到信号的电压与同时记录的噪声电压的比值. 信噪比越

2022-08-03收到原稿, 2022-11-10收到修改稿

*国家自然科学基金项目(11973022、61273248、61075033)和校一般科研项目(KY2021-36)资助

†zjq@scnu.edu.cn

‡lixiangru@m.scnu.edu.cn

高, 即信号强度相对噪声更大, 信噪比也是识别脉冲信号的主要依据. 参数特征通常是在接收天体信号并做初步处理的时候直接记录在数据文档中, 后续可以直接读取或者通过简单计算获得, 其特点是特征获取简单、含义明确, 对脉冲信号分类效果影响明显.

统计特征是指通过对数据进行观察计算后, 人工设计出的一些具有描述意义的量化特征. 例如, Lyon等人基于脉冲轮廓曲线和DM-S/N曲线分别计算了4个无偏统计特征, 分别为曲线的均值、标准差、超额峰度与偏度, 这些特征在单脉冲信号分类中具有较好的性能[5]. Tan等人在Lyon等[5]无偏统计特征的基础上, 新增加了基于时间-相位图、频率-相位图和脉冲轮廓图的相关统计特征, 在分类时, 极大地降低了假阳率这一评价指标[6]. 统计特征的特点是含义明确, 但是其设计受经验影响大, 并且容易遗漏掉重要的统计特征.

抽象特征是指那些不需要人工设计, 直接由算法自动提取的特征. 目前, 基于卷积神经网络的卷积运算是最常用的抽象特征提取方法. 它利用不同的卷积核, 经过多层卷积运算, 最终输出一系列特征, 这些特征没有明确含义, 但对模型分类具有良好的效果, 正成为各领域应用的主流, 在单脉冲信号识别方面也发挥着越来越重要的作用. 例如, Zhu等人设计了一个基于图像的脉冲星分类系统PICS (Pulsar Image based Classification System), 该系统通过PRESTO (PulsaR Exploration and Search Toolkit)软件输出的4幅子图进行脉冲星信号的筛选, 并使用卷积神经网络从脉冲星候选体中自动学习脉冲星的特征, 再利用支持向量机、人工神经网络(Artificial Neural networks, ANN)、逻辑回归等分类算法进行脉冲星信号的分类[7]. Wang等人根据PICS (the pulsar image-based classification system)系统提出了PICS-ResNet (Residual Networks)模型, 主要思路是使用ResNet替换了原来的CNN (Convolutional Neural Networks), 通过在FAST (the Five-hundred-meter Aperture Spherical radio Telescope)与GBNCC (Green Bank North Celestial Cap)等观测数据上进行实验, 获得了更高的分类性能[8]. 2020年, Agarwal等人基于8种深度网络模型, 如VGG (Visual Geometry Group)和Densenet (Dense Convolutional Network)等网络结构提取的特征, 组建了11个深度学习模型, 已探测到了超过20颗脉冲星的2000多个单脉冲信号[9–10].

应用表明, 基于卷积神经网络的抽象特征, 可以有效进行脉冲信号的分类识别, 但是其可解释性差, 含义不明确. 另外, 利用卷积神经网络提取的特征, 经常包含冗余特征, 不但消耗计算资源, 而且在一定程度上影响分类结果的准确性. 因此, 如何充分利用参数特征、统计特征和抽象特征各自的优势, 对单脉冲信号的分类具有重要意义. 本文的目标是设计一种集成多元特征的选择和评价方法, 为基于机器学习的单脉冲信号分类提供特征选择的方法和依据.

2 数据来源

在本文中, 直接使用Michilli等[11]工作中已标注的单脉冲数据集进行实验分析. 该数据集来源于低频射电联合阵列巡天(LOFAR tied-array all-sky survey, LOTAAS)项目, 具体形成过程可见参考文献[11–12]. 该数据集包含脉冲信号记录374万条, 归属于53066个脉冲事件, 其中35063个为射频干扰事件, 18003个属于47个已知脉冲星的脉冲事件. 属于同一个脉冲的信号事件组成一个弥散脉冲组.

3 集成特征选择方法设计

集成特征选择的基本思路是从参数特征、统计特征和抽象特征构成的特征集合中选择最适合单脉冲信号分类的最优特征组合. 其总体技术流程如图1所示, 主要分为3步: 第1步是分别计算参数特征、统计特征和抽象特征, 形成多元原始特征集合; 第2步是利用单一特征选择方法分别从混合特征集中提取最优特征子集; 第3步是利用贪心策略从多个最优特征子集中筛选最优集成特征子集.

3.1 多元混合特征集的构建

3.1.1 参数特征和统计特征设计

根据脉冲信号数据的特点, 结合已有研究[11–12]中的特征设计, 本文应用的参数特征及统计特征如表1所示.

