

基于 LightGBM 的 EAST 杂质破裂预警^①



孙召宏^{1,2}, 胡文慧¹, 袁旗平¹, 高彬富¹, 丁锐¹, 曾龙¹, 肖炳甲¹

¹(中国科学院合肥物质科学研究院等离子体物理研究所, 合肥 230031)

²(中国科学技术大学, 合肥 230026)

通信作者: 胡文慧, E-mail: huwh@ipp.ac.cn

摘要: 对全超导托卡马克核聚变实验装置东方超环 (EAST) 运行放电期间发生的杂质破裂进行预测对未来的聚变装置的长脉冲稳态放电有重要意义. 根据杂质破裂的物理特性筛选出的 2018 年的 334 炮杂质破裂炮数据以及 2021 年的 1 628 炮非破裂炮作为训练炮, 再由等离子体平衡、密度、电流以及辐射等 8 种诊断信号组成的训练样本以 LightGBM 算法训练出杂质破裂预测模型. 实验结果表明 LightGBM 算法模型可以对杂质破裂进行准确预测 (成功预测率 96.29%), 非破裂炮的误判率 6.87%. 研究结果证明利用 LightGBM 进行 EAST 等离子体杂质破裂预警是可行的方案.

关键词: 全超导托卡马克; EAST; 杂质破裂; 破裂预警; LightGBM

引用格式: 孙召宏, 胡文慧, 袁旗平, 高彬富, 丁锐, 曾龙, 肖炳甲. 基于 LightGBM 的 EAST 杂质破裂预警. 计算机系统应用, 2023, 32(1): 50-60. <http://www.c-s-a.org.cn/1003-3254/8922.html>

Prediction of EAST Impurity Disruption Using LightGBM

SUN Zhao-Hong^{1,2}, HU Wen-Hui¹, YUAN Qi-Ping¹, GAO Bin-Fu¹, DING Rui¹, ZENG Long¹, XIAO Bing-Jia¹

¹(Institute of Plasma Physics, Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei 230031, China)

²(University of Science and Technology of China, Hefei 230026, China)

Abstract: The prediction of impurity disruption during the discharge period of experimental advanced superconducting tokamak (EAST) is of great significance for the long-pulse steady-state discharge of future EAST. According to the physical characteristics of impurity disruption, the data of 334 impurity disruptive discharges in 2018 and 1 628 non-disruptive discharges in 2021 are selected as training discharges. Then, the training samples composed of eight diagnostic signals, including plasma equilibrium, density, current, and radiation signals, are used to train the impurity disruption prediction model by LightGBM. The test results reveal that the LightGBM model can accurately predict the impurity disruption, with a success rate of 96.29%, while for non-disruptive discharges, the false positive rate is 6.87%. The research results indicate that it is feasible to use LightGBM to predict plasma impurity disruption of EAST.

Key words: all-superconducting tokamak; experimental advanced superconducting tokamak (EAST); impurity disruption; disruption prediction; LightGBM

1 引言

托卡马克磁约束聚变装置是当下核聚变领域最具前景的实验装置, EAST (experimental advanced superconducting tokamak) 是我国自主研发的全超导托卡马

克装置^[1]. 在托卡马克磁约束聚变等离子体放电过程中, 等离子体破裂是一种十分危险却常见的现象, 会对装置第一壁和面向等离子体的材料及组件产生严重损害^[2]. 破裂过程中产生的热流^[3]和粒子流负荷能对第一壁材

① 基金项目: 国家磁约束核聚变能发展研究专项 (2018YFE0302100); 国家自然科学基金 (12005264, 12075285, U1867222, 12022511)

收稿时间: 2022-06-02; 修改时间: 2022-07-01; 采用时间: 2022-07-29; csa 在线出版时间: 2022-11-16

CNKI 网络首发时间: 2022-11-18

料造成辐照损伤,破裂时还会产生晕电流,导致真空部件有很大的电磁力荷载,严重时会导致组件脱落,损坏装置.破裂事件造成的原因主要有杂质进入等离子体、垂直位移事件、磁流体不稳定性、等离子体运行极限(包括比压极限,密度极限,安全因子极限)等.经统计^[4,5],如表1所示,在EAST装置2018–2019年的放电实验中,杂质破裂在所有破裂事件中占比接近50%,高于其他所有类型的破裂.研究针对杂质破裂的预警模型,并且将其运用到等离子体放电中去提前预警破裂,及时采取破裂缓解手段有助于EAST实现高参数长脉冲稳态放电.

表1 EAST完整放电炮、破裂炮和杂质破裂炮统计

实验年度	破裂炮	杂质破裂炮	完整放电炮	杂质破裂炮占破裂炮比例 (%)
2017	500	237	1 517	47.4
2018	726	292	1 606	40.2
2019a	450	136	1 926	30.2
2019b	673	287	1 780	42.6
总计	2 349	952	6 829	40.5

托卡马克磁约束实验中的杂质是指工作气体离子或原子以外的离子或原子.例如对于一个氢等离子体而言,除了氢离子(或原子)以外的离子(或原子)都属于杂质.其中,造成杂质破裂的主要为核电荷数大于18的高电荷数杂质,例如钨、铁和铜等,而在EAST上高电荷数杂质以钨杂质为主.托卡马克装置上杂质的来源主要有两种:第1种是等离子体与壁、限制器和偏滤器靶板的相互作用产生的,第2种是外界主动注入的杂质和放电实验中产生的聚变产物.其中,高电荷数杂质主要来源于第1种.EAST托卡马克内部腔室和第一壁材料构成如图1所示.EAST的面向等离子体材料接近于下一代托卡马克ITER^[6],而ITER上可预见的主要破裂依然是杂质破裂,因此EAST杂质破裂预测的研究对于下一代聚变装置的稳定运行具有重要意义.

相比于传统通过某些诊断信号或者物理模型来预警杂质破裂,基于数据驱动的机器学习算法可以同时兼顾实时性和准确性,即在保证运算速度以满足实时预警需求的时候还能保证预测的准确性.破裂过程分为先兆、热猝灭和电流猝灭^[7]这3个阶段.破裂预警是指希望在先兆阶段发生的初期就能准确预测到破裂的发生并由预警系统发出警报^[8].本文希望能够至少在破裂发生前30 ms预测到破裂的发生,让系统有时间做出相应的措施缓解或者避免破裂发生.

