DOI:10.20033/j.1003-7241.(2025)09-0039-05

基于改进梯度提升决策树的配电网信息化评审中心故障分类方法

张艳春，李 阳，王世君，邱湘可

（国网冀北电力有限公司承德供电公司，河北 承德 067000）

摘 要：为提升配电网故障分类的准确性和效率，研究提出一种改进的梯度提升决策树模型。该模型通过实时电流采样和经验模态分解技术，结合最小相关熵准则进行特征筛选，有效提高了故障检测的性能。在UCI机器学习库的4个数据集上，改进模型相较于传统梯度提升决策树模型实现了平均准确率约2.72%的提升，且和其他模型进行对比时，在F_1分数、精确率和召回率上分别达到了90.9%、93.8%和91.2%。研究结果表明，该方法在故障定位和分类任务中表现出色，验证了其在配电网故障管理中的应用潜力。

关键词：GBDT；配电网；故障分类；故障定位

中图分类号：TP183；TP277.2 **文献标识码：**A **文章编号：**1003-7241(2025)09-0039-05

Fault Classification Method of Distribution Network Information Review Center Based on Improved Gradient Boosting Decision Tree

ZHANG Yanchun, LI Yang, WANG Shijun, QIU Xiangke

(Chengde Power Supply Company, State Grid Hebei Electric Power Co., Ltd., Chengde 067000, China)

Abstract: In order to improve the accuracy and efficiency of distribution network fault classification, an improved gradient boosting decision tree model is proposed. This model effectively improves the performance of fault detection through real-time current sampling and empirical mode decomposition technology, combined with the minimum correlation entropy criterion for feature screening. On the four data sets of the UCI machine learning library, the improved model achieves an average accuracy improvement of approximately 2.72% compared to the traditional gradient boosting decision tree model, and when compares with other models, it achieves better results in F_1 score, accuracy and the recall rates reaches 90.9%, 93.8% and 91.2% respectively. The research results show that the method performs well in fault location and classification tasks, verifying its application potential in distribution network fault management.

Keywords: gradient boosting decision tree; distribution network; fault classification; fault location

0 引言

随着智能电网技术的发展，配电网信息化评审中心在保障电力系统稳定运行中扮演着越来越重要的角色[1]。李铁柱等人采用数据驱动方法对配电网故障定位展开研究，提高了故障检测的准确性和效率[2]。张大波等提出一种基于边缘计算和深度学习的单相接地故障区段定位方法，得到离线训练好的定位模型并储存在边缘节点以实现快速故障定位[3]。Fang等学者提出一种基于粒子群优化和极限梯度Boosting树算法的配电网故障风险预测方法，缓解了测试精度低的问题[4]。但目前还存在配电网故障分类中，处理大规模、高维度和动态变化的数据捕捉复杂数据模式能力较低等问题。同时，这些方法的性能和泛化能力还有提升空间。因此，研究提出一种基于改进梯度提升决策树（gradient boosting decision tree，GBDT）的配电网故障分类方法。该方法不仅继承了GBDT在处理结构化数据上的优势，还通过引入增量学习机制，增强了模型对动态数据流的适应能力和学习效率。同时，研究创新性地结合了最小相关熵准则进行特征选择，进一步提升了模型在故障识别和分类任务中的性能。期望能为配电网故障分类提供新的解决方案。

1 配电网故障分类的集成学习模型研究与应用

1.1 改进的增量式梯度提升决策树在配电网故障分类中的应用

在配电网信息化评审中心的故障分类领域，研究提出了一种基于改进梯度提升决策树的方法。这种方法利用了梯度提升机（gradient boosting machine，GBM），一种通过集成多个弱模型来构建强模型的集成学习技术[5]。尽管GBM可以采用多种算法作为其基础模型，但由于决策树在解释性和预测速度方面的优势，它们常被作为GBM

*基金项目：国网承德供电公司科技项目（B30106230006）

收稿日期：2024-04-25

的首选弱学习器，形成了GBDT[6]。然而，GBDT在处理大规模数据集和动态数据流方面存在一定的局限性，因为它不具备神经网络的增量学习能力[7]。为了提高配电网故障分类的效率和准确性，研究参考卷积神经网络的灵活性和训练机制，开发了一种能够适应增量学习的GBDT模型，模型如图1所示。