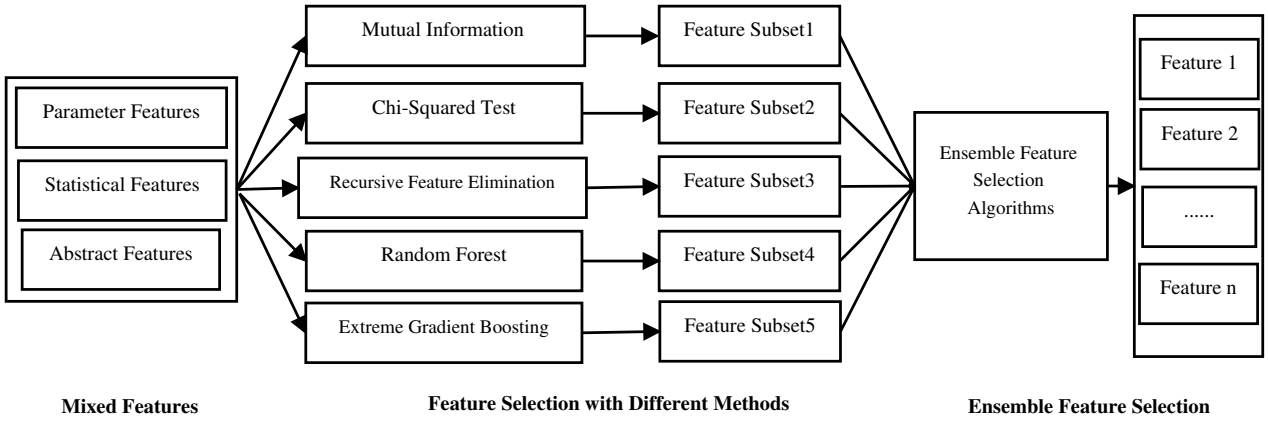

图 1 集成特征选择总体流程

Fig. 1 Overall framework of ensemble feature selection

表 1 参数特征及统计特征

Table 1 Parameter features and statistical features

No.	Features	Meaning
1	DM	Dispersion measure, the integrated column density of free electrons between an observer and a pulsar in unit of $\mathrm{pc}\cdot\mathrm{cm}^{-3}$.
2	S/N	The ratio of signal to noise, that is, the ratio of the voltage value of the signal received by the radio telescope to the noise voltage recorded at the same time.
3	Duration	The window width of the boxcar function used for peak detection of time series signals, that is, the time range of the window.
4	DM_Extent	The DM value extent corresponding to all signal events in a dispersed pulse group.
5	TimeIndex	The index designed according to the generation time of the pulse signal.
6	Time_Extent	The value extent of time corresponding to all signal events in a dispersed pulse group.
7	N_Events	Number of signal events contained in a dispersed pulse group
8	aDM	Average value of DM for all signal events within the same dispersed pulse group, calculated by $\mathrm{aDM}=\frac{\sum_{\mathrm{e}}\mathrm{DM_e}}{\mathrm{N\_Events}}$, where $\mathrm{DM_e}$ is the dispersion measure corresponding to the signal event (e).
9	wDM	Weighted average value of DM for all signal events within the same dispersed pulse group, calculated by $\mathrm{wDM}=\frac{\sum_{\mathrm{e}}\left(\mathrm{DM_e S/N_e}\right)}{\sum_{\mathrm{e}}\mathrm{DM_e}}$.
10	aTime	Average time of all signals forming a dispersed pulse group.
11	KurtSigma	Excess kurtosis of S/N distribution curve, calculated by $k_{\mathrm{S/N}}=\frac{\sum_{\mathrm{e}}\left(\mathrm{DM_e}-\overline{\mathrm{DM}}\right)^4 W_{\mathrm{e}}}{\sigma^4(\mathrm{S/N_e})\sum_{\mathrm{e}}\mathrm{S/N_e}}-3$, where W_{e} is duration, and $\mathrm{S/N_e}$ is S/N corresponding to the signal event (e) respectively. σ is the standard deviation of the S/N of all events in a dispersed pulse group; $\overline{\mathrm{DM}}$ is the mean value of DM.
12	Time	Signal reception time of the strongest event in a dispersed pulse group.

3.1.2 基于卷积神经网络的抽象特征提取

卷积神经网络通过利用卷积、激活、池化等处理, 可以从不同的感受野进行多层特征提取, 在图像分类识别等领域取得了成功的应用. 利用卷积神经网络对单脉冲信号的数据分布图进行抽象特征提取, 将大大增强单脉冲信号的特征来源. 本文搭建深度残差收缩网络, 并将每个弥散脉冲组数据的信噪比与窗口宽度分布曲线形态图像作为网络模型的输入, 依此提取单脉冲信号的抽象特征.

本文设计的深度残差收缩网络(Residual Shrinkage Distribution curve Feature extraction Network, RSDFNet)以He等人提出的深度残差神经网络为基础[13], 在卷积神经网络的基础上引入了残差模块. 其模型结构如图2所示, 以RSDFNet最后一层隐藏层作为特征提取层, 获取从信噪比与窗口宽度分布曲线形态图像中学习的抽象特征.