各大装置都在进行人工智能模型预测的探索,作为集成学习典型代表的随机森林已经被成功运用到EAST的高密度破裂预警与缓解的实时实验中^[9],成功预测率达到92%,误判率9.9%,可优化空间较大.并利用CNN和LSTM两种神经网络对EAST密度极限破裂进行对比实验分析^[5,10],CNN的准确率为82.7%,而LSTM的准确率为95%,误判率接近8%,但是所选特征信号不能实时采集故无法实时化,详细数据如表2所示.同时在JET、DIII-D、TJ-II和J-TEXT等装置都利用支持向量机、随机森林和神经网络等算法建立在线和离线的预测.JET利用SVM对进行破裂预测^[11],成功预测率81%.TJ-II装置上利用集成学习算法Ada-Boost算法进行放电识别和波形分类^[12],虽然不是对破裂的预测但是研究方法和步骤有很大参考价值.集成学习算法LightGBM被用于对J-TEXT和HL-2A两个装置进行跨装置破裂预测^[13],在J-TEXT上准确率为91.9%,误判率为10.1%,在HL-2A上为准确率和误判率分别为83.2%和11.1%,HL-2A上还利用卷积神经网络对破裂预警进行研究^[14],预测的真实阳性率达到92.2%,真实阴性率达到97.5%.

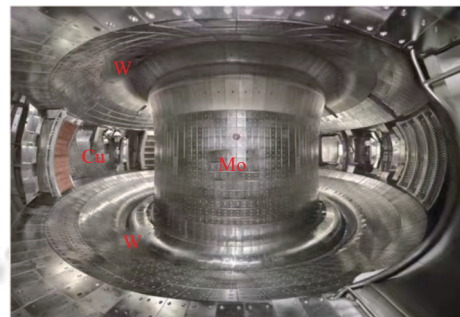


图1 EAST托卡马克内部腔室和第一壁材料构成

表2 EAST破裂预测模型参数对比

参数	CNN	LSTM	随机森林
准确率 (%)	82.7	95.37	92
误判率 (%)	6.7	7.9	9.9
使用信号数	14	11	9

基于以上,本文主要研究对EAST杂质破裂进行离线预测算法的研究,为后续的实时预测提供测试结果和理论基础.在结合文献^[5,9,10]研究的基础上进行改进,鉴于实时性需求,综合考虑到LightGBM算法^[15]是一款基于决策树算法的,被广泛用于分类和异常检测问题的轻量级集成学习算法,采用LightGBM对

EAST 放电过程中的等离子体状态进行杂质破裂事件的识别进行研究,对于现阶段借助机器学习探索核聚变物理性质有重要意义。

文章组织内容如下:第2节为数据预处理及杂质破裂数据库构建;第3节为破裂预警模型训练;第4节为模型测试结果分析;第5节总结了本研究的主要结果,并讨论了下一步研究计划。

2 数据预处理及杂质破裂数据库构建

2.1 杂质破裂特征信号选择

EAST 装置作为现有国内最先进的全超导托卡马克模型之一,经过十几年的探索拥有了种类繁多且全面的上百种诊断信号,选出展现杂质破裂特性的诊断信号需要长时间的调研以及对物理特性透彻的研究。在经过调研之后,基于物理特性和可实时获取特性最终根据如下3点选取标准选择了9种诊断信号进行实验:

- (1) 和杂质破裂特性密切相关。
- (2) 可以在等离子体控制系统实时获取。
- (3) 信号应该有各自特性,尽量做到物理特性不重叠。

在此准则下选择了归一化比压信号 (betaN), 等离

子体拉长比信号 (Kappa) 形状, 等离子体自感 (Li) 信号, 中心通道杂质辐射信号 (Pxuv32), 等离子体边界安全因子 (q95), 环电压信号 (Vloop), 等离子体储能 (Wmhd), 电流误差归一化值 ($I_p_error_norm$), 等离子体密度与 Greenwald 密度极限之比 ($Greenwald_fraction$) 这9个信号特征, 选取信号特征数值信息如表3所示^[16], 由表3中信号的物理含义可以看到所选诊断信号包括等离子体平衡、密度、电流以及辐射等多个物理方面, 所选信号符合上述的选择标准。其中 $I_p_error_norm$ 和 $Greenwald_fraction$ 是计算所得信号特征, 采样频率参考原使用信号频率。对于 $I_p_error_norm$ 信号和 $Greenwald_fraction$ 数学处理公式如下:

$$I_p_error_norm = (I_p - I_p_target) / I_p_target \quad (1)$$

$$Greenwald_fraction = n_e / n_{GW} \quad (2)$$

$$n_{GW} = I_p / (\pi a^2) \quad (3)$$

其中, I_p 为等离子体电流, I_p_target 为电流控制的目标值, $I_p - I_p_target$ 是电流控制的误差, n_{GW} 是 Greenwald 密度极限, n_e 是密度。挑选上述与放电过程中杂质破裂很相关的信号, 还对这些信号的采集数据进行了初步预处理, 只留下在物理特征允许的正常范围的样本, 剔除异常样本。

表3 特征信号基本信息

信号名	物理含义及包含的等离子体信息	训练集中最大值	训练集中最小值	原始采样率 (kHz)
betaN	归一化比压, 是等离子体压强比上磁场压强的归一化值, 包含了等离子体压强和相对于磁场压强大小的信息	2.21998	0.02287	0.02
Kappa	等离子体拉长比, 是等离子体横截面的纵向总长度比上横向总长度, 包含了等离子体几何形状的信息	1.9972	0.98984	0.02
Li	等离子体内自感, 包含了等离子体电流密度剖面分布的信息	2.29165	0.35031	0.02
Pxuv32 (V)	中心杂质辐射信号, 是等离子体横截面中间的一道辐射计测量弦的辐射信号强度, 包含了等离子体杂质含量的信息	125.5405	0	100
q95	磁边界安全因子, 表示等离子体边界的磁面上磁力线绕磁面形成的螺旋形轨迹从出发点, 再次回到出发点所旋转的角度比上 2π , 包含了等离子体平衡的信息	14.42347	1.94253	0.02
Wmhd (kJ)	等离子体储能, 包含等离子体温度、密度的信息	323405.3	2869.89	0.02
Vloop (V)	等离子体环向回路电压测量信号, 通过测量与等离子体平行的环形线圈两端的电压来确定, 包含了等离子体电阻和焦耳热的信息	7.94313	-10.0097	1
$I_p_error_norm$	等离子体电流控制误差的归一化值, 计算方法为 $(I_p - I_p_target) / I_p_target$, 包含了等离子体电流控制误差和等离子体电流稳定性信息	0.24845	-0.42522	1
$Greenwald_fraction$	等离子体 Greenwald 密度极限因子, 计算方法为 n_e / n_{GW} , n_e 为等离子体密度, n_{GW} 为 Greenwald 密度极限, 包含了等离子体密度和距离密度极限边界有多远的信息	9.2876	-0.12218	0.02

图2 是一个典型非破裂炮 (#97456), 0–2 s 为电流爬升段, 2–7.91 s 为平顶段, 7.91 s 后电流进入下降段,

电流缓慢降低至零, 这一炮放电结束。图3 是典型的杂质破裂炮 (#77083), 0–1.7 s 是电流爬升段, 1.7 s 后电流

进入平顶段, 4 s 之后由于杂质出现, 导致辐射信号 Pxuv32 增强, 到 4.93 s 时电流陡降为零, 电流猝灭, 等离子体发生破裂, 等离子体放电中止。

