·电动汽车锂离子电池安全技术专题·

基于生成对抗网络的电动汽车电池数据增强和故障诊断*

李洁[1] 张震豪[1] 董亚冰[2] 陈旭迎[2]

（1.湖南大学，长沙 410082；2.河南省新融高速公路建设有限公司，洛阳 471000）

【摘要】针对电动汽车动力电池故障数据稀缺导致诊断模型泛化能力差的问题，提出了基于生成对抗网络（GAN）的数据增强方法，根据增强后的数据，利用随机森林（RF）模型结合贝叶斯优化（BO）方法设计故障诊断方案，形成GAN-RF-BO电池故障诊断框架，并在真实故障数据集上与常用的多层感知机（MLP）模型、支持向量机（SVM）模型和梯度提升决策树（GBDT）模型进行泛化能力对比，结果表明，所提出的故障诊断方案准确率较MLP模型、SVM模型和GBDT模型分别提高19.66%、19.71%及16.31%，GAN-RF-BO框架能有效利用稀缺数据诊断动力电池故障。

主题词：动力电池 数据增强 生成对抗网络 故障诊断

中图分类号：U471 **文献标识码：**A **DOI:** 10.19620/j.cnki.1000-3703.20230177

Electric Vehicle Battery Data Augmentation and Fault Diagnosis Based on Generative Adversarial Networks

Li Jie[1], Zhang Zhenhao[1], Dong Yabing[2], Chen Xuying[2]

（1. Hunan University, Changsha 410082; 2. Henan Xinrong Expressway Construction Co., Ltd., Luoyang 471000）

【Abstract】A solution was proposed to address the issue of low generalization ability of diagnostic models for electric vehicle power battery faults caused by sparse data, which utilized a data augmentation method based on Generative Adversarial Networks (GAN). According to the augmented data, a fault diagnosis scheme was designed using the Random Forest (RF) model combined with the Bayesian Optimization (BO) method to form a GAN-RF-BO battery fault diagnosis framework. The proposed fault diagnosis approach was compared with the common Multilayer Perceptron (MLP), Support Vector Machine (SVM) and Gradient Boosting Decision Tree (GBDT) model on a real fault dataset. The results show that the accuracy of the proposed method is improved by 19.66%, 19.71% and 16.31% compared to the MLP, SVM, and GBDT models respectively. The GAN-RF-BO framework can better utilize sparse data to troubleshoot problems with power batteries.

Key words: Power battery, Data augmentation, Generative Adversarial Networks (GAN), Fault diagnosis

【引用格式】李洁, 张震豪, 董亚冰, 等. 基于生成对抗网络的电动汽车电池数据增强和故障诊断[J]. 汽车技术, 2023(8): 1-6.

LI J, ZHANG Z H, DONG Y B, et al. Electric Vehicle Battery Data Augmentation and Fault Diagnosis Based on Generative Adversarial Networks[J]. Automobile Technology, 2023(8): 1-6.

1 前言

锂离子电池具有使用寿命较长、能量密度高、自放电率低等优点，是当前应用最广泛的电动汽车动力电池[1-2]。电池组电压是动力电池的关键参数，电压异常故障可能导致电池的热失控[3]，甚至引发燃烧、爆炸事故。及时、准确的电动汽车动力电池故障诊断方法可以提高电动汽车的整体安全性，增加公众对电动汽车的信赖程度[4]，进而推动传统交通向绿色低碳交通的转变。

近年来，国内外学者使用机器学习[5]、信息融合[6]等方法针对动力电池故障诊断展开了广泛的研究，取得了系列成果。Yang等[7]基于电动汽车电池组的真实运行

*基金项目：湖南省科学技术厅重点研发项目（2022SK2096）；河南省交通厅科技项目（2020G11）。

通讯作者：张震豪（1999—），男，硕士研究生，主要研究方向为新能源汽车与交通大数据，zhangzhenhao_hun@hnu.edu.cn。

数据，使用组内相关系数分析电池的端电压，对比充、放电过程中单体电池电压的排序差异，确定发生故障的单体电池，探究故障形成的原因。Hong等[8]结合天气、车辆和驾驶员信息，利用长短期记忆神经网络对电动汽车电池系统的电压进行预测，实现对动力电池安全性能的评估。刘鹏等[9]根据实车运行数据，采用快速傅里叶变换和异常系数评估方法对动力电池的电压故障进行诊断。宋哲等[10]使用主成分分析法、支持向量机模型和粒子群优化算法，对锂离子电池的健康状态进行了预测。Li等[11]将经验模态分解法和样本熵相结合，根据提取到的汽车动力电池电压信号，实现对电池故障的识别和定位。