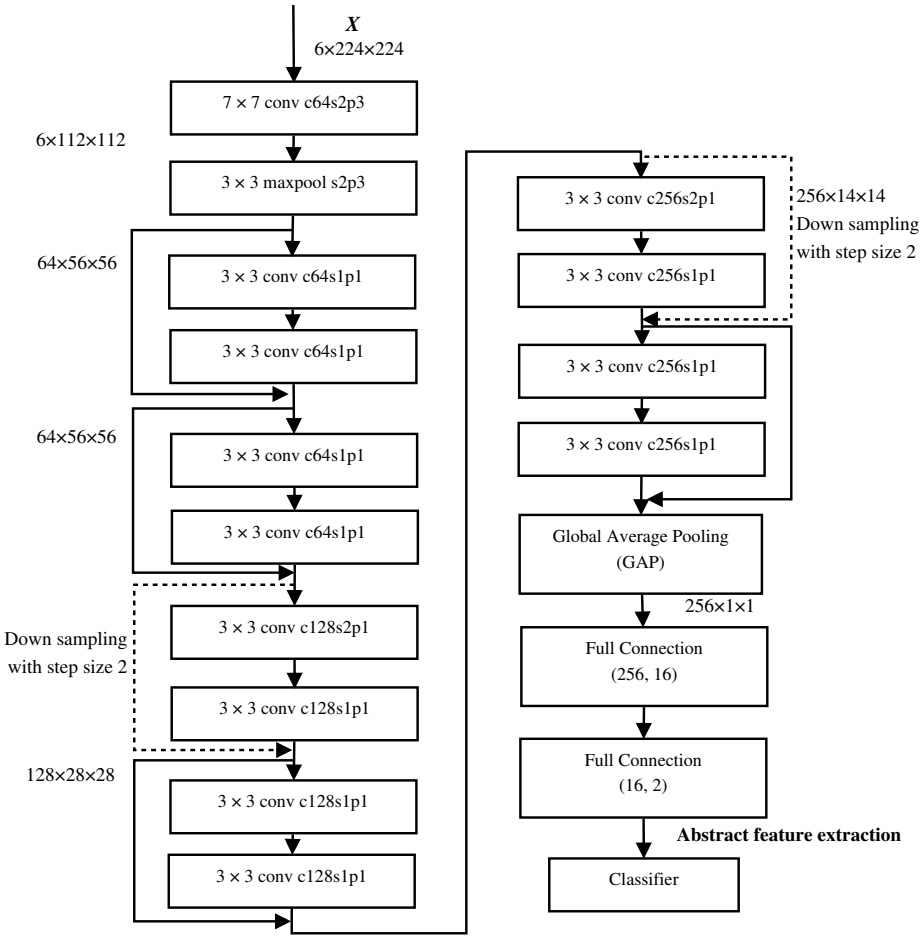

图 2 RSDFNet结构示意图. 图中, $\boldsymbol{X}$为6×224×244大小的输入特征矩阵, conv表示执行卷积操作, c表示通道, s表示步长, p表示池化窗口, c64s2p3即表示对64个通道数据执行步长为2的池化, 池化窗口为3×3. maxpool表示最大值池化, GAP为全局平均池化, 图中两条虚线表示的残差运算, 因通道数不同, 需要降采样处理使通道数一致.

Fig. 2 Structure of RSDFNet. In the figure, $\boldsymbol{X}$ stands for input feature matrix with size of 6×224×244, conv represents convolution operation, c represents channel, s represents step size, and p represents pooling window. For example, c64s2p3 represents performing pooling with step size 2 on 64 channels data, and pooling window is 3×3. Maxpool represents maximum pooling, GAP represents global average pooling, and dashed lines represent different number of channels during residual operation. Down sampling is required to ensure a consistent number of channels.

3.2 基于单一方法的特征子集构建

多元混合特征集中不可避免地存在着众多冗余特征和无效特征, 这些冗余特征不但会降低模型的运算效率, 造成维数灾难, 而且会影响模型的准确性. 因此, 如何筛选出最有用的特征, 对模型计算有重要意义, 然而如何评价一个特征对分类任务的重要性, 却有众多不同的方法. 本文首先利用卡方检验[14]、互信息[15]、递归特征消除[16]、嵌入式特征选择等方法进行单一方法特征选择, 分别筛选出每种方法的最优特征子集. 然后, 将多种方法的最优特征子集进行筛选集成, 形成最优集成特征组合, 以实现各种特征选择方法的优势互补.

3.2.1 基于卡方检验的特征子集选择

卡方检验的基本思想是通过观察实际值与理论值的偏差来确定理论值正确与否. 具体做法是先假设两个变量是独立的("原假设"), 然后观察实际值(观察值)与理论值的偏差程度, 如果偏差足够小则认为两者确实是相互独立的, 此时就接受原假设; 如果偏差大到一定程度, 则认为两者是相关的, 即否定原假设而接受备择假设. 在进行单脉冲信号特征选择的时候, 使用"提取的特征与待识别的单脉冲信号不相关"来做原假设, 计算出的卡方值越大, 说明对原假设的偏离越大, 此时, 倾向认为原假设的反面是正确的, 也就是卡方值越大, 特征与单脉冲信号的相关度越高. 卡方计算公式如(1)式所示:

$$\chi^2 = \sum \frac{(A-E)^2}{E}, \tag{1}$$

其中, A为基于某项特征计算的实际值, E为理论值.