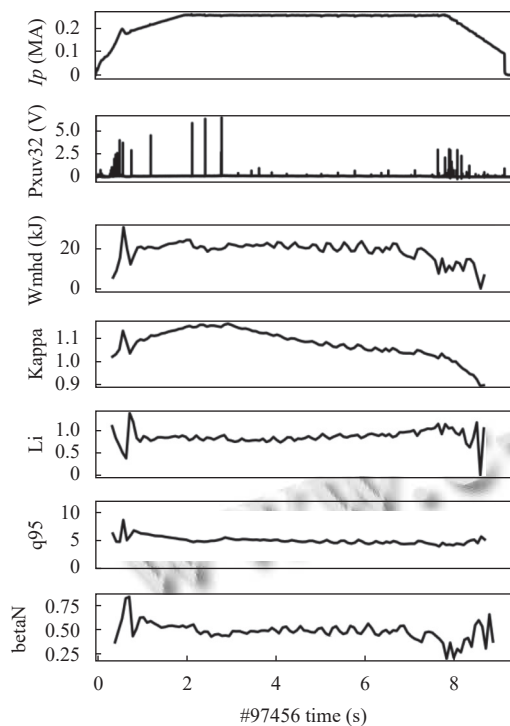


图2 典型非破裂炮诊断信号展示 (#97456)

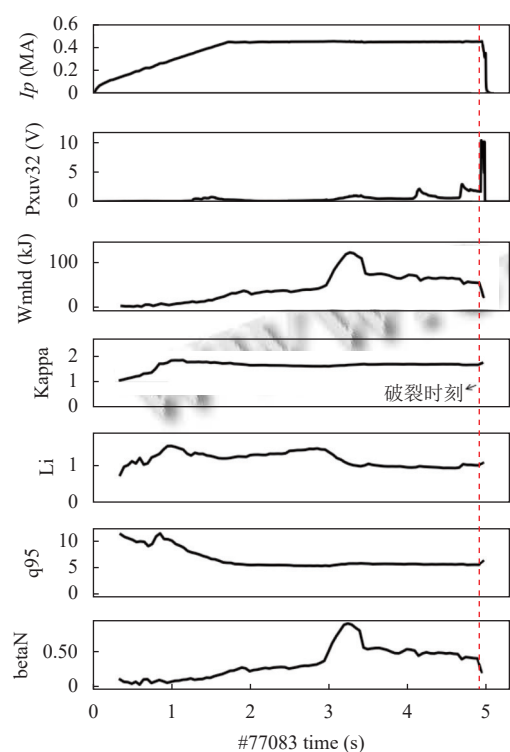


图3 典型杂质破裂炮诊断信号展示 (#77083)

2.2 数据采集和预处理

EAST 上采用 MDSplus 程序^[17] 对数据进行采集存储, MDSplus 作为采集、存储和处理复杂科学数据的一套软件系统, 在磁约束聚变研究领域被广泛的使用. 数据采集、清洗、预处理是本文针对 MDSplus 开发了一套基于 Python 的数据采集、清洗、预处理系统, 使用 MDSplus 自带的 Python 接口和 Pandas, Numpy 等开源库实现本文所需破裂时刻点计算功能^[18].

由图 3 可知破裂会让电流在极短的时刻发生猝灭, 在电流稳定的时刻称为平顶端, 往往平顶端结束极短的时间就会发生电流的急剧下降, 破裂时刻的数值被定义为: 等离子体电流一阶导数最大值出现的时刻. 如图 3 是典型杂质破裂炮, 可以看到相对于图 2 典型非破裂炮其电流是呈直线下降, 且很快就下降为 0, 而非破裂炮会在数秒时间内较为平稳的湮灭. 根据上述破裂的特性^[17], 开发了相应程序对破裂时刻进行准确计算, 其算法流程如图 4 所示, 其核心流程^[18] 如下。

(1) 采集 I_p 电流信号并且做中值滤波去除噪音并判断平顶段时长是否大于 0.6 s, 这是稳定放电的判定标准。

(2) 由引言可知电淬灭阶段持续阶段不会超过 0.01 s, 电流会骤降到 100 kA 以下, 寻找电流值大于 200 kA 的最后时段 t 并计算前后 0.05 s 的 $\text{abs}(dI_p/dt) > 3.3 \text{ MA/s}$ 的时段, 并且此时段总长度要小于 50 ms, 破裂时刻发生在此时段内。

(3) 判断得到时刻 15 ms 后电流值是否小于 100 kA, 以及时刻电流值是否大于 100 kA 同时大于放电最大电流的 1/3, 这是为了排除电淬灭之后的小电流破裂^[19]。

2.3 杂质破裂训练数据库建立

本文依据 Pxuv32 辐射信号和 I_p 电流信号建立 EAST 杂质破裂数据库, 分为 3 个步骤。

(1) 首先对实验炮区分破裂炮和非破裂炮. 主要依据等离子体电流信号 I_p , 对于非破裂炮, 电流有爬升、平顶段、下降阶段, 而对于破裂炮, 电流只有爬升和平顶段, 本文需要的实验数据是平顶段时刻。

(2) 紧接着从所有破裂炮中挑选出杂质破裂炮, 挑选依据是根据破裂时刻 Pxuv32 辐射信号变化是否陡增, 这需要人工判定。

(3) 最后剔除第 2.2 节中缺失信号的实验炮 (针对所有实验炮), 得到完整实验数据集, 针对数据集的划分如表 4 所示。

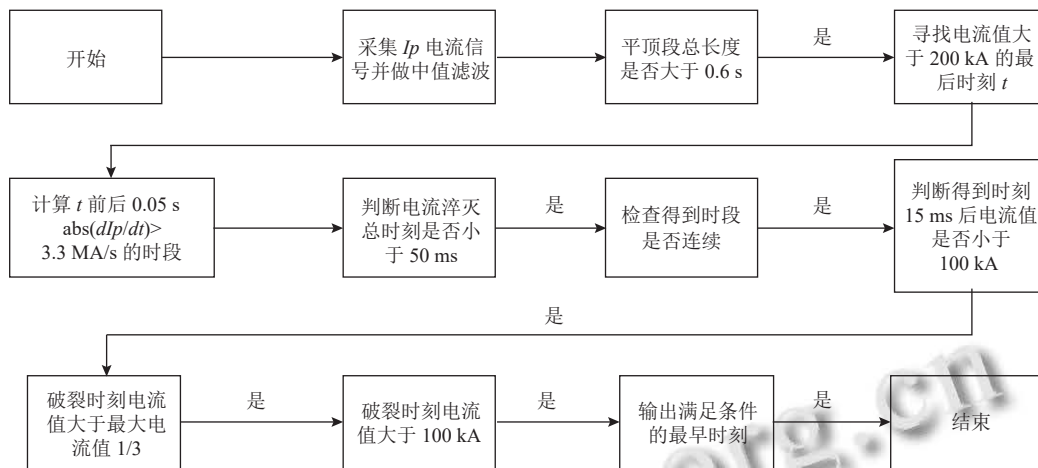


图 4 破裂时刻计算步骤

表 4 LightGBM 训练数据集的划分

划分	训练集	验证集	测试集	合计
破裂炮	334	334	405	1073
非破裂炮	1624	1624	1628	4876
总计	1958	1958	2033	5949

2.4 杂质破裂数据样本处理

在得到完整的数据样本之后,需要对采集的数据进行数据重采样使信号采样率一致,考虑到破裂发生的时刻极短,预警时间需要控制在实际破裂 10 ms 之前,调用了 Python 的 interp1d 库函数进行重插值将实验数据时间间隔设置在 1 ms 得到新的样本数据 [20].