当前的电动汽车动力电池故障诊断方法大都基于海量的车辆历史运行数据构建模型，很少有研究关注此类数据集中故障异常数据样本占比过少的现象。样本的不均衡可能影响模型的有效性和普适性[12]。针对这一问题，本文提出一种基于生成对抗网络（Generative Adversarial Networks，GAN）的电动汽车电池数据增强方法。根据增强后的数据，采用随机森林（Random Forests，RF）模型进行故障诊断，使用贝叶斯优化（Bayesian Optimization，BO）方法对模型的超参数进行优化，提出GAN-RF-BO电动汽车电池故障诊断框架，并将其与常用的故障诊断模型在真实的电池故障数据集上进行对比，验证其有效性。

2 电动汽车数据分析

本文的研究数据是上海市新能源汽车公共数据采集与监测研究中心提供的20辆纯电动汽车的真实运行数据，采样时间为2021年7月至12月，采样间隔为10 s。

2.1 电压异常故障占比

本文使用的20辆纯电动汽车的运行数据共有14 787 935条，其中电压异常故障数据有5 376条，占数据总量的0.35 ‰。20辆车中电压异常故障发生数量最多的车辆有643条故障数据，占该车辆全部数据的0.82 ‰。统计分析20辆纯电动汽车发生电压异常故障数量的分布情况，大部分车辆在采样期间发生故障的数量在100~400起范围内，如图1所示。

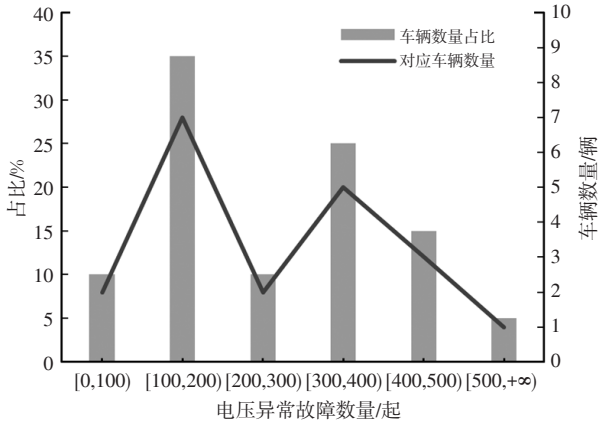


图1 电压异常故障分布

2.2 电压异常故障诊断关键因素提取

本文采用的纯电动汽车真实运行数据集包括38个数据字段，其中一些字段（如挡位、温度探针数量等）与电池电压异常故障没有直接联系，不宜引入数据增强和故障诊断模型。本文采用随机森林算法从38个特征变量中筛选出导致动力电池电压异常故障的关键因素。

随机森林模型[13]从分类回归树（Classification and Regression Tree，CART）扩展而来，由多棵决策树组成，适合进行分类和回归分析。随机森林模型在建立决策树时，利用自助重抽样方法（Bootsrap）构建数据集。

自助重抽样方法从给定的包含m个样本的数据集合D中随机选取1个样本放入采样集D_i，再把该样本放回原数据集D中。如此经过m次操作，可以得到包含m个样本的采样集D_i。每次抽样未被选择的数据称为袋外数据（Out-Of-Bag，OOB），用于对决策树的性能进行评估，以OOB作为测试集进行分类预测的错误率称为袋外数据误差。计算特征重要性时，对OOB中相应特征变量加入噪声，再次计算袋外数据误差。根据2次袋外数据误差可以计算对应特征的重要度：

