

[大数据]

基于大数据的用电检查异常分析方法研究

文/钱文静

国网安徽省电力有限公司巢湖市供电公司

提出一种基于大数据的用电检查异常分析方法，涵盖数据收集与预处理、异常特征检测以及模型构建与评估优化等环节。通过引入随机森林、支持向量机和时间序列分析等先进技术，在实验中验证了该方法的有效性和优越性。实验结果表明，基于大数据的方法在准确率和响应时间上均优于传统方法，为电力系统的智能化发展和用电检查工作的优化提供了新的解决方案。

随着社会经济的快速发展和科技的不断进步，电力系统在现代社会中扮演着越来越重要的角色。用电量的增加不仅推动了电力行业的蓬勃发展，同时也带来了用电检查工作中的诸多挑战。传统的用电检查方法主要依赖于人工巡检和经验判断，不仅效率低下，而且难以应对日益复杂的用电环境和多样化的用电需求。

本文研究的重点在于如何利用大数据技术对用电检查异常构建一个更有效、更优越的分析方法。本文将详细探讨数据收集与预处理、异常特征检测和异常分析模型构建等方面的内容，并通过实验研究验证所提方法的有效性，以期为电力行业的用电检查工作提供新的思路和方法，推动电力行业的智能化发展。

基于大数据的用电检查异常特征提取

1. 数据收集与预处理

在基于大数据的用电检查异常特征提取中，数据收集与预处理至关重要。本文首先从各种数据源中收集用电数据，包括智能电表、用电管理系统和传感器等数据源。收集到的数据包括用电量、电压、电流和用电时间等信息。为了确保数据的质量

和完整性，建立统一的数据收集标准和规范，明确数据的采集频率、范围和格式。

接下来，对数据进行预处理，确保数据质量。包括数据清洗、整合和转换。对用电量数据进行归一化处理，将其转换到[0, 1]的范围内，消除不同量纲和量级带来的影响，提高模型的训练效率和准确性^[1]。对于数据集中的某个特征 x ，归一化后的值 x' 通过以下归一化处理公式计算

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (1)$$

式中， $\min(x)$ 和 $\max(x)$ 分别为该特征的最小值和最大值。

此外，数据中还可能存在缺失值，本文通过插值法、均值填充等方法进行填补，确保数据的完整性，为后续的分析 and 建模提供可靠的数据基础。在均值填充方法中，将缺失值填充为该特征的均值，公式为

$$\text{填充值} = \frac{\sum_{i=1}^n x_i}{n} \quad (2)$$

式中， x_i 为该特征的非缺失值； n 为非缺失值的数量。

对于插值法，本文使用的是线性插值方法，

适用于数据点之间变化较为线性的情况。设已知数据点为 (x_0, y_0) 和 (x_1, y_1) ，需要插值的点为 x ，且 $x_0 < x < x_1$ ，插值公式为

$$y = y_0 + \frac{(y_1 - y_0)(x - x_0)}{(x_1 - x_0)} \quad (3)$$

式中， y 为插值点 x 对应的值。

由于用电数据的来源多样，数据格式和结构可能存在差异，本文还将这些数据按照一定的规则进行整合，形成一个包含各种用电信息的完整数据集，便于后续的特征提取和分析。本文使用的方法是将数据进行标准化，对每个特征进行缩放，使其均值为0，方差为1。标准化公式为

$$x' = \frac{x - \bar{x}}{s} \quad (4)$$

式中， x 是原始数据； $\bar{x}$ 是该特征的均值； s 是该特征的标准差； x' 是标准化后的数据。

此外，还对数据进行特征工程处理，提取出有用的特征，为异常特征提取提供更为直观和有效的数据支持。在特征工程处理中，计算用户的日用电量均值、用电量波动率等特征。日用电量均值的计算公式为

$$\text{日用电量均值} = \frac{\sum_{i=1}^n \text{日用电量}_i}{n} \quad (5)$$

式中，日用电量 $_i$ 是第 i 天的用电量； n 是天数。

用电量波动率的计算公式为

$$\text{用电量波动率} = \frac{\text{用电量标准差}}{\text{用电量均值}} \quad (6)$$

通过高效的数据收集和科学的预处理方法，为下面的异常特征检测和分析提供高质量的数据保障。

2. 异常特征检测

在基于大数据的用电检查中，异常特征检测是识别和处理异常用电行为的关键步骤。本文从大量的用电数据中，采用以下几种异常特征检测方法，识别出与正常用电模式显著不同的数据点，这些数据点可能代表偷电、设备故障或其他异常情况^[2]。

(1) 统计学方法

统计学方法主要基于数据的概率分布来识别异常值，使用标准差来判断数据点是否异常。在数据集 $X = \{x_1, x_2, \dots, x_n\}$ 中，其均值为 $\bar{x}$ ，标准差为 s ，异常值检测公式为

$$\text{异常值} = \{x_i \mid |x_i - \bar{x}| > k \cdot s\} \quad (7)$$

式中， k 为设定的阈值，本文取3。这种方法能够有效地识别出偏离均值较远的数据点。

(2) 机器学习方法

机器学习方法通过构建模型来识别异常特征。本文采用孤立森林这种无监督学习算法，专门用于检测异常值。因为异常值在特征空间中是稀疏的，与其他数据点的距离相对较大。通过构建随机决策树并测量每个数据点的可隔离性，有效地检测异常值。孤立森林的异常检测公式为

$$\text{异常值} = \{x_i \mid \text{pathlength}(x_i) < \text{meanpathlength} + k \cdot \text{stdpathlength}\} \quad (8)$$