3.2.2 基于互信息的特征子集选择

互信息(Mutual Information)可以用来度量两个随机特征变量之间的相互依赖程度[15], 通常用于评价一个事件的出现对另一个事件出现所贡献的信息量. 在分类中, 可看作是某个特征对于某个类别区分的贡献度. 当变量$\boldsymbol{X}$与$\boldsymbol{Y}$为离散随机特征变量时, 计算公式如下:

$$I(\boldsymbol{X},\boldsymbol{Y}) = \sum_{x\in\boldsymbol{X}}\sum_{y\in\boldsymbol{Y}} p(x,y)\lg\left[\frac{p(x,y)}{p(x)p(y)}\right], \tag{2}$$

在公式中, $I(\boldsymbol{X},\boldsymbol{Y})$表示$\boldsymbol{X}$和$\boldsymbol{Y}$之间的互信息量, $p(x,y)$为$\boldsymbol{X}$和$\boldsymbol{Y}$的联合概率分布函数, $p(x)$和$p(y)$分别为$\boldsymbol{X}$和$\boldsymbol{Y}$的边缘概率分布函数, x表示变量集$\boldsymbol{X}$中的一个成员, y表示变量集$\boldsymbol{Y}$中的一员. 若互信息值为零, 则表明两个随机变量之间互相不提供任何信息, 相互独立. 互信息值越大, 则表明这两个变量之间的依赖程度越高.

3.2.3 基于递归特征消除的特征子集选择

递归特征消除法是指在给定的特征集上训练一个模型, 根据模型的结果从特征集中移除最不重要的特征, 接着在剩余特征集上继续训练, 不断重复该过程, 直到集合中的特征数量达到指定值, 即可选出最优特征子集[16]. 在本文中, 选择LightGBM (Light Gradient Boosting Machine)模型, 进行递归特征消除. LightGBM是一个基于决策树的梯度提升框架, 在传统的GBDT (Gradient-Boosting Decision Tree)算法上进行了优化, 支持多线程的并行计算, 在保证准确率的同时降低了内存的消耗, 训练速度也得到了极大程度的提高, 从而达到高效处理海量数据的目的[16].

3.2.4 嵌入式特征选择

嵌入式特征选择是在给定基学习器的情况下, 将特征数据与模型结合在一起, 在模型的训练过程中筛选掉系数为零的特征数据, 其计算代价较低, 特征选择速度快, 能极大程度上对数据进行降维. 本文选择随机森林和XGBoost (Extreme Gradient Boosting)学习器, 分别作为基模型, 进行特征选择[17–18]. 这两种嵌入式学习器都能较好地对特征间的非线性关系进行建模, 在特征选择的过程中, 模型会计算特征的相关性系数和对模型性能的贡献度指标, 当相关性系数或贡献度指标低于设定阈值时, 自动舍弃该特征.

3.3 基于贪心策略的集成特征选择

单一方法的特征选择无法全面地对数据特征进行评价, 而综合利用多种特征选择方法的优势, 是弥补单一方法局限的有效途径. 为此, 本文提出基于贪心策略的集成特征选择方法, 具体做法如下:

(1)使用每个单一方法提取的特征, 按重要性从大到小排序. 假设第i个方法给出的特征子集为$\boldsymbol{S}_i = \{s_{i,j}\,|i=1,2,3,4,5\,;j=1,2,3,\cdots,m\}$, $s_{i,j}$即表示第i个方法给出的特征子集中排序为j的特征, m表示该特征子集中特征总个数, 方法总数为n;

(2)取出各特征子集中排在首位的特征, 放入缓冲集合$\boldsymbol{B}$中, 对$\boldsymbol{B}$包含的特征进行去重后逐一输入至LightGBM分类模型, 得到对应的分类性能, 筛选出分类性能最好的特征, 记为c_1. 将c_1从$\boldsymbol{B}$中取出, 添加进集成特征集$\boldsymbol{C}$中;

(3)取出各特征子集中排在第2位的特征, 即s_{i2}, $i=1,2,3,4,5$. 将新选择的5个特征继续放入集合$\boldsymbol{B}$中并去重, 然后从$\boldsymbol{B}$中逐一取出元素与第1轮已经筛选出的最优特征c_1进行组合并输入至LightGBM分类模型中, 得到对应的分类性能. 从中筛选出性能最好的特征组合, 将第2个筛选出的特征记为c_2, 并将其从$\boldsymbol{B}$中取出, 添加进集成特征集$\boldsymbol{C}$中;

以此类推, 筛选出特征$c_3, c_4, \cdots, c_m$, 直到筛选出特征子集中包含的所有特征. 最后, 得到按特征重要性排序的集成特征集$\boldsymbol{C}=\{c_i|i=1,2,3,\cdots,m\}$. 集成特征选择方法的算法流程如下所示.