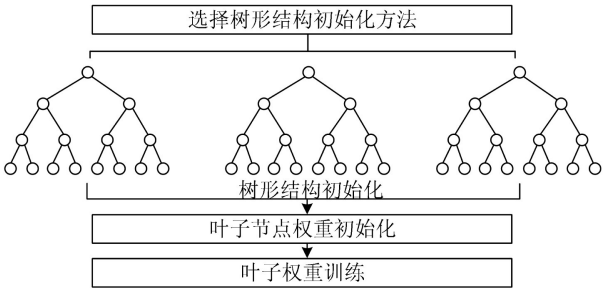


图1 可增量式学习的GBDT模型

如图1所示，改进的增量学习GBDT模型的建立分为初始化树结构和模型训练两个关键步骤。在初始阶段，用户需从五种预设树结构中选取一个，包括随机法、预训练法和中值法等。确定结构后，系统将自动设定决策树叶节点。随后，在训练阶段，模型会通过梯度下降的反向传播方法进行参数优化[8]。在构建决策树模型的过程中，一种关键的初始化技术是基于信息增益原则的。该方法的核心在于将先验知识整合入树结构的构建初期，以期提升模型的性能和解释性。信息熵作为量化系统不确定性的度量，为评估信息的混乱程度提供了一个客观标准。相应地，信息增益则作为衡量两个概率分布差异的指标，其在机器学习领域内的应用，体现在对分类任务贡献度的评估上。在面对一个包含k个类别的训练数据集时，对于数据集中的某一特征a_i，信息增益计算如式(1)所示。

$$H(C)=-\sum_i^k p(c_i)\log p(c_i) \tag{1}$$

式中，$H(C)$表示数据集C的信息熵，$p(c_i)$是第i个类别在数据集中出现的概率。$H(C|T)$表示计算划分出T数据集后的信息熵，具体如式(2)所示。

$$H(C|T)=p(a_i)H(C|a_i)+p(a_i)H(C|a_i) \tag{2}$$

式中，a_i表示新第i个类别在划分出的新数据集中出现的概率。$H(C|a_i)=-\sum_j^k p(c_j|a_i)\log p(c_j|a_i)$，其中，$p(c_j|a_i)$代表当$a_i$存在的情况下，数据集属于类别$c_j$的概率。$a_i$在未出现的某一条件下，数据集样本不属于某类别时，如式(3)所示。

$$IG(a_i)=H(C)-H(C|T) \tag{3}$$

在决策树构建算法中，基于信息增益原则的初始化策略要求首先对数据集中的每个特征值进行信息增益的量化评估[9]。信息增益的计算基于特征值对数据集分类不确定性的减少程度。完成这一评估后，算法将根据计算得出的信息增益值对特征值进行降序排列，以便在树的构建过程中优先考虑具有较高信息增益的特征值。在算法实施者明确指定树的最大深度参数后，可以进一步推导出树结构中非叶子节点的数量[10]。对于具有深度depth的决策树，depth是一个正整数，非叶子节点的数量如式(4)所示。

$$IN(\text{depth})=2^{\text{depth}-1} \tag{4}$$

在完成特征值的信息增益排序之后，算法将从排序列表中选取前$IN(\text{depth})$个特征，其中IN表示选取的特征数量，depth代表树的预定深度。接着，采用层次遍历的策略对树的每个非叶子节点进行初始化，这一过程遵循从根节点到叶子节点的顺序。该初始化策略虽然在计算复杂度上相对较高，但由于避免了在每个节点上寻找最佳分裂点的贪心搜索过程，因此在预处理阶段的计算成本实际上低于那些需要进行详尽搜索的算法。

1.2 基于GBDT架构的电力分配网络故障检测与分类集成模型

在电力分配网络中，故障的迅速检测与准确分类对于实现有效的故障隔离、提升电力供应的可靠性以及减少停电影响范围和持续时间具有显著的重要性[11]。随着分布式发电资源的广泛接入，传统的故障定位技术遭遇了新的制约因素[12]。现有的故障定位技术在精确性方面虽有所成就，但往往未能充分融合网络拓扑结构的复杂性和动态变化，导致在故障区域定位与分类方面的端到端集成模型尚未成熟[13]。因此，研究结合GBDT架构的集成模型，实现从配电网信息处理中心到故障区域定位以及故障类型识别的流程。在模型表示中，h_i表示第i个节点的向量表示，H则代表由所有节点向量构成的矩阵。如图2所示，模型采用二进制编码，其中0代表配电区域的正常运行状态，而1代表故障发生状态。