$$I_i = \frac{1}{N}\sum(O_T^i - O_T) \tag{1}$$

式中，I_i为特征i的重要度；N为随机森林中决策树的总数量；O_T为决策树T的袋外数据误差；O_T^i为决策树T中特征i加入噪声干扰后计算的袋外数据误差。

特征重要度反映了特征变量与电池电压故障的关联性，重要度越高则关联性越强。对于数据增强和故障诊断模型，输入维度较低的特征变量难以提取关键信息，影响模型的鲁棒性，过高的特征变量维度又会带来冗余特征，增加模型的复杂度。筛选出关联性较强的特征变量，能够减少计算量，提高模型的精度和效率[14]。为保证数据增强和故障诊断模型的性能，本文以特征重要度0.05为阈值，从38个潜在特征变量中筛选出11个关键特征变量，即电压异常故障诊断关键影响因素，如图2所示。

3 数据增强模型

20辆纯电动汽车中动力电池电压异常故障数据占比不足0.1%，是典型的稀缺数据。针对电动汽车真实

运行数据集中动力电池电压异常故障数据的样本不均衡问题，本文使用生成对抗网络进行数据增强。

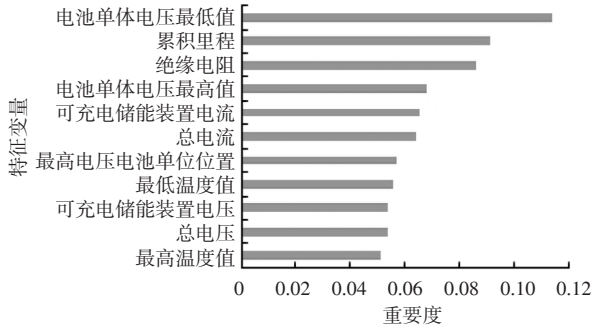


图2　关键特征变量

3.1　生成对抗网络

生成对抗网络是Goodfellow等[15]根据零和博弈思想在2014年提出的一种生成模型，已经应用于图像生成[16]、交通标志识别[17]以及交通事故检测[18]等多个领域。GAN的基本结构如图3所示，主要由生成器和判别器组成。生成器的主要作用是学习真实数据的分布，判别器根据输入的数据判断该数据是真实数据或是生成器生成的数据。根据判别结果，生成器G和判别器D不断优化，最终达到纳什均衡，即判别器D无法判断输入数据是真实数据还是生成数据。

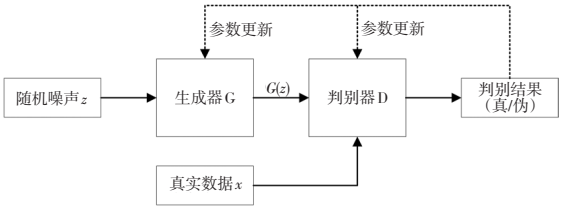


图3　GAN基本结构

GAN优化过程的目标函数为：

$$\min_G \max_D L(G,D) = E_{x\sim P_{\text{data}(x)}} \log\big(D(x)\big) + E_{z\sim P_{z(z)}} \log\big(1 - D\big(G(z)\big)\big) \tag{2}$$

式中，E为下标中指定分布的数学期望；$P_{\text{data}(x)}$为真实数据x的分布；$P_{z(z)}$为生成数据的分布；$D(x)$为判别器的判别函数；$G(z)$为生成器生成的数据。

生成器G的优化目标是使生成数据的分布无限近似于真实数据，判别器D的优化目标是最大可能地判别真实数据和生成数据。

本文中GAN的生成器和判别器均采用全连接神经网络，激活函数为修正线性单元（Rectified Linear Unit，ReLU）函数，ReLU函数能有效缓解梯度消失问题[19]。生成器和判别器使用Adam优化器[20]优化，Adam优化器具有计算高效、适用于不稳定的目标函数等优点。