Algorithm: Ensemble feature selection method

Input: Ordered feature set list $\boldsymbol{S}_i$; Number of single feature selectors n; Number of features m

Output: Selected feature set $\boldsymbol{C}$ using ensemble feature selection method

```
Initialize temporary collection B and results feature collection C
for i = 1 to n:
    for j = 1 to m:
    B ← [B; s_ij]
    end for
    Remove duplicate features in B
    for k = 1 to size (B):
        Get out the k^th feature b_k
        if b_k is not in C:
            Compute classification performance of subsets {c_1, ..., c_{i-1}, b_k}
    end for
    Record the best performance feature b_l based on combination of b_k and C
    C ← [C; b_l]
end for
return C
```

4 结果与讨论

在实验时, 分别将属于单脉冲的弥散脉冲组和射频干扰弥散脉冲组按照6∶2∶2的比例进行随机分组, 划分成训练集、验证集和测试集, 然后以综合了精确率和召回率的F1值为主要评价指标, 主要实验结果如下.

4.1 不同神经网络模型的分类效果分析

随着卷积神经网络的发展, 涌现了越来越多的网络模型, 本文选取了部分代表性的网络模型, 进行单脉冲信号分类效果对比, 从而确定最优的网络结构并进行抽象特征的提取. 在实验时首先对每个网络通过自动搜索方式单独进行参数调优, 获得最佳效果. 各模型的实验结果如表2所示.

表 2 不同卷积神经网络的分类结果比较
Table 2 Classification results for different Convolutional Neural Networks

Model	Accuracy	Precision	Recall	F1-score
VGG16	0.937	0.955	0.872	0.903
MobileNet	0.940	0.900	0.922	0.911
GoogleNet	0.957	**0.963**	0.915	0.938
ResNet50	0.961	0.959	0.929	0.943
ResNet34	0.959	0.911	0.965	0.937
ResNet18	0.962	0.957	0.934	0.945
ResDFNet	0.966	0.932	**0.966**	0.949
RSDFNet	**0.968**	0.945	0.960	**0.953**

从中可以看出, 本文所使用的RSDFNet模型, F1值达到了95.3%, 在这些模型中的整体性能表现最好, 证明了RSDFNet具有较好的从信噪比与窗口宽度的分布曲线形态上学习和提取特征的能力. 与ResDFNet相比, RSDFNet引入了残差收缩模块后F1值提高了0.4%. 分析认为由于弥散脉冲组中的信号事件是分别基于信号事件表中每条记录的信号时间和色散值的邻近程度直接聚类进行分组而来, 没有考虑其中的相关性, 其分布曲线形态图上就容易存在非相关的噪声事件点特征. RSDFNet通过注意力机制从分布曲线形态图像上聚焦到这些不合理的特征点, 通过软阈值处理将其置为零, 进而加强了模型在这些分布曲线形态图上提取特征的能力.

4.2 卷积神经网络提取特征个数对单脉冲信号分类结果的影响

本文主要通过RSDFNet模型的最后一层隐藏层提取输入图像包含的抽象特征, 该层的节点数不同, 对模型的分类性能也有一定影响. 为此, 通过调整抽象特征数量, 即对RSDFNet最后一层隐藏层的节点数进行调整, 观察模型的分类性能以寻求更有质量的特征, 得到的实验结果如图3所示.

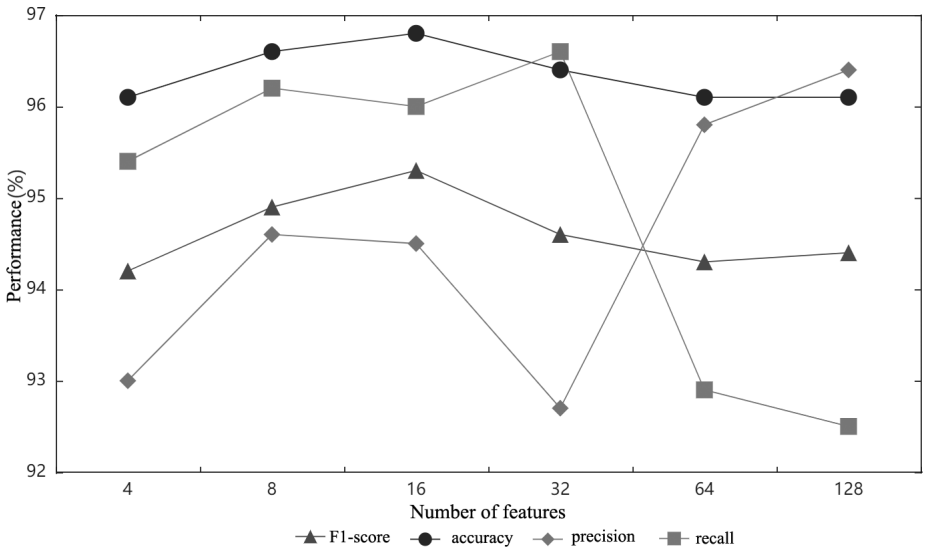

图 3 模型性能随抽象特征个数的变化

Fig. 3 Variations of model performance with the number of abstract features

图中可直观地看到, 随着抽象特征个数的增加, 模型的性能随之而提升; 当抽象特征个数达到16个时, 模型的F1最高. 此后, 随着特征个数的增加, 模型性能不再提升, 反而持续下降. 因此, 本文设置提取的抽象特征个数为16.