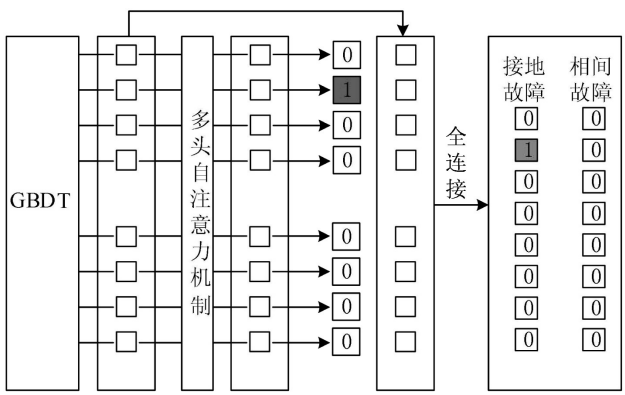


图2 配电网故障定位与识别模型

在图2所示的模型架构中，GBDT模型被选作强学习器，整合了节点的固有属性与配电网络的拓扑特性，以生

成对节点的精确数学表述。在故障定位阶段，模型部署了多头自注意力机制，该机制通过并行处理多个注意力头，能够从多个视角捕获节点间的复杂相互作用及其内在属性，进而实现对故障位置的精确识别。在故障分类任务中，模型结合了由故障定位确定的故障区域信息以及GBDT模型强学习器提供的节点数学表达，通过全连接层进一步构建故障类型识别模型，确保了故障分类的准确性和可靠性。

在模型预测的领域中，决策树结构的预测方法主要分为两类。对于第一类，即通用决策树的预测，模型启动一个迭代过程，该过程涉及对预定义树结构的遍历。在此过程中，模型采用GBDT模型，并实施并行计算，通过为每个样本和每棵决策树分配独立的线程来加速预测任务[14-15]。并行化的最高程度由线程数量决定，计算公式为$|T|\times n$，其中$|T|$代表决策树的总数，n代表待预测的样本数。模型提供了对树结构的自定义选项，使得预测过程更为高效。对于第二类，即特征遗忘决策树的推理优化，模型专门针对此类决策树的特性进行了优化。由于特征遗忘决策树在每个节点层保持一致的特征集和分裂规则，模型采用了一种高效的存储机制，通过两个数组分别存储特征标识符和相应的分裂阈值，以此代表树的每一层。在预测阶段，模型摒弃了传统的树遍历技术，转而直接访问数组中的存储信息，由于内存访问的时间复杂度为常数，提升了预测的效率。在预测环节，模型采用了一种直接数据存取策略，取代了以往的树形遍历方法。模型为每个样本在决策树的每个层级都配置了独立的线程，线程的数量根据样本的维度计算为$d\times|T|\times n$，其中d表示树的最大深度，$|T|$表示树的总数，n表示待预测样本的数量。

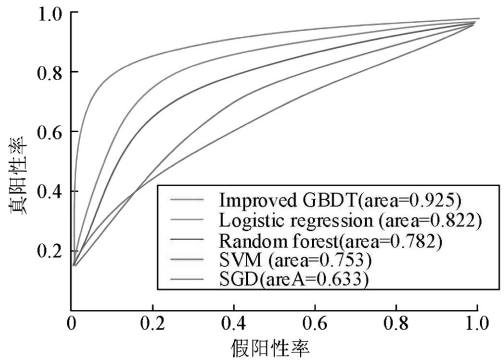
2 基于改进梯度提升决策树的配电网故障检测与分类方法验证

研究采用了Dev-C++这一集成开发环境来实现并测试改进的梯度提升决策树算法。所使用的计算机配置包括Intel Core i7系列的CPU，8 GB的内存以及运行Windows 10操作系统。实验数据选取自UCI机器学习库，从中选择了四个不同的样本数据集进行实验，它们分别是Balance Scale、Car、Adult和Chess。Balance Scale数据集包含625个样本，分为3个类别，每个样本有4个条件属性；Car数据集由1 728个样本组成，样本类别同样为4个，但每个样本包含6个条件属性；Adult数据集是最大的一个，包含15 060个样本，分为2个类别，每个样本有14个条件属性；Chess数据集则包含3 196个样本，2个类别，以及最多的36个条件属性。传统GBDT模型与改进GBDT模型对比如表1所示。