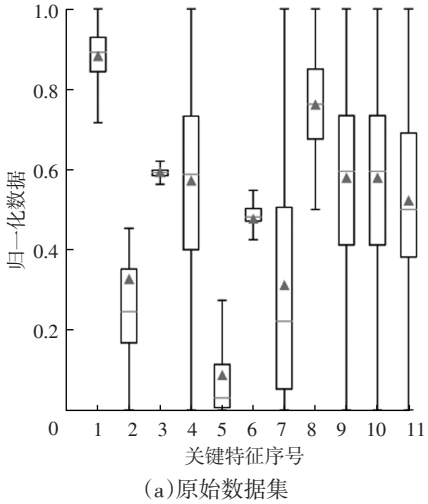
3.2　数据增强结果

本文保留由随机森林算法得到的11个关键特征数据字段。从14 787 935条车辆真实运行数据中按照不同车辆的数据占比随机抽样获取2 010条电压异常故障数据和18 090条正常运行数据，得到故障数据占比为10%的原始数据集。使用Python和Pytorch库构建生成对抗网络，输入原始数据集进行电压异常故障数据的数据增强。模型批次大小，即每次训练选取的样本数量设置为64个。生成器和判别器的学习率设置为0.01，生成器和判别器隐藏层的神经元数量设置为128个。迭代训练100 000轮，最终得到包含8 040条电压异常故障数据和12 060条正常运行数据的GAN扩充数据集。为了检验本文提出的GAN的性能，使用归一化方法处理GAN扩充数据集和原始数据集以消除量纲影响，方便比较两类数据集中数据的分布。归一化处理公式为：

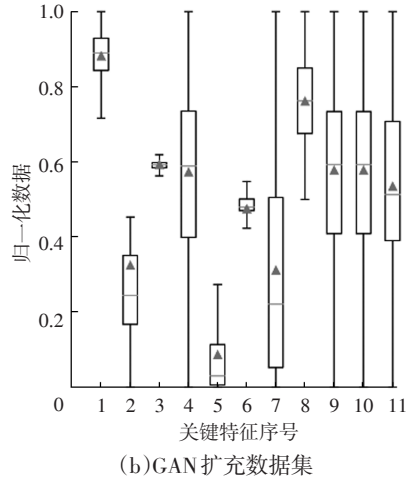
$$y_i = \frac{x_i - \min\limits_{1\leqslant j\leqslant n}\{x_j\}}{\max\limits_{1\leqslant j\leqslant n}\{x_j\} - \min\limits_{1\leqslant j\leqslant n}\{x_j\}} \tag{3}$$

式中，x_i、y_i分别为归一化处理前和处理后的数据。

归一化处理能够将数据缩放至[0,1]范围内，且不改变数据的分布情况。本文中归一化处理后的原始数据集和GAN扩充数据集的分布如图4所示。



（a）原始数据集



（b）GAN扩充数据集

图4　两类数据集中的数据分布

箱型图中从上至下的5条横线分别代表数据的最大值、上4分位数、中位数、下4分位数和最小值,三角形标识符代表数据的均值。由图4可知,GAN扩充数据集与原始数据集仅在关键特征11上存在细微差别,其余特征数据分布情况基本一致。本文提出的GAN数据增强模型能学习原始数据的分布情况,可以应用于电动汽车动力电池电压异常故障数据的数据增强。

4 故障诊断模型

根据增强后的数据,使用随机森林分类模型进行动力电池电压异常故障诊断,采用贝叶斯优化方法优化模型的超参数。

4.1 随机森林

随机森林分类模型的结构如图5所示。随机森林模型可以有效处理高维数据,对数据噪声具有较高的容忍度,且不易出现过拟合,常用于分类和回归分析。对于回归问题,随机森林模型采用平均法进行预测;对于分类问题,随机森林模型采用投票法得出最终结果。本文使用随机森林分类模型对电动汽车动力电池的电压异常进行故障诊断。

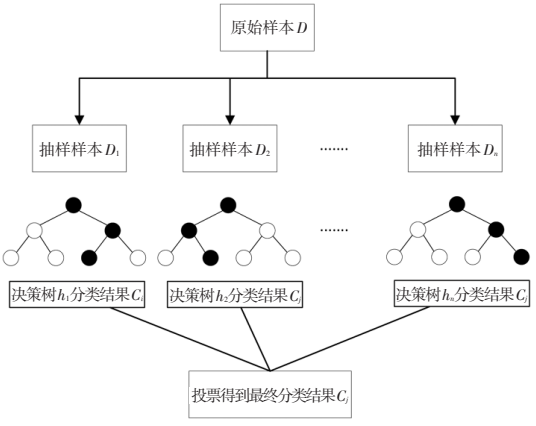


图5 随机森林基本结构

随机森林分类采用投票法获得结果,假设随机森林模型包含n个决策树模型$\{h_1,h_2,\cdots,h_n\}$,m个类别集合$\{C_1,C_2,\cdots,C_m\}$。为保证分类的可靠性,根据绝对多数投票法输出结果,即当某个类别在所有决策树中得到超过半数投票,则预测为该类别,否则拒绝分类预测:

