

随机森林模型电池热失控预警研究

周昊 魏瑜欣 周敏春

安徽江淮汽车集团股份有限公司

摘要: 新能源汽车热失控问题正受到各方关注。本文通过模型逻辑、模型初始化等研究，证实随机森林模型在电池热失控预警问题上的可行性和适用性。随机森林模型通过随机生成的多个决策树组合进行预测的，决策树的随机性降低了样本数据处理难度，减少了过拟合风险，提高了模型泛化能力。

关键词: 电池热失控；随机森林模型；样本数据

0 研究背景

在新能源汽车动力电池安全事故中，热失控是影响动力电池安全的最大诱因。实现电池热失控预警功能，其难点在于电池热失控预警模型搭建及调优。建立热失控预警模型，重点是基于机器学习算法对电池进行风险评估，依据不同指标因素及诊断方式，对电池可能发生热失控的程度进行判断，输出热失控风险评估结果。随机森林模型就是典型的机器学习算法，通过对训练数据进行采样和调整权重来平衡正负样本量之间的不平衡问题，具有较高的预测准确性，适合电池热失控预警情景。

1 随机森林模型预测逻辑

1.1 核心思想

随机森林模型是由众多决策树组成的一种集成学习方法。如图 1 所示，该模型将若干个弱分类器的分类结果进行投票选择，从而组成一个强分类器。

1.2 样本集

(1) 不平衡数据处理

由于随机森林模型算法旨在解决正负样本不平衡类问题，而此类问题数据常常是极度不平衡且正负样本比例大，影响模型训练效果，因此需要对这些不平衡的样本数据进行处理。不平衡样本数据处理方法主要是随机过采样、信息性过采样和随机欠采样。针对电池热失控预警问题，由于车辆电池数据为定时上报，正负样本量均较多，只是比例不平衡，但热失控对车的安全影响较大，故不采用随机过采样和信息性过采样方法，所以最终应当采用随机欠采样方法，通过随

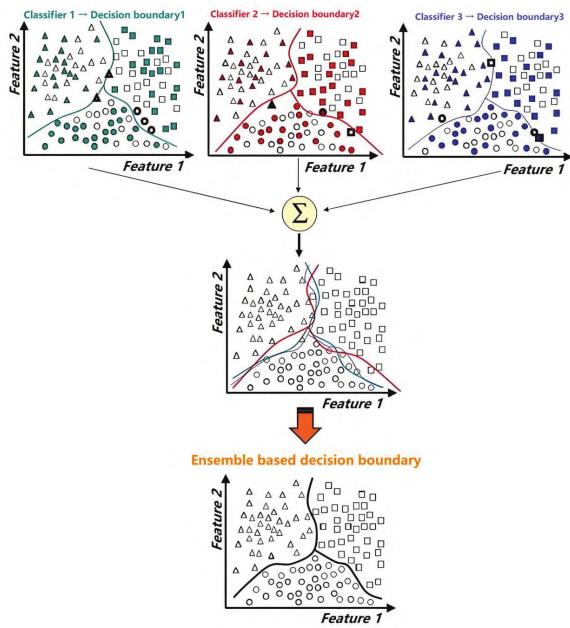


图 1 随机森林模型的核心思想

机不重复的从多数类样本中抽取适当数量样本，将其与少数类样本相结合，生成新的平衡数据集。

(2) Bootstrap 采样

Bootstrap 采样是在平衡后的样本集中，随机且有放回地从训练集中抽取 N 个训练样本，作为决策树的训练集。这样森林中的每一棵决策树的训练集都不相同，而且里面包含重复的训练样本，这样可以使得每棵决策树的分类结果存在一定的相关性但又不会完全相同。

1.3 决策树

(1) 决策树简介

决策树归纳的基本算法是贪心算法，自上向下递归方式构造决策树，如图 2 所示。基本思想是一种二分递

归分割方法，在计算过程中充分利用二叉树，在一定的分割规则下将当前样本集分割为两个子样本集，使得生成的决策树的每个非叶节点都有两个分裂，这个过程又在子样本集上重复进行，直至无法再分成叶节点为止。