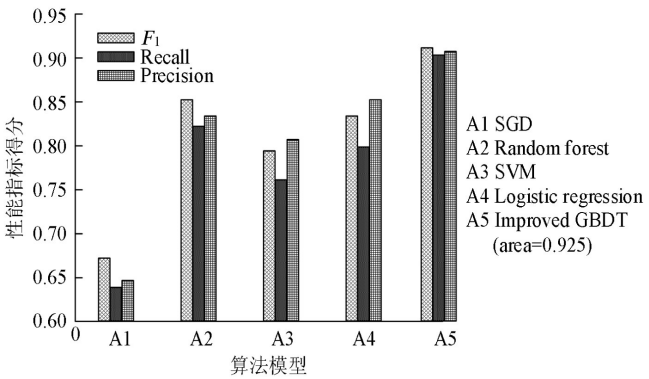
表1 传统GBDT模型与改进GBDT模型对比

数据集名称	决策树分类平均准确率/%		决策树平均叶子节点数	
	传统GBDT模型	改进GBDT模型	传统GBDT模型	改进GBDT模型
Chess	82.73	83.91	716.95	683.42
Balance Scale	74.61	76.85	163.50	120.31
Adult	82.98	85.96	861.59	826.12
Car	76.15	80.63	224.94	223.13

如表1所示，在Chess数据集上，研究提出的可增量式学习的GBDT模型分类平均准确率为83.91%，相比传统的GBDT模型提高了1.18%；在Balance Scale数据集上，传统的GBDT模型分类准确率为74.61%，而可增量式学习的GBDT模型的分类准确率为76.85%，提升了2.24%；在Adult数据集上，传统的GBDT模型决策树分类平均准确率为82.98%，可增量式学习的GBDT模型的决策树分类平均准确率是85.96%；在Car数据集上，传统的GBDT模型决策树分类平均准确率为76.15%，可增量式学习的GBDT模型的决策树分类平均准确率是80.63%。可增量式学习的GBDT模型的决策树分类平均准确率在四种数据集上平均提升了约2.72%，表现出更好的性能，说明具有良好的泛化能力，能够适应不同类型的数据集。研究还将可增量式学习的GBDT模型与新颖的算法模型进行了对比，具体结果如图3所示。