$$H(x)=\begin{cases}C_j, & \sum_{i=1}^{n} h_i^j(x) > \frac{1}{2}\sum_{k=1}^{m}\sum_{i=1}^{n} h_i^k(x)\\ \text{reject}, & \text{其他}\end{cases} \tag{4}$$

式中,$H(x)$为投票结果;$h_i^j(x)$表示决策树h_i对输入x的类别判断为C_j;reject表示拒绝进行分类预测。

4.2 贝叶斯优化

搭建故障诊断模型时,需要对模型的超参数进行设置,但是人工设置的超参数不一定能使模型的性能达到最优,需要选择合适的模型超参数搜索方法。贝叶斯优化属于全局优化算法,是近年来机器学习领域中常用的超参数优化方法之一[21]。贝叶斯优化由概率代理模型和采集函数2个部分构成。概率代理模型用于代理复杂的未知目标函数,以最大化采集函数为目标选择下一个评估点[22]。本文使用高斯过程作为贝叶斯优化的概率代理模型,概率改进函数作为采集函数,对随机森林分类模型中的决策树的数量、决策树的最大深度、节点拆分时考虑的特征数量以及节点拆分所需的最小样本数进行超参数优化,提高故障诊断模型的性能。贝叶斯优化的流程如图6所示。

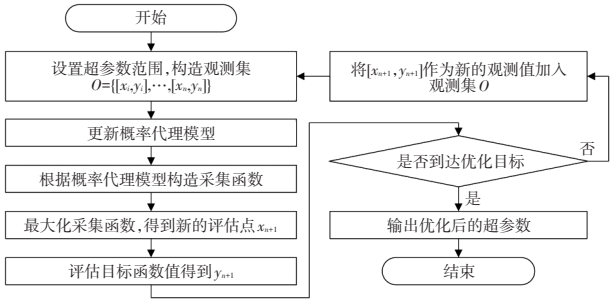


图6 贝叶斯优化流程

4.3 故障诊断结果

本文采用Python构建随机森林分类模型进行电动汽车动力电池电压异常故障诊断,分别在故障数据占比为10%的原始数据集和故障数据占比为25%的GAN扩充数据集上进行训练。使用准确率(Accuracy)作为模型的评价指标,准确率是所有诊断正确的类(包括正类和负类)样本数量占正类和负类样本总数的比例:

$$A=\frac{T_P+T_N}{P+N} \tag{5}$$

式中,A为诊断准确率;T_P为正类判定为正类的样本数量;T_N为负类判定为负类的样本数量;P为正类的样本数量;N为负类的样本数量。

训练模型时,按照6:2:2的比例将数据集划分为训练集、验证集和测试集。训练集用于训练故障诊断模型,验证集用于调试模型的超参数,测试集用于检测模型的精确度。随机森林分类模型的待优化超参数取值区间为:决策树的数量[100,1 000],决策树的最大深度[50,250],节点拆分时考虑的特征数量[1,11],节点拆分所需的最小样本数[2,8]。以最大故障诊断准确率为优化目标,经贝叶斯优化迭代150轮,最终得到优化后的超参数为:决策树的数量为456,决策树的最大深度为96,节点拆分时考虑的特征数量为3,节点拆分所需的最小样本数为2。

为了验证本文提出的RF-BO故障诊断模型的有效性，从车辆真实运行数据集（除去用于训练和数据增强的原始数据集）中按比例抽样得到真实故障数据集，此数据集包含2 010条故障数据和6 030条正常运行数据，故障数据占比为25%，用于评价故障诊断模型的泛化能力。选择故障诊断领域常用的多层感知机（Multilayer Perceptron，MLP）模型、支持向量机（Support Vector Machine，SVM）模型和梯度提升决策树（Gradient Boosting Decision Tree，GBDT）模型，在同样的数据集上训练，在同一个真实故障数据集上对比各模型的泛化性能，结果如图7所示。