4.3 单一特征选择与集成特征选择的结果对比

根据本文第3部分描述的单一特征选择和集成特征选择方法, 对单脉冲信号的所有混合特征进行筛选, 不同特征选择方法得出的特征重要性排序如表3所示.

表 3 不同方法选择特征的重要性排序

Table 3 Feature importance ranking for different feature selection methods

Selection methods	Feature ranking based on importance from large to small
Chi-square test	f15, f2, f1, f14, f13, f7, f12, Duration, f11, f8, f9, f6, aDM, DM, S/N, TimeIndex, Time_Extent, Time, aTime, DM_Extent, KurtSigma, N_Events, wDM
Mutual Information	N_Events, wDM, S/N, aTime, Time, Time_Extent, TimeIndex, DM_Extent, KurtSigma, aDM, Duration, DM, f8, f12, f11, f14, f9, f2, f13, f15, f6, f1, f7
Recursive feature elimination	f9, f6, f14, f2, f15, f12, f8, DM, f1, Duration, f11, S/N, aDM, wDM, DM_Extent, Time_Extent, N_Events, KurtSigma, f7, aTime, f13, TimeIndex, Time
Random forest	f15, f9, f6, f2, f8, f13, f7, f1, f11, f14, f12, Duration, DM, aDM, Time_Extent, S/N, wDM, KurtSigma, DM_Extent, N_Events, Time, aTime, TimeIndex
XGBoost	f9, f15, f11, f6, f14, f2, f12, f8, Duration, DM, f1, S/N, wDM, DM_Extent, aDM, Time_Extent, KurtSigma, f7, N_Events, f13, Time, TimeIndex, aTime
Ensemble feature selection	f9, wDM, S/N, aTime, f8, f11, TimeIndex, Time, f6, DM_Extent, Time_Extent, f1, f15, f2, f13, f12, f14, f7, N_Events, Duration, KurtSigma, aDM, DM

表3中, 以“f+数字”命名的特征是基于深度残差收缩网络提取的抽象特征, 其他方式命名的为参数特征和统计特征. 从表中可以看出, 每种方法计算出的特征重要性排序明显不同. 以互信息方法选择的特征, 把统计特征和参数特征作为重要的特征, 而基于随机森林的嵌入式特征选择方法则把抽象特征作为重要的特征. 从集成特征选择的结果看, 抽象特征f9是最重要的特征, 然后是统计特征和参数特征. 总体上看, 单纯依靠一类特征, 例如只使用统计特征或者只使用深度残差收缩网络的抽象特征, 都不是最好的特征集合. 通过对多元特征进行集成选择是构建最优特征集的有效方法.

4.4 特征个数对模型性能的分析

在上一节中, 虽然得出了不同方法下特征的重要性排序, 但是能让分类模型得到最优结果的输入特征数量仍不确定. 因此, 本节继续讨论输入特征个数对模型性能的分析. 我们以LightGBM模型为例, 使用F1值为模型评价指标, 分别计算模型在不同输入特征个数下的F1值. 选择LightGBM是因为该模型相比于XGBoost等其他模型具有更快的训练速度和更高的效率, 而且适用于大规模数据的处理[19]. 另一方面, LightGBM本质是一种基于树的模型, 模型本身存在着较多的超参数, 这些超参数会影响树的结构、训练的速度以及模型的拟合度等. 同时, 这些超参数之间还存在相互影响, 如: 参数num_leaves既影响决策树结构, 又可以控制拟合程度; max_bin既与效率相关, 也与准确率相关, 还与拟合程度相关. 因此在应用时尽量避免手动调整参数, 最好是通过自动搜索的方式确定超参数. 本文选出了LightGBM模型中8个常用的超参数, 使用麻雀搜索算法对这些参数进行自动调整, 这8个参数及其取值搜索范围如表4所示.

为了分析特征个数对模型性能的影响, 并比较

单一特征选择方法和集成特征选择方法的表现, 进行相同特征个数下的对比实验分析. 按照表3中的特征重要性排序, 由小到大, 依此构建不同特征选择方法的特征子集, 分别输入LightGBM模型进行训练和分类结果预测, 基于分类结果计算5种单一特征选择方法的F1值, 取每个特征数量下5种单一方法特征子集的最大F1值和集成特征子集的F1值进行比较. 其值随着特征个数的变化如图4所示.