对样本贴标签是监督学习的必要步骤,本研究针对杂质破裂炮数据样本进行稳定和 non-稳定样本的划分,并编码为 0 和 1. 划分 0/1 样本的目的是让模型学习杂质

破裂和非破裂时刻样本信息的差异. 为了能够准确区分出破裂时刻和非破裂时刻,本文应用基于物理特性的手动样本划分方法,将破裂发生前杂质信号波动时刻作为 0/1 样本的分界时刻,在电流平顶端开始到分界时刻数据都为 0 样本,而在此时刻之后到电流破裂时刻数据都为 1 样本,这样的标签判定使得样本数据的划分更为准确.

如图 5 所示是一个典型的杂质破裂炮,在 3.65 s 发生因杂质抖升造成破裂,为了达到提前预测的效果,本文在 3.502 s 左右先兆阶段将来来临杂质波动时刻进行 0/1 标签的判定,从平顶段开始到 3.502 s 时刻数据都是 0 样本,在 3.502 s 以后到平顶段结束破裂时刻即将开始都是 1 样本. 对于非破裂炮数据,由于最终放电结束电流能够逐渐湮灭没有任何破裂情况的发生,所以对于非破裂炮平顶段数据所有样本都判定为负样本.

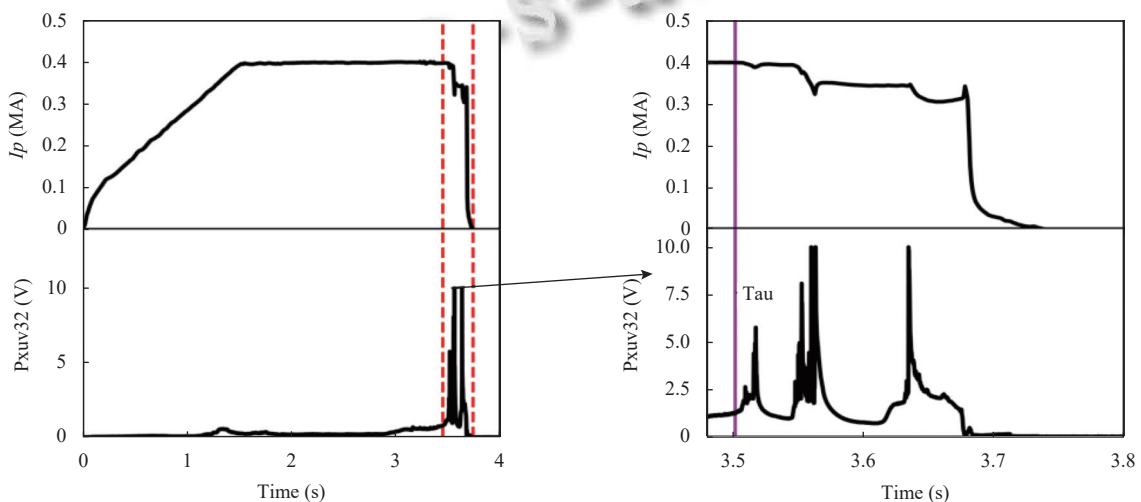


图 5 杂质破裂炮 0/1 标签判定方法 (#77069) (Tau 表示标签时间点)

3 模型训练

3.1 AdaBoost 以及 GBDT 算法介绍

AdaBoost 算法即自适应增强算法^[21], 它的分类思想为当下主流 Boosting 算法所运用 (包括 LightGBM 算法). AdaBoost 算法的自适应在于: 前一个基学习器分错的样本会得到加强, 加强后的全体样本再次被用来训练下一轮基本分类器, 到指定的迭代次数, 根据分类器分类效果对每个弱学习器赋予权重, 结合成强学习器, 其算法流程如图 6 所示.

图 6 中的计算单元数学公式及原理如下. D_i 集合是指第 i 次训练权重分布, w_{ij} 是指第 i 次迭代过程第 j 个数值的权重, $G_m(x)$ 是第 m 次训练的弱分类器, e_m 是 $G_m(x)$ 在样本训练集上的分类误差率, a_m 是 $G_m(x)$ 弱分类器上的权重系数, $f(x)$ 是最终学习的强分类器线性组合.

$$D_i = (w_{i1}, w_{i2}, \dots, w_{ij}, w_{iN}) \quad (4)$$

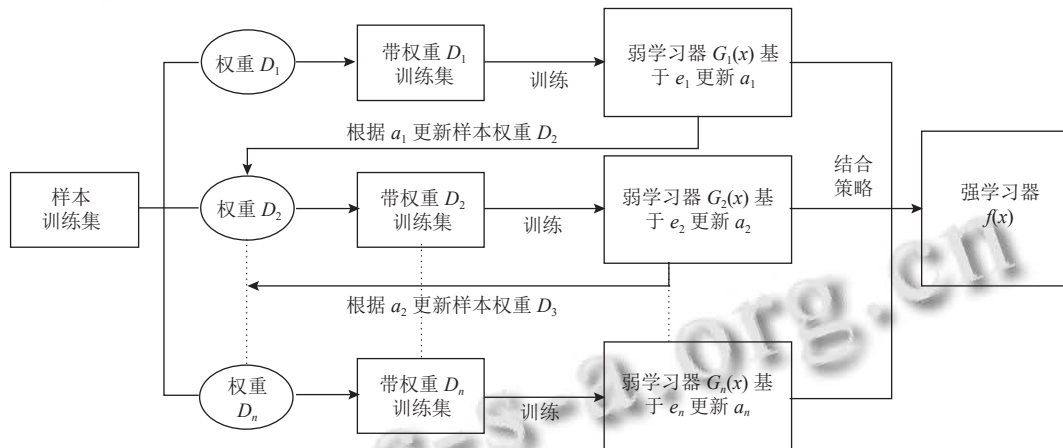


图 6 AdaBoost 算法流程

3.2 LightGBM 算法介绍

LightGBM 是 Microsoft 在 2017 年提出的开源机器学习模型^[15], 它是 GBDT 算法的改进算法, 是基于决策树的梯度提升集成学习框架. LightGBM 的分割点选择是基于直方图的算法, 先把连续的特征值离散化成 k 个整数, 同时构造一个宽度为 k 的直方图. 对于连续特征来说, 装箱处理就是特征工程中的离散化: 如 $[0, 0.5)$ s 时间段内的样本都定义为 0, $[0.5, 1.0)$ s 时间段内的样本都定义为 1, 这样大大提高了选择分割点的速度. 对于决策树的生长策略, LightGBM 是使用 leaf-