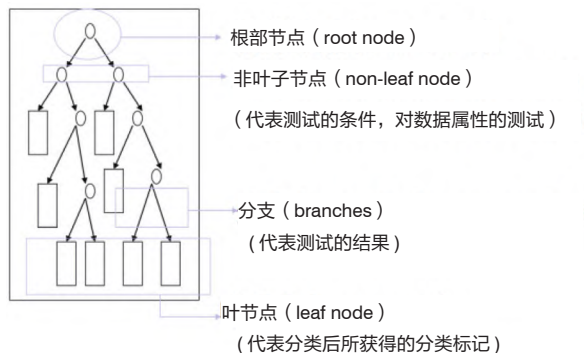


图2 决策树结构

基于不纯度的减少来作为分裂准则，即通过最小化节点不纯度的减少来确定最优分裂变量和最优分裂点。其中，衡量节点不纯度的指标有：

$$\text{错分率: } \frac{1}{n_m} \sum_{x_i \in R_m} I(y_i \neq k(m)) = 1 - \hat{p}_{m,k(m)}$$

$$\text{Gini指数: } \sum_{k \neq k'} \hat{p}_{m,k} \hat{p}_{m,k'} = \sum_{k=1}^K \hat{p}_{m,k} (1 - \hat{p}_{m,k})$$

$$\text{熵: } - \sum_{k=1}^K \hat{p}_{m,k} \log \hat{p}_{m,k}$$

（2）随机森林模型唯一参数

如果每个样本的特征维度为 M ，设定常量 m ，即每棵决策树的特征维度均是随机地从 M 个特征中选取 m 个特征的子集。所以随机森林的关键问题在于如何选择最优的 m 。减小特征选择个数 m ，树的相关性和分类能力也会相应地降低；增大 m ，两者也会随之增大。选择最优的 m ，主要依据计算误分率 oob error。但针对热失控分类问题，一般默认使用 $m = \text{pow}(M, 2)$ 。

由于使用 Bootstrap 方法抽取训练样本并随机选取 m 个特征维度构建决策树，使得随机森林不容易陷入过拟合，所以随机森林中每棵树都可以最大限度地生长，并且没有剪枝过程。

2 随机森林模型初始化

2.1 训练集准备

针对热失控预警问题，基于电池五个主要参数：电压、温度、电流、电阻和 SOC 进行衍生，确定样本的特征维度。基于故障车辆和正常车辆上报的电池数

据，计算包含特征维度带标签的数据集并对异常值进行数据清洗，然后按照一定比例划分成训练样本集和测试样本集。

对于不平衡的训练样本集，采用随机欠采样方法，通过随机不重复地从正常车辆样本中抽取适当数量样本，将其与故障车辆样本相结合，生成新的平衡训练数据集。

对平衡的训练数据集，按照项目资源情况，设定合适的训练集样本量 N 和决策树量 k 。采用 Bootstrap 方法重复 k 次，随机且有放回地从训练集中抽取 N 个训练样本，得到 k 个样本量为 N 的训练集。

2.2 决策树构建及集成

从样本所有的 M 个特征维度中，随机选取 m 个特征子集，作为一颗决策树的节点。使用 m 个节点和样本量为 N 的训练集，创建决策树。重复 k 次上述步骤，得到 k 个决策树，组合成一个随机森林模型。

2.3 测试样本预测

对测试样本进行预测，每个决策树都得到一个预测结果，最后综合少数服从多数，得到最终的决策结果。将全部测试样本的决策结果与样本标签进行对比，计算得到该随机森林模型的混淆矩阵，如表所示。根据矩阵中的假阳性率（FPR）和真阳性率（TPR），绘制 ROC 曲线，如图 3 所示，采用 ROC 曲线下方的面积（area under the ROC curve, AUC）来评判随机森林模型的性能。

表 随机森林模型混淆矩阵

		预测分类		
		- 或零	+ 或非零	总计
真实分类	- 或零	真阴性值 (TN)	假阳性值 (FP)	N
	+ 或非零	假阴性值 (FN)	真阳性值 (TP)	P
	总计	N*	P*	