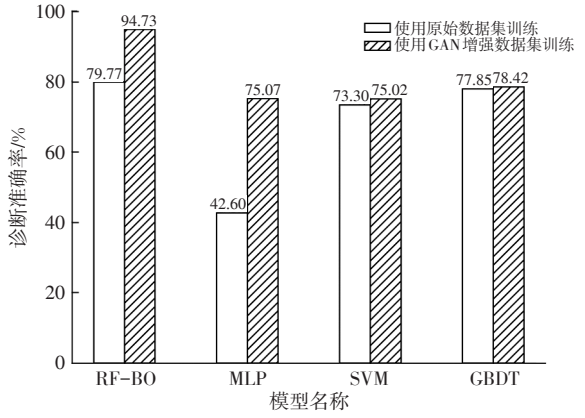


图7　模型泛化性能对比

图7中，RF-BO模型使用GAN扩充数据集训练后，故障诊断准确率可达94.73%，对比使用原始数据集训练的模型，诊断准确率提升14.96百分点。MLP模型、SVM模型、GBDT模型使用GAN扩充数据集训练后，在真实故障集上验证的准确率均有不同程度的提升，证明采用GAN增强动力电池电压异常故障稀缺数据可以提高故障诊断模型的准确性。RF-BO的模型诊断准确率较MLP模型、SVM模型及GBDT模型分别提升19.66百分点、19.71百分点和16.31百分点，体现了GAN-RF-BO故障诊断框架的优越性。

5　结束语

本文基于20辆纯电动汽车的真实运行数据，利用生成对抗网络对动力电池故障数据进行数据增强，采用随机森林模型结合贝叶斯优化进行电池故障诊断。利用随机森林算法计算38个特征数据字段的重要度，保留了11个显著特征变量进行数据增强和故障诊断，降低了模型的复杂度，缩短了运行时间。针对数据集的数据不均衡现象，提出使用生成对抗网络对动力电池电压异常故障数据进行数据增强。将扩充后的数据集与原始数据集进行比较，数据分布基本一致，说明GAN数据增强方法能够对稀缺数据样本进行有效扩充，提高故障诊断模型的准确率。

使用随机森林模型进行故障诊断，通过贝叶斯优化得到最优超参数组合。对比验证结果表明，本文提出的RF-BO模型在诊断准确率和泛化性能上明显优于MLP模型、SVM模型和GBDT模型。

动力电池故障与驾驶行为、天气条件、道路工况等多方面因素相关，且不同类型故障之间存在一定联系，本文仅根据实车运行数据对动力电池电压异常故障展开分析，未来将重点探究不同类型电池故障之间的耦合关系，并结合驾驶行为、天气信息等多源数据，深入研究电动汽车电池故障的形成机理，提高驾驶安全性与交通安全性。

参　考　文　献

[1] DUH Y S, TSAI M T, KAO C S. Characterization on the Thermal Runaway of Commercial 18650 Lithium- Ion Batteries Used in Electric Vehicle[J]. Journal of Thermal Analysis and Calorimetry, 2017, 127(1): 983-993.

[2] ZHANG J N, ZHANG L, SUN F C, et al. An Overview on Thermal Safety Issues of Lithium-Ion Batteries for Electric Vehicle Application[J]. IEEE Access, 2018, 6: 23848-23863.

[3] YE J N, CHEN H D, WANG Q S, et al. Thermal Behavior and Failure Mechanism of Lithium Ion Cells during Overcharge under Adiabatic Conditions[J]. Applied Energy, 2016, 182: 464-474.

[4] 孙正良，江帆，虞力英. 新能源汽车发展对城市交通管理的影响[J]. 交通信息与安全，2016，34(6)：108-113.
SUN Z L, JIANG F, YU L Y. Influences of Development of New Energy Vehicles on Management of Urban Traffic[J]. Journal of Transport Information and Safety, 2016, 34(6): 108-113.

[5] ZHAO Y, LIU P, WANG Z P, et al. Fault and Defect Diagnosis of Battery for Electric Vehicles Based on Big Data Analysis Methods[J]. Applied Energy, 2017, 207: 354-362.