$$w_{ij} = \frac{w_{i-1j}}{Z_m} \exp(-a_m y_i G_m(x_i)) \quad (5)$$

$$G_m(x) : x \rightarrow -1, +1 \quad (6)$$

$$e_m = \sum_{i=1}^N w_{mi} I(G_m(x_i) \neq y_i) \quad (7)$$

$$a_m = \frac{1}{2} \log \frac{1 - e_m}{e_m} \quad (8)$$

$$f(x) = \sum_{m=1}^N a_m G_m(x) \quad (9)$$

GBDT (gradient Boosting decision tree) 算法是以分类树或回归树为弱学习器的集成学习算法, 它也运用了 AdaBoost 算法的思想, 并且每一轮训练是拟合上一轮的残差 (即平方损失函数) 来生成新一轮的决策树, gradient Boosting 意味着沿着梯度方向进行拟合将每一轮的目标函数看做梯度下降的参数来优化.

wise 的策略, 每次分裂选择收益最大的点分裂, 而不是每层对每一个点都进行分裂, 可以提高模型的精度. LightGBM 的效率更高、内存使用率更低、模型效果更好, 支持并行学习^[22], 同时它提供 C, C++ 等语言接口, 为实时预警提供了实现方案.

3.3 参数调优

线下模型的训练是实验最为关键的步骤, 在选取并处理好需要的信号特征及样本数据之后, 我们选用五折交叉检和超参数优化的方法对模型进行调参训练.

五折交叉检验法是 K-CV 交叉验证法的一种, 该

类方法最大的优点就是很好地规避了过拟合和欠拟合现象的发生^[23]。五折交叉验证是将总样本随机分割成5个大小一致的子样本,一个单独的子样本被保留作为验证模型的数据,其他4个样本用来训练。对上述步骤重复5次,每一个子样本都作为一次验证数据,最后计算5次求得分类错误率的平均值,作为该模型或者假设函数的真实分类错误率。

贝叶斯超参数优化^[24]是目前最高效的自动化调参方法之一,在机器学习中定义模型属性或者定义训练过程的参数称为超参数。贝叶斯优化算法克服了网格搜索的速度缓慢占用资源过多和随机搜索搜索空间不完整的缺陷,会充分利用之前测试点的信息规避不必要的资源浪费。具体步骤为:首先根据先验分布,假设一个搜索函数;然后每次使用新的测试点来测试目标函数时,利用这个信息来更新目标函数的先验分布;最后,算法测试由后验分布给出的全局最优值可能出现的位置的点。在这里本文选择高斯过程分布作为待优化函数的先验分布是因为高斯过程具有良好的数学特性,使得基于贝叶斯推导的分布更新有解析解,方便算法构建。

在使用贝叶斯超参数优化找到最佳参数之后开始模型的训练,这里经过重复测试发现230轮左右模型迭代到最优模型,图7为迭代230次后训练集上的训练效果,横坐标为迭代次数,纵坐标为训练AUC值。

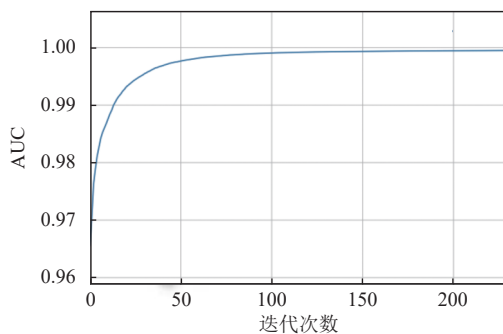


图7 模型迭代训练图

本文对于训练模型采用AUC (area under ROC curve) 统计,ROC曲线的AUC是机器学习领域一个很重要的模型判定标准,ROC曲线的横坐标为假正率FPR,纵坐标为真正率TPR,通过设定不同的判定样本的阈值(threshold),可以得到不同的TPR和FPR点对,将一系列点对连接成平滑的曲线,则为ROC曲线。而

AUC为ROC曲线与x轴以及直线 $x=1$ 构成图形面积,AUC值越接近1说明拟合效果越好。FPR、TPR定义如下:

(1) 伪阳性率 (FPR): 判定为正例却不是真正例的概率,即真负例中判为正例的概率, $FPR=FP/N=FP/(FP+TN)$ 。

(2) 真阳性率 (TPR): 判定为正例也是真正例的概率,即真正例中判为正例的概率(也即正例召回率),其中 $TPR=TP/P=TP/(TP+FN)$ 。

TP、FP、FN、TN定义如表5所示。

表5 混淆矩阵

真实结果	预测结果	
	预测正例	预测负例
正例 (P)	真正例 (TP)	假反例 (FN)
反例 (N)	假正例 (FP)	真反例 (TN)

离线模型在测试集上对于单个样本时刻预测准确率为99.95%。训练模型由9个特征信号230棵决策树组成,图8为其中第1棵决策树前5层模型。同时使用支持向量机(SVM)和简单BP神经网络对数据集进行拟合分析,支持向量机选用惩罚参数13的高斯核函数模型,BP神经网络第1层隐层节点200个,第2层50个节点,激活函数是Sigmoid,损失函数为MSE。拟合效果如表6所示,可以看到SVM与简单BP神经网络拟合效果不如LightGBM。

3.4 信号特征重要性分析

本文基于LightGBM自带的plot_importance库函数对9个信号特征进行数据敏感度分析,也就是每个特征在对实验数据进行预测时所占的重要程度,计算结果如图8所示,可以看到在杂质破裂中q95磁表面安全因子的变化最具代表性,betaN、Kappa、Li、Wmhd是很重要的平衡特征,Pxuv32是辐射信号,Ip_error_norm和Greenwald_fraction(GW_fraction)这两种归一化信号也起到重要作用。如图9所示是各信号所占重要性的比重,总比重为1。

在特征分析中发现Vloop的重要程度偏低,本文认为Vloop信号在该模型的作用很小。为了进一步验证假设,本文针对有无Vloop信号分别训练了两个LightGBM模型并进行了离线测试对比,结果如表7所示,可以看到不使用Vloop信号的模型在数据集上的表现效果更好,最终模型也采用只有8个信号230棵决策树的LightGBM模型。