(a) 不同分类模型的ROC曲线

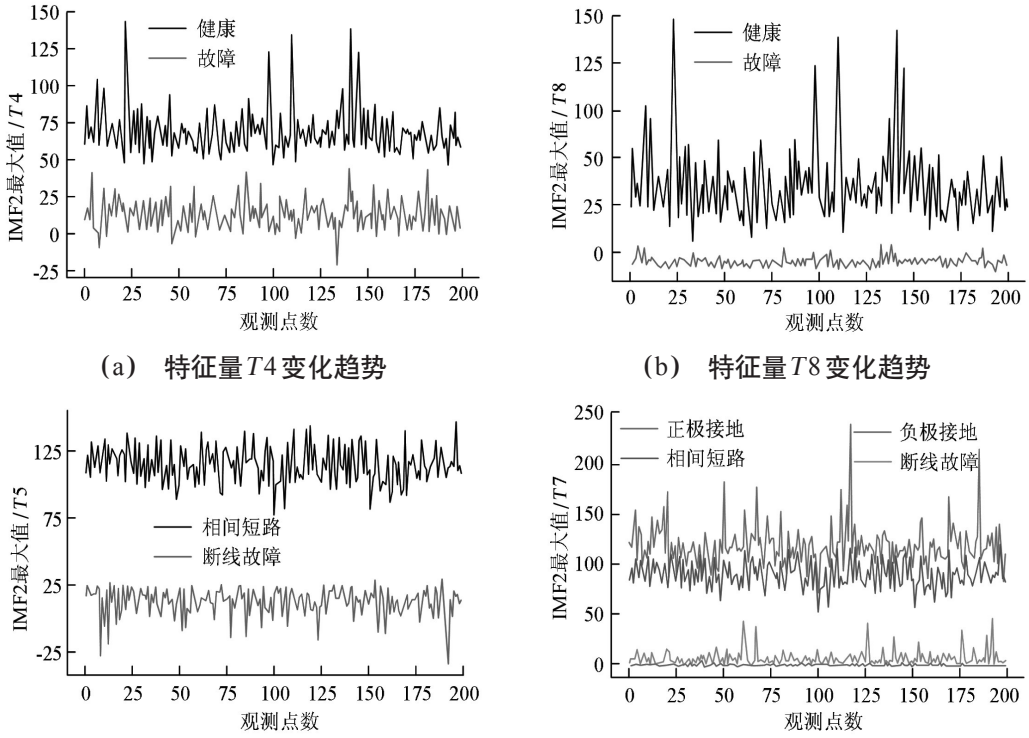


(b) 不同分类模型的性能指标得分

图3 不同分类模型的ROC曲线及不同性能指标得分

如图3(a)所示,研究提出的方法在AUC指标上达到了0.921,高于逻辑回归的0.822、随机森林的0.782、支持向量机的0.753以及随机梯度下降的0.633,在分类预测任务中性能表现良好。如图3(b)所示,在回归训练任务中,各项指标得分的排序同样突显了研究提出的方法的优势。在未实施聚类、关联规则分析及模型参数优化之前,分类预测模型的F1分数、精确率和召回率分别记录为64.5%、66.6%和67.2%。经过这些优化措施,相应的性能指标分别提升了24%、27.2%和26.4%,优化后的模型在F_1分数、精确率和召回率上分别达到了90.9%、93.8%和91.2%。

研究通过实时监控电流采样,首先应用经验模态分解技术对电流信号进行分解,进而通过最小相关熵准则筛选关键特征。在特征提取之后,利用决策树方法对故障电流进行检测和分类。为了实现故障的精确定位,研究进一步使用梯度提升决策树对故障电流特征进行分析,以区分不同的故障类型,4种故障类型为相间短路PP、正极接地PPG、负极接地NPG和断线故障DF。研究应用可增量式学习的GBDT模型对电流特征进行故障分析,选取$T1$至$T18$特征量进行模型训练。设定故障电阻(R_r)为0.5 ~ 2 Ω,故障位置(L)为0.1 ~ 3 km,故障时刻(T)在0.05 ~ 0.15 s之间变化。图4为不同特征量的变化趋势。



(a) 特征量$T4$变化趋势　(b) 特征量$T8$变化趋势

(c) 特征量$T5$变化趋势　(d) 特征量$T7$变化趋势

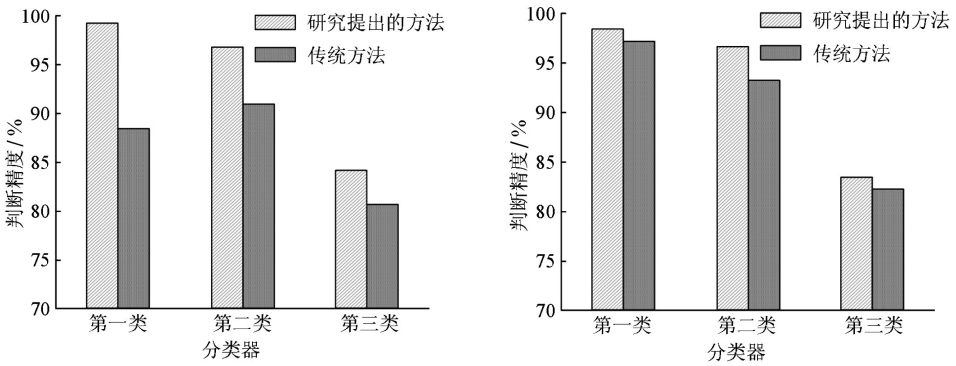
图4 不同特征量的变化趋势

图4(a)中,健康电流的$T4$特征量保持较高且稳定,反映正常运行状态;故障电流的$T4$在特定观测点显著上升,指示故障时刻。图4(b)显示健康电流的$T8$特征量维持在IMF2 0 ~ 150之间,故障电流$T8$则接近0,波动较小。图4(c)中,$T5$特征量在PP故障中IMF2值多在75 ~ 150之间,DF故障则在-25 ~ 25之间,有效区分两类故障。图4(d)通过$T7$特征量展示4种故障类型,PP和PPG故障电流归为一类,NPG和DF故障电流为另一类,模型准确捕捉故障特征差异。研究选取三类分类器进行线路故障判断精度和故障分类精度的研究,如图5所示。