[6] ODENDAAL H M, JONES T. Actuator Fault Detection and Isolation: An Optimised Parity Space Approach[J]. Control Engineering Practice, 2014, 26: 222-232

[7] YANG J, JUNG J, GHORBANPOUR S, et al. Data-Driven Fault Diagnosis and Cause Analysis of Battery Pack with Real Data[J]. Energies, 2022, 15(5).

[8] HONG J C, WANG Z P, YAO Y T. Fault Prognosis of Battery System Based on Accurate Voltage Abnormity Prognosis Using Long Short-Term Memory Neural Networks[J]. Applied Energy, 2019, 251.

[9] 刘鹏，吴志强，张照生，等. 基于电压频域特征和异常系数的动力电池故障诊断方法[J]. 中国公路学报，2022，35(8):

89-104.
LIU P, WU Z Q, ZHANG Z S, et al. Fault Diagnosis for Battery Systems Based on Voltage Frequency- Domain Indicator and Abnormal Coefficient[J]. China Journal of Highway and Transport, 2022, 35(8): 89-104.

[10] 宋哲, 高建平, 潘龙帅, 等. 基于主成分分析和改进支持向量机的锂离子电池健康状态预测[J]. 汽车技术, 2020(11): 21-27
SONG Z, GAO J P, PAN L S, et al. Prediction for the State of Health of Lithium- Ion Batteries Based on PCA and Improved SVR[J]. Automobile Technology, 2020(11): 21-27.

[11] LI X Y, DAI K W, WANG Z P, et al. Lithium-Ion Batteries Fault Diagnostic for Electric Vehicles Using Sample Entropy Analysis Method[J]. Journal of Energy Storage, 2020, 27(5).

[12] HE H B, GARCIA E A. Learning from Imbalanced Data[J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(9): 1263-1284.

[13] BREIMAN L. Random Forests[J]. Machine Learning, 2001, 45(1): 5-32.

[14] YU K, GUO X J, LIU L, et al. Causality- Based Feature Selection: Methods and Evaluations[J]. ACM Computing Surveys, 2020, 53(5): 1-36.

[15] GOODFELLOW I J, POUGET- ABADIE J, MIRZA M. Generative Adversarial Nets[C]// 27th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press, 2014: 2672-2680.

[16] 陈佛计, 朱枫, 吴清潇, 等. 生成对抗网络及其在图像生成中的应用研究综述[J]. 计算机学报, 2021, 44(2): 347-369.
CHEN F J, ZHU F, WU Q X, et al. A Survey Image Generation with Generative Adversarial Nets[J]. Chinese Journal of Computers, 2021, 44(2): 347-369.

[17] 高忠文, 于立国. 基于生成对抗网络改进的更快速区域卷积神经网络交通标志检测[J]. 汽车技术, 2020(7): 14-18.
GAO Z W, YU L G. Improved Faster R-CNN Traffic Sign Detection Based on Generative Adversarial Network[J]. Automobile Technology, 2020(7): 14-18.

[18] LIN Y, LI L C, JING H L, et al. Automated Traffic Incident Detection with a Smaller Dataset Based on Generative Adversarial Networks[J]. Accident Analysis and Prevention, 2020, 144.

[19] 蒋昂波, 王维维. ReLU 激活函数优化研究[J]. 传感器与微系统, 2018, 37(2): 50-52.
JIANG A B, WANG W W. Research on Optimization of ReLU Activation Function[J]. Transducer and Microsystem Technologies, 2018, 37(2): 50-52.

[20] KINGMA D, BA J. Adam: A Method for Stochastic Optimization[EB/OL]. (2017-01-30)[2023-04-20]. https://arxiv.org/abs/1412.6980.

[21] GHAHRAMANI Z. Probabilistic Machine Learning and Artificial Intelligence[J]. Nature, 2015, 521: 452-459.

[22] 崔佳旭, 杨博. 贝叶斯优化方法和应用综述[J]. 软件学报, 2018, 29(10): 3068-3090.
CUI J X, YANG B. Survey on Bayesian Optimization Methodology and Applications[J]. Journal of Software, 2018, 29(10): 3068-3090.

（责任编辑　斛　畔）

修改稿收到日期为2023年4月20日。