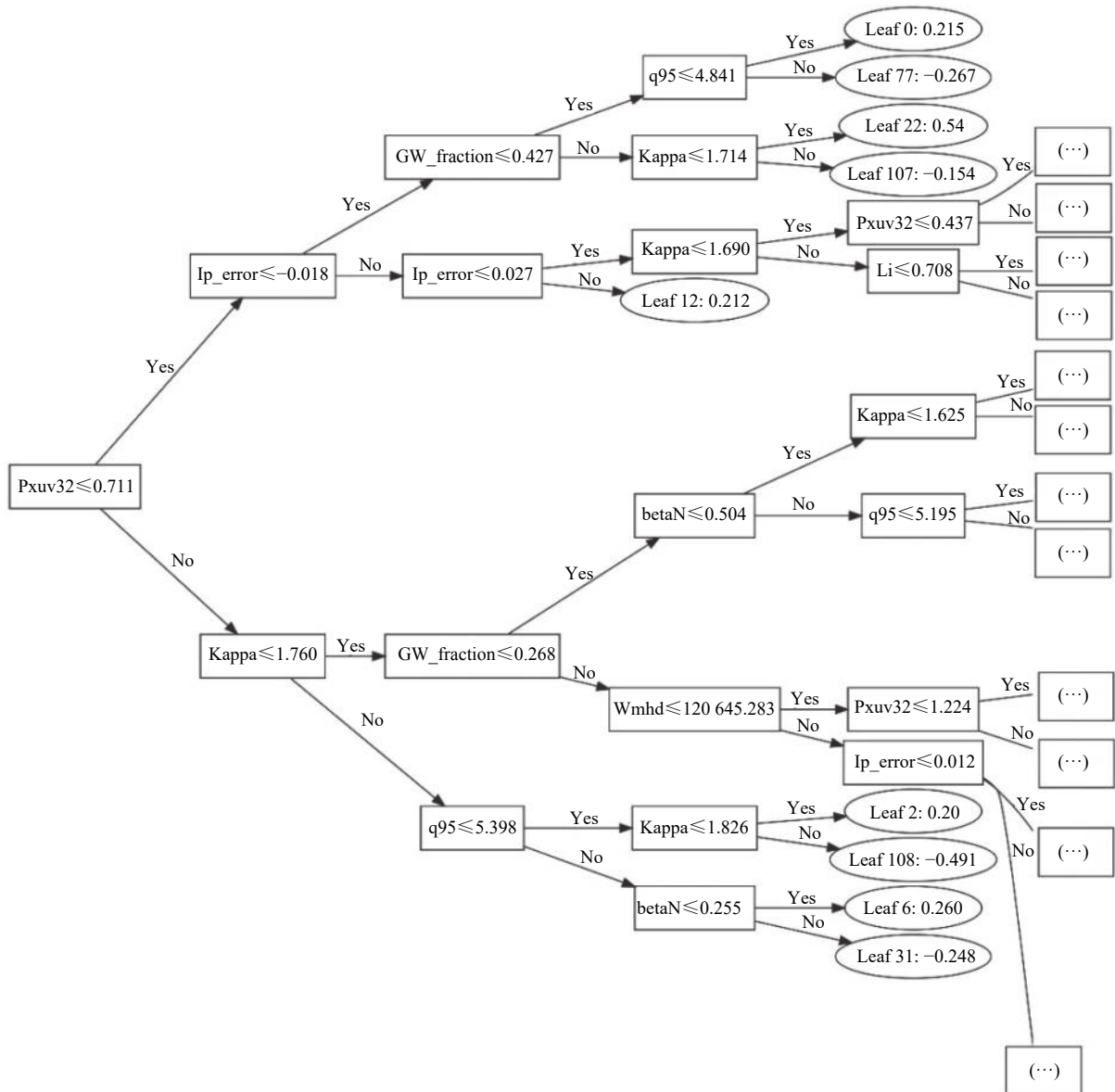


图8 模型第1棵树前5层模型

表6 模型性能度量 (%)

选用模型	SVM	BP神经网络	LightGBM
AUC评估指标	96.82	98.3	99.95

3.5 判定标准

对于杂质断裂预测炮的预警, 希望判定算法准确及时的预测到杂质断裂炮 (正例) 而对于非断裂炮 (反例) 尽可能少的发生错误的判定. 判定算法不能以单个样本的判定结果作为判定依据, 因为不是每次杂质波动都会造成断裂, 需要综合考虑在长脉冲稳定放电后所有的放电时刻数据, 同时要综合考虑时间变化的因

素, 其本质是一种时序预测的模型. 输入给模型的数据是当前时刻 8 种特征信号的数值, 从放电稳定开始, 每隔 1 ms, LightGBM 模型用 8 种特征信号作为输入, 计算一个断裂概率作为输出, 输出的断裂率组成了一个时间序列, 当断裂概率大于提前设置预警阈值, 并持续时间达到 10 ms 时会产生一个断裂预警信号. 对于离线预测样本根据相关的研究定义以下指标.

(1) 对于杂质断裂炮 (正例)

① 成功预警炮: 实际为杂质断裂炮, 预测也为杂质断裂炮. 预测模型至少在杂质断裂发生前 10 ms 触发

预警,而重插值时间间隔为 1 ms,即对于连续 10 个输入样本数据都达到阈值,该杂质破裂炮被成功预警。

② 漏判炮:实际为杂质破裂炮,预测为非破裂炮或者在实际破裂 10 ms 之前未发出破裂预警。破裂预测与其他预测在此有一定差别,因为本文希望在预测到破裂的发生后有至少在杂质破裂前 10 ms 内未触发预警或者在破裂最后 10 ms 内才触发了预警,此时破裂缓解系统没有足够的时间避免或者缓解破裂。也认为始终没有触发预警。

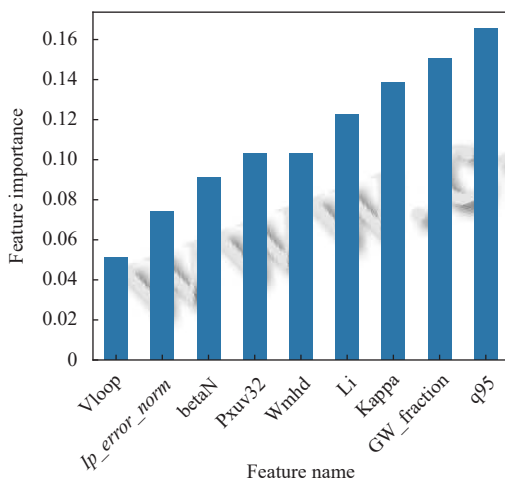


图9 信号重要性分析

表7 LightGBM (决策树数量=230) 在有无 Vloop 特征信号下在测试集上表现效果

阈值	特征信号数=9		特征信号数=8	
	成功预警炮 (%)	误判炮 (%)	成功预警炮 (%)	误判炮 (%)
0.9	95.8	6.1	96.2	5.4
0.8	96.2	7.9	97.5	7.1
0.7	97.5	9.2	98.2	8.6