从图4可以看出, 随着输入特征个数的增加, F1值也迅速提升, 大概在输入8个特征的时候, 集成特征方法的F1值达到最大值, 在输入10个特征的时候, 单一特征方法的F1值达到最大值. 后面随着特征个数的增加, F1值都趋于平缓并略微下降. 这说明后续增加的特征可能属于冗余特征或者无效特征, 由此看出, 单脉冲分类时并不是使用的特征个数越多越好.

表 4 LightGBM超参数及取值搜索范围

Table 4 Super parameters and their value ranges for search of LightGBM

Parameters	Value ranges	Parameter meaning
max_depth	[3, 10]	The max depth for tree model
num_leaves	[7, 1023]	Max number of leaves in one tree
min_data_in_leaf	[0.0005, 0.05]	Minimal number of data in one leaf
learning_rate	[20, 60]	Shrinkage rate
bagging_fraction	[0.5, 1.0]	Randomly select part of data without resampling
feature_fraction	[0.5, 1.0]	Randomly select a subset of features on each iteration
reg_alpha	[0, 200]	Also named as Lambda_l1 which is a floating-point number that represents the L1 regularization coefficient.
reg_lambda	[0, 200]	Also named as Lambda_l2 which is a floating-point number that represents the L1 regularization coefficient.

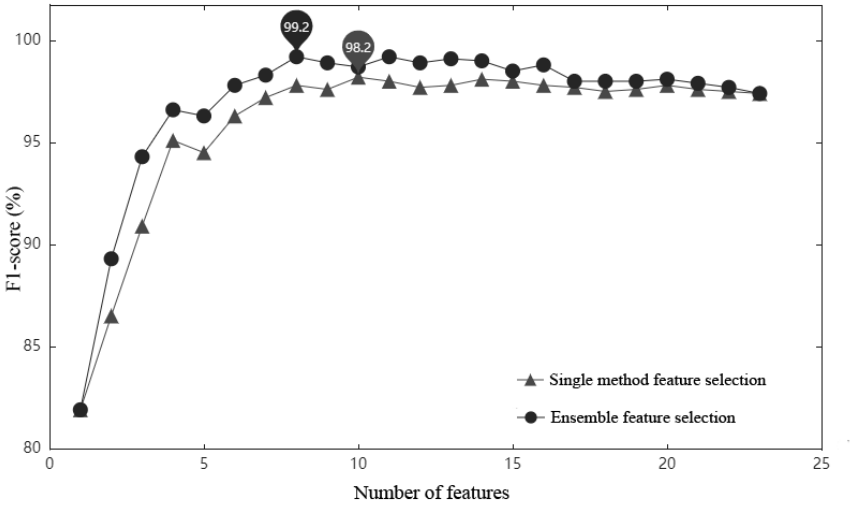

图 4 单一特征选择与集成特征选择方法的F1值随特征个数的变化

Fig. 4 Variations of F1-score with feature numbers for single method feature selection and ensemble feature selection

从单一特征选择和集成特征选择的对比来看, 在使用相同的特征数量下, 集成特征选择方法的分类结果都比单一特征选择的表现要好. 集成特征子集的F1值最高达到了99.2%, 在相同的特征数量下, 集成特征选择的F1值比单一特征选择的F1值最高可提升1.8%, 说明了集成特征选择方法的有效性. 集成特征选择方法结合了多种单一特征选择的结果, 更容易找到区分能力较强的特征.

4.5 抽象特征对不同模型的性能增益分析

根据前面的分析, 可以看出筛选出的最优特征集包含3个抽象特征和5个自定义特征. 这一方面说明依靠经验设计的一些统计特征是有效的, 同时也说明仅仅依靠人工特征不一定能取得最佳效果. 本节进一步分析神经网络提取的抽象特征对不同模型的性能增益情况. 我们通过实验对比只利用表1中的人工特征和结合RSDFNet提取的16个抽象特征之后, 在SVM (support vector machines)、KNN (K-Nearest Neighbors)、AdaBoost (Adaptive Boosting)和LightGBM等模型上对单脉冲信号的分类效果, 利用准确度和F1值分析抽象特征对不同模型的分类效果性能增益情况. 这些模型的参数均通过自动搜索的方式取得最优值, 各模型实验结果如表5所示.

表 5 抽象特征对不同模型的性能增益(UDF表示用户自定义特征)

Table 5 Performance improvement of abstract features on different models (UDF stands for User defined features)

Model	Accuracy			F1-score		
	UDF	UDF + abstract features	Changes	UDF	UDF+abstract features	Changes
SVM	0.954	0.95	−0.42%	0.901	0.928	3.00%
KNN	0.848	0.937	**10.50%**	0.784	0.902	**15.05%**
AdaBoost	0.966	0.969	0.31%	0.95	0.954	0.42%
LightGBM	0.975	**0.982**	0.72%	0.963	**0.974**	1.14%

通过表5可以看出, 增加抽象特征的输入后, 各模型的准确率和F1值大都出现了相应的提升, 尤其是对KNN模型的提升最大, F1值最高提升了15%. 虽然KNN的准确率和F1值提升最大, 但是LightGBM模型的准确率和F1值在增加抽象特征之前和之后都是最高的. SVM模型的准确率并没有提升, 反而出现了略微下降, 一方面可能是因为SVM分类界面通过少量特征就可以构建, 另一方面可能是因为抽象特征中包含了一些无效或冗余特征. 通过集成特征选择, 可以进一步筛选出最优特征组合.