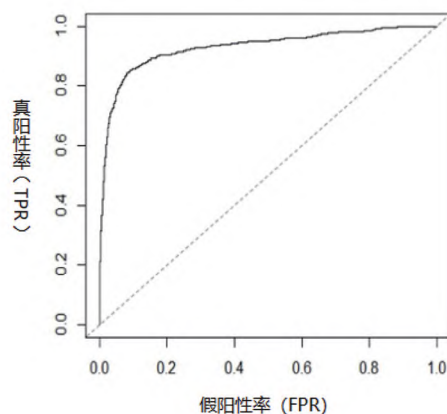


图3 ROC曲线

（下转第 33 页）

5 结语

综上，增程式变速器电机选型所需基本参数初步确认完成，需要基于初版参数同电机厂商交流电机电磁开发方案，输出电机详细参数特性，同时收集确认用于 Cruise 和 Simulink 仿真的其他系统相关零部件具体参数，共同搭建动力总成系统零部件数学模型和整车动力学模型验证零部件设计，寻优整车控制策略，指导整车能量管理标定。同步需要考虑最高持续车速、亏电最高持续车速以及高速爬坡等工况下整车热管理的散热性能边界是否达标。

总之，增程式变速器的开发，不仅要通过一系列整车动力性边界计算确定电机参数，还需要考虑满足电机硬件设计边界下的控制策略、能量管理策略以及热管理边界，通过多轮仿真寻求最优解，最终确认增

程式电动汽车驱动电机的选型。

参考文献

- [1] . 基于 MATLAB 的新能源汽车仿真实例. 北京: 化学工业出版社, 2020.4
- [2] 赵航, 史广奎. 混合动力电动汽车技术. 北京: 机械工业出版社, 2012.5
- [3] 宋柯, 章桐. 增程式纯电驱动汽车动力系统研究 [J]. 汽车技术, 2011 (7): 14-19.
- [4] 焦磊. 增程式电动汽车动力总成参数匹配及控制策略研究 [D]. 西安: 长安大学, 2013.
- [5] Li Z H, Khajepour A, Song J H. A comprehensive review of the key technologies for pure electric vehicles [J]. Energy, 2019, 182: 824-839.

(上接第 19 页)

3 结语

随机森林模型通过组合多个决策树的预测结果，可以得到较高的准确性。由于每个决策树都是基于不同的随机样本和随机特征子集构建的，随机森林可以减少过拟合风险，提高模型的泛化能力。由于其随机性，模型对样本数据的处理要求低，可不进行特征归一化和缺失值处理。每个决策树的特征维度不同，可以根据决策树的结果分析得到样本的各个特征维度对决策结果的贡献度，这样有助于进行对特征维度的选择。

随机森林模型是由多个决策树构成的，训练时间长且占用资源多，这就导致随机森林模型受资源影响较大。资源越充足，训练样本数量 N 和决策树个数 k 就可以越大，模型预测结果就更加准确。在随机森林模型的基础上，还可以采用树提升算法，如 Adaboost、

GBDT 和 xgboost 等，通过分析集成的随机森林模型对训练样本的判错情况，优化随机森林模型。

参考文献

- [1] 方匡南, 吴见彬, 朱建平, 等. 随机森林方法研究综述 [J]. 统计与信息论坛, 2011, 26(3): 7. DOI: 10.3969/j.issn.1007-3116.2011.03.006.
- [2] 叶玲, 张永军. 一种基于随机森林回归预测算法的路灯智能节能方法: CN201610922265.6 [P]. CN107979900A [2023-07-21].
- [3] 袁博, 刘石, 姜连勋, 等. 基于随机森林回归算法的住房租金预测模型 [J]. 电脑编程技巧与维护, 2020(1): 3. DOI: CNKI: SUN: DNBC.0.2020-01-009.
- [4] 王仁超, 朱品光. 基于随机森林回归方法的爆破块度预测模型研究 [J]. 水力发电学报, 2020(1): 13. DOI: 10.11660/slfdbx.20200110.