(2) 若为非破裂炮 (反例)

① 误判炮:实际为非破裂炮,预测为破裂炮。如果模型在长脉冲放电平稳端任意时刻触发预警,但是这次放电实际是一次非破裂炮,则此非破裂炮被误判为杂质破裂炮,因此为错误预警炮,称为误判炮。

② 安全炮:实际为非破裂炮,预测也是非破裂炮。如果模型在长脉冲放电平稳端任意时刻未触发预警,则此非破裂炮正确预测,称为正确未预警炮。

对于评价标准:

PFA (percent of false alarm, 误判率): 错误预测的总非破裂炮数占非破裂炮的比重。

PSA (percent of successful alarm, 成功率): 成功预测的总破裂炮数占破裂炮总数的比重。

PMA (percent of missed alarm, 遗漏率): 遗漏预测的总破裂炮数占破裂炮总数的比重。

预警时间: 预测破裂时间与实际预警时间之间的间隔。

4 模型离线预测结果分析

如图 10 所示展现的是在 0.7 到 0.95 不同阈值下正确预警率、误判率以及预警时间的变化,实验给出的判定要求是正确预警率大于 95%,误判率低于 10%,预警时间在 100 ms 以上。本文综合考虑最终挑选 0.8, 0.9, 0.95 这 3 个阈值作为理想判定阈值,此时的正确率、误判率以及预警时间都在理想状态。在阈值 0.8、0.9 和 0.95 的时离线测试结果成功率分别为 97.53%、96.28% 和 95.06%,误判率分别为和 7.12%、4.91% 和 3.5%,平均预警时间为 1.986 s、1.740 s 和 1.545 s,在实时控制系统中可以根据实验状况调整阈值^[25]。

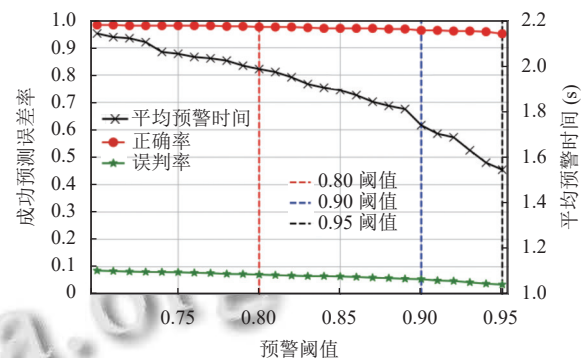


图10 成功预测率和误判率以及预警时间随阈值变化

最终选取模型是离线数据构建的 LightGBM 模型一共有 230 棵每棵树有 125 叶子节点组成的决策树。图 11 展示的是#98275 非破裂炮的预测结果,可以看到在该炮平顶段时刻稳定放电都没有触发预警,模型准确未发生误判。图 12 展示的是#90126 炮杂质破裂炮的预测结果,可以看到该炮在 3.43 s 发生杂质破裂,LightGBM 模型在 3.18 s 触发破裂预警系统,提前 0.25 s 准确预警杂质破裂的发生。进一步研究模型对杂质破裂炮的预警时间分布,如图 13 所示,横坐标是预警时间与破裂时刻差值,纵坐标为炮号数,可以看到破裂预警时间集中在 1 s 左右,离线测试数据最早预测时间为 36 ms。