5 总结

机器学习已成为单脉冲信号探测和识别的主要方法, 对脉冲信号的特征抽取成为影响机器学习效果的重要方面. 为此, 本文在参数特征、统计特征和抽象特征的基础上, 设计了集成特征的选择方法. 该方法首先利用卡方检验、互信息、递归特征消除、嵌入式特征选择等方法筛选出不同侧面的最优特征子集, 然后利用贪心策略从最优特征合集中, 筛选出用于最终分类的特征组合.

根据对实验结果的分析, 可以得出, 不同的特征选择方法, 其特征重要性的排序明显不同, 特征选择方法对分类精度有明显影响. 当特征数量较少时, 不同特征选择方法对分类结果的影响较大. 当特征数量超过10个时, 不同特征筛选方法的分类性能开始趋同. 与单一特征选择方法相比, 基于集成特征的F1值可提高1.8%, 说明集成特征选择对单脉冲分类精度有较好的提升.

从集成特征的构成来看, 集成方法选择的特征包含了神经网络提取的抽象特征、参数特征和统计特征. 这说明单纯依靠卷积神经网络的抽象特征或者单纯依靠人工设计的统计特征, 都很难达到最优的分类效果. 对多元特征进行混合应用是提升单脉冲信号分类的有效手段. 本文的工作, 给基于机器学习的单脉冲信号分类一种全新的认知, 通过选

取有效集成特征, 不但降低了特征个数而且提升了分类性能. 特征个数的降低进一步减少了模型数据处理的计算量, 在高速大规模巡天背景下对提升海量天文数据的处理效率具有重要意义.

参 考 文 献

[1] Cordes J M, McLaughlin M A. ApJ, 2003, 596: 1142

[2] McLaughlin M A, Lyne A G, Lorimer D R, et al. Nature, 2006, 439: 817

[3] Lorimer D R, Bailes M, McLaughlin M A, et al. Science, 2007, 318: 777

[4] 王元超, 郑建华, 潘之辰, 等. 深空探测学报, 2018, 5: 203

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

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

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

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

[9] Agarwal D, Aggarwal K, Burke-Spolaor S, et al. MNRAS, 2020, 497: 1661

[10] Agarwal D, Lorimer D R, Surnis M P, et al. MNRAS, 2020, 497: 352

[11] Michilli D, Hessels J W T, Lyon R J, et al. MNRAS, 2018, 480: 3457

[12] 凌毓, 张金区, 李乡儒, 李慧. 天文研究与技术, 2022, 19: 264

[13] He K, Zhang X, Ren S, et al. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016. Las Vegas, 2016: 770

[14] Kumar V, Minz S. Smart Computing Review, 2014, 4: 211

[15] Hira Z M, Gillies D F. Adv Bioinformatics, 2015, 2015: 198363

[16] Albawi S, Abed Mohammed T A, Al-Zawi S. 2017 International Conference on Engineering and Technology (ICET), August 21-23, 2017. Antalya, 2018: 1

[17] Kwak N, Choi C H. ITPAM, 2002, 24: 1667

[18] Chen T, Guestrin C. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: Association for Computing Machinery, 2016: 785

[19] Ke G, Meng Q, Finley T, et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree//Guyon I, Von Luxburg U, Bengio S, et al. Advances in Neural Information Processing Systems. NewYork: Curran Associates Inc., 2017: 3146

Ensemble Feature Selection Method for Single Pulse Classification

ZHANG Jin-qu[1] LING Yu[1] DU Ping[2] LI Xiang-ru[1] LI Hui[1]

(1 School of Computer Science, South China Normal University, Guangzhou 510631)

(2 School of Building Information, Guangdong Construction Vocational Technology Institute, Qingyuan 511500)

Abstract Affected by a large number of radio frequency interference signals, it has become an important task for astronomical data processing to quickly and accurately identify single pulse signals from massive observation data. Designing and extracting effective data features is the key issue for efficient identification of single pulse signals using machine learning. This paper proposes an ensemble feature selection method for single pulse signal classification. The method first mixed three types of features, including the parametric features, statistical features and abstract features of single pulse signals, and then used five independent feature selection methods to select the corresponding optimal feature set, respectively. At last, the features selected by the five independent methods are mixed and the greedy strategy was used to select the optimal ensemble feature set. The experimental results show that the ensemble feature set can improve F1-score by value of 1.8% at most and can obtain higher accuracy than the features selected by independent methods. Under the background of high-speed and large-scale sky survey, the ensemble feature selection method plays an important role in reducing the number of features, improving classification performance and speeding up data processing.

Key words pulse signal, radio pulsar, action variable, methods: data analysis