图5(a)为两种方法在直流线路上的故障判断精度,在第一类分类器上,研究提出的方法故障判断精度为99.2%,相对于传统方法高出12.34%,在第二类分类器上高出6.49%,在第三类分类器上高出4.46%,证明了研究提出方法的优越性。如图5(b)所示,研究在第一类分类器上的故障分类精度为99.5%,在第二类上为96.7%,在第三类分类器上为83.3%,说明研究提出的方法在处理更为复杂或难以识别的故障类型时,仍然能够提供比传统方法更高的精度。

3 结束语

研究成功开发并验证了一种改进的GBDT模型,用于配电网故障分类。通过引入实时电流采样和经验模态分解技术,结合最小相关熵准则进行特征选择,改进的GBDT模型在UCI库的4个数据集上平均准确率提升了2.72%,AUC指标达到0.921,优于其他算法。模型在故障



(a) 直流线路故障判断　　(b) 线路故障分类

图5 线路故障判断和分类

电流特征分析中，通过$T4$和$T8$特征量有效区分健康和故障电流，并通过$T5$和$T7$特征量印证了模型对不同故障类型的分类能力，故障判断和分类进一步证明了其在配电网故障定位中的有效性和鲁棒性，为快速定位和解决配电网故障提供了新的途径。未来的研究需要在更广泛的数据集上验证模型的性能，考虑更多的实际运行条件，并探索更有效的正则化技术来防止过拟合，同时优化模型结构以适应更大规模的数据处理需求。

参考文献：

[1] 王超，张海涛，郑皓天，等. 考虑客户用电安全隐患的配电网脆弱线路风险自动预警方法[J]. 自动化技术与应用，2024，43(2)：45-48，67.

[2] 李铁柱，王伟，丁超前，等. 基于一维密集连接卷积网络的配电网故障定位研究[J]. 制造业自动化，2024，46(1)：41-44.

[3] 张大波，李雪婷，陶维青. 基于边缘计算和深度学习的有限信息配电网单相接地故障区段定位[J]. 电力系统保护与控制，2023，51(24)：22-32.

[4] FANG J，WANG H，YANG F，et al. A failure prediction method of power distribution network based on PSO and XGBoost [J]. Australian journal of electrical and electronics engineering，2022，19(4)：371-378.

[5] 赵新，芦博，王丹丹，等. 基于组件式GIS的配电网故障自动化定位系统[J]. 自动化技术与应用，2023，42(12)：146-149.

[6] 岳宝强，袁森，彭静，等. 基于自然语言处理技术的智能配电网调度优化方法研究[J]. 微型电脑应用，2023，39(11)：55-59.

[7] 吴方权，代湘蓉，刘亦驰. 基于ACNN的配电网故障识别与定位研究[J]. 微型电脑应用，2024，40(2)：149-153.

[8] 李世明，顾东健，余志文，等. 基于改进GRU模型的配电网故障线路区段检测[J]. 微型电脑应用，2023，39(10)：105-109.

[9] 胡学强，董伟锋. 一种极端天气条件下配电网预警系统设计与研究[J]. 微型电脑应用，2023，39(11)：122-125.

[10] ZEROUALI B，ZEROUALI B. A novel comprehensive framework for analyzing and assessing water quality and failure consequences based on Bayesian networks[J]. Water Environment Research，2021，93(5)：738-749.

[11] 何冰，谢天祥，胡园园，等. 基于批量匹配算法模型的配电网故障诊断方法[J]. 计算机测量与控制，2022，30(8)：19-24.

[12] 郭兴，徐武，唐文权，等. BPSO优化算法在含DG配电网中的故障诊断研究[J]. 计算机测量与控制，2021，29(3)：88-92.

[13] 姚荃，徐倩，何延强，等. 融合UPFC的隔离多环路配电系统线损最小化方法[J]. 制造业自动化，2024，46(2)：54-60，102.

[14] 李秉宇，杜旭浩，郭小凡，等. 面向新型配电系统电压支撑的微电网供电性能提升控制策略[J]. 可再生能源，2023，41(12)：1665-1674.

[15] 盛万兴，刘科研，李昭，等. 新型配电系统形态演化与安全高效运行方法综述[J]. 高电压技术，2024，50(1)：1-18.

作者简介：张艳春(1971—)，女，本科，高级工程师，研究方向：变电二次设计技术。