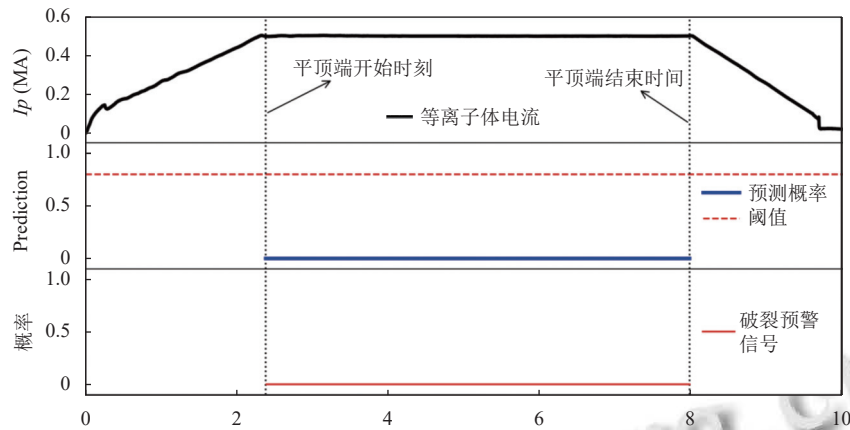


图 11 #98275 非破裂炮模型预测结果

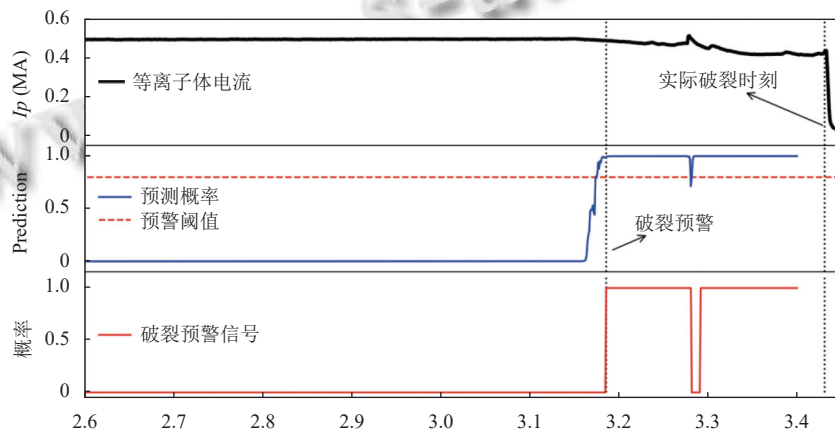


图 12 #90126 杂质破裂炮模型预测结果

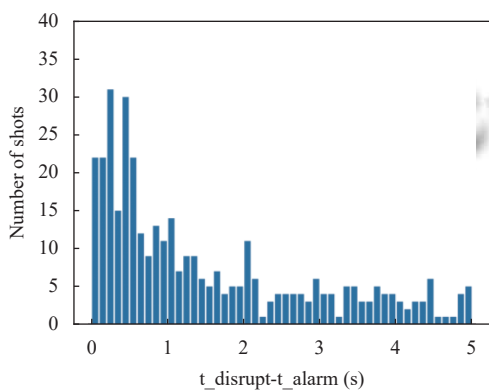


图 13 Threshold=0.8 下杂质破裂炮预警时间统计

5 总结

通过实验表明基于 LightGBM 的杂质破裂预测模型能够达到 EAST 托卡马克实验装置对于杂质破裂的

准确率, 误判率预测要求, 对于杂质破裂的预警对未来更大规模托卡马克装置有重要意义. 本文首次针对 EAST 采用 LightGBM 集成学习进行杂质破裂预测研究. 首先通过对杂质破裂物理性质的调研筛选出符合要求的信号特征, 同时开发基于 Python 的数据采集、清洗、预处理程序并建立杂质破裂数据库和数据集, 手动对采集到的训练数据进行标签判定, 采用贝叶斯优化和五折交叉检验方法训练出离线模型, 并根据模型的分析结果进行特征数据再次调整. 在阈值 0.8、0.9 和 0.95 的时离线测试结果成功率分别为 97.53%、96.28% 和 95.06%, 误判率分别为 7.12%、4.91% 和 3.5%, 平均预警时间为 1.986 s、1.740 s 和 1.545 s, 满足实时破裂的预设需求. 实验结果说明利用 LightGBM 建立杂质破裂离线预测算法的可行性. 下一步研究计划是将杂质破裂预警模型嵌入 EAST 的控制系统, 实现杂质破裂概率的在线计算, 实时监测杂质破裂事件.

参考文献

- 1 李建刚. 托卡马克研究的现状及发展. 物理, 2016, 45(2): 88–97. [doi: 10.7693/wl20160203]
- 2 何开辉, 潘传红, 冯开明. 托卡马克等离子体大破裂及防治综述. 中国核科技报告, 2002: 117–129.
- 3 Lehnen M, Arnoux G, Hartmann N, *et al.* Disruption heat loads and their mitigation in JET with the ITER-like wall. *Journal of Nuclear Materials*, 2013, 438: S102–S107. [doi: 10.1016/j.jnucmat.2013.01.016]
- 4 Gao BF, Ding R, Xie H, *et al.* Plasma-facing components damage and its effects on plasma performance in EAST tokamak. *Fusion Engineering and Design*, 2020, 156: 111616. [doi: 10.1016/j.fusengdes.2020.111616]
- 5 Guo BH, Shen B, Chen DL, *et al.* Disruption prediction using a full convolutional neural network on EAST. *Plasma Physics and Controlled Fusion*, 2021, 63(2): 025008. [doi: 10.1088/1361-6587/abcbab]
- 6 Wang L, Guo H Y, Ding F, *et al.* Advances in plasma—Wall interaction control for H-mode operation over 100 s with ITER-like tungsten divertor on EAST. *Nuclear Fusion*, 2019, 59(8): 086036. [doi: 10.1088/1741-4326/ab1ed4]
- 7 Greenwald M. Topical review: Density limits in toroidal plasmas. *Plasma Physics and Controlled Fusion*, 2002, 44(8): R27–R53. [doi: 10.1088/0741-3335/44/8/201]
- 8 陈大龙, 沈飙. EAST装置上 halo 电流特性分析与偏滤器载荷估算. 核聚变与等离子体物理, 2013, 33(1): 66–71. [doi: 10.3969/j.issn.0254-6086.2013.01.012]
- 9 Hu WH, Rea C, Yuan QP, *et al.* Real-time prediction of high-density EAST disruptions using random forest. *Nuclear Fusion*, 2021, 61(6): 066034. [doi: 10.1088/1741-4326/abf74d]
- 10 陈俊杰, 胡文慧, 肖建元, 等. 基于神经网络的 EAST 密度极限破裂预测. 计算机系统应用, 2020, 29(11): 21–28. [doi: 10.15888/j.cnki.csa.007668]
- 11 Ratta GA, Vega J, Murari A, *et al.* An advanced disruption predictor for JET tested in a simulated real-time environment. *Nuclear Fusion*, 2010, 50(2): 025005. [doi: 10.1088/0029-5515/50/2/025005]
- 12 Farias G, Fabregas E, Martínez I, *et al.* Nuclear fusion pattern recognition by ensemble learning. *Complexity*, 2021, 2021: 1207167.
- 13 Zhong Y, Zheng W, Chen ZY, *et al.* Disruption prediction and model analysis using LightGBM on J-TEXT and HL-2A. *Plasma Physics and Controlled Fusion*, 2021, 63(7): 075008. [doi: 10.1088/1361-6587/abfa74]
- 14 Yang ZY, Xia F, Song XM, *et al.* A disruption predictor based on a 1.5-dimensional convolutional neural network in HL-2A. *Nuclear Fusion*, 2020, 60(1): 016017.
- 15 Ke GL, Meng Q, Finley T, *et al.* LightGBM: A highly efficient gradient Boosting decision tree. *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Long Beach: Curran Associates Inc., 2017. 3149–3157.
- 16 Wesson J. 托卡马克. 王文, 译. 第 4 版, 北京: 清华大学出版社, 2021.
- 17 Fredian T, Stillerman J, Manduchi G, *et al.* MDSplus yesterday, today and tomorrow. *Fusion Engineering and Design*, 2018, 127: 106–110. [doi: 10.1016/j.fusengdes.2017.12.010]
- 18 唐剑寅, 肖炳甲, 袁旗平. 对 EAST PCS 系统延迟的分析与改善. 计算机系统应用, 2018, 27(11): 115–119. [doi: 10.15888/j.cnki.csa.006606]
- 19 谭胜均, 张洋, 叶民友, 等. EAST 上由垂直不稳定性引发破裂的分析与预测. 核聚变与等离子体物理, 2019, 39(2): 104–111. [doi: 10.16568/j.0254-6086.201902002]
- 20 柴文婷, 肖炳甲, 袁旗平, 等. EAST 破裂炮的判定与波形显示. 计算机系统应用, 2018, 27(5): 156–160. [doi: 10.15888/j.cnki.csa.006333]
- 21 Zheng YY, Zhang ZH, Li Q, *et al.* Design of an energetic particle radiation diagnostic spectroscopy system based on national core chips and Qt on Linux in EAST. *Nuclear Science and Techniques*, 2021, 32(7): 68. [doi: 10.1007/s41365-021-00906-x]
- 22 Mohapatra D, Subudhi B, Daniel R. Real-time sensor fault detection in tokamak using different machine learning algorithms. *Fusion Engineering and Design*, 2020, 151: 111401. [doi: 10.1016/j.fusengdes.2019.111401]
- 23 李航. 统计学习方法. 北京: 清华大学出版社, 2012.
- 24 南东亮, 王维庆, 王海云. 基于消息队列的 LightGBM 超参数优化. 计算机工程与科学, 2019, 41(8): 1360–1365. [doi: 10.3969/j.issn.1007-130X.2019.08.004]
- 25 Cannas B, Fanni A, Murari A, *et al.* Automatic disruption classification based on manifold learning for real-time applications on JET. *Nuclear Fusion*, 2013, 53(9): 093023. [doi: 10.1088/0029-5515/53/9/093023]

(校对责编: 牛欣悦)