式中， $\text{pathlength}(x_i)$ 是数据点 x_i 在决策树中的路径长度， meanpathlength 和 stdpathlength 分别为路径长度的均值和标准差， k 为设定的阈值。

(3) 时间序列分析方法

对于用电数据这类具有时间序列特性的数据，本文使用时间序列分析方法来检测异常特征。利用移动平均法用来平滑数据并识别异常波动。移动平均值的计算公式为

$$\text{MA}_t = \frac{1}{n} \sum_{i=1}^n x_{t-i+1} \quad (9)$$

式中， x_t 是时间点 t 的用电量； n 是移动平均的窗口大小。通过比较实际用电量与移动平均值的偏差，识别出异常波动^[3]。

基于大数据的用电检查异常分析模型构建

1. 模型的选择与训练

在构建基于大数据的用电检查异常分析模型时，基于高效的数据收集和科学的预处理后，针对数据的高维性和复杂性，本文采用随机森林和支持向量机两种机器学习算法进行模型选择。

随机森林是一种集成学习算法，通过构建多个决策树并结合它们的预测结果来提高模型的准确性和稳定性^[4]。

首先，从原始数据集中随机抽取样本，形成多个训练集。每个训练集的大小通常与原始数据集相同。设原始数据集包含 N 个样本，抽样的过程表示为

$$D_{\text{train}}^{(i)} = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\} \quad (10)$$

式中， (x_1, y_1) 表示第 i 个样本及其标签。通过这

种方式，可以生成 B 个不同的训练集， $D_{\text{train}}^{(1)}, D_{\text{train}}^{(2)}, \dots, D_{\text{train}}^{(B)}$ ，每个训练集都用于训练一棵决策树。

接下来，在每个节点的分裂过程中，随机选择一部分特征进行分裂。设原始数据集 M 个特征，随机选择的特征数量为 m ，在每个节点的分裂时，选择特征的过程表示为

$$F_{\text{split}} = \{f_1, f_2, \dots, f_m\} \quad (11)$$

式中， F_{split} 是选出的特征集合， $m < M$ 。随机特征选择有助于减少模型的相关性，提高模型的泛化能力。

在每棵决策树的构建过程本文用递归的方式描述。设在节点 t 上有 N_t 个样本，特征集合为 F_{split} ，采用信息增益这个分裂标准选择一个特征 $f_j \in F_{\text{split}}$ 进行分裂。信息增益的计算公式为

$$IG(D, f_j) = H(D) - H(D|f_j) \quad (12)$$

式中， $H(D)$ 是数据集 D 的熵， $H(D|f_j)$ 是在特征 f_j 的条件下数据集的熵。选择信息增益最大的特征进行分裂。

本文采用的支持向量机（SVM）是一种强大的分类和回归模型，最大化地将不同类别的样本分隔开来。

通过选择合适的核函数和参数，将输入样本映射到一个更高维的特征空间，使得在这个空间中样本变得线性可分^[5]。本文使用有线性核、多项式核和径向基函数（RBF）等核函数。设输入样本为 x 和 z ，核函数 $K(x, z)$ 表示为

$$K(x, z) = \phi(x)\phi(z) \quad (13)$$

式中， ϕ 是映射函数。对于径向基函数（RBF），核函数的计算公式为

$$K(x, z) = \exp(-\gamma \|x - z\|^2) \quad (14)$$

式中， γ 是核函数的参数。

然后，通过求解一个二次规划问题，找到最优的超平面和对应的支持向量。设训练数据集为 $\{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$ ，其中 $y_i \in \{-1, 1\}$ ，SVM的优化问题表示为

$$\min_{w, b, \xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^N \xi_i \quad (15)$$

$$\text{subject to } y_i(w \cdot x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, 2, \dots, N$$

式中， w 是超平面的法向量； b 是偏置项； ξ_i 是松弛变量； C 是惩罚因子。

支持向量是离决策边界最近的那些样本点。这些点满足

$$y_i(w \cdot x_i + b) = 1 \quad (16)$$

通过这些支持向量定义决策边界，使决策边界能够最大化地分隔不同类别的样本^[6]。

2. 模型评估、优化与部署应用

在构建基于大数据的用电检查异常分析模型后，本文对模型进行评估、优化和部署应用，确保模型有效性和实用性。采用的评估指标包括准确率、召回率、 $F1$ 分数和ROC曲线下面积^[7]。

准确率是模型正确预测的样本数占总样本数的比例，计算公式为

$$\text{准确率} = \frac{TP + TN}{TP + TN + FP + FN} \quad (17)$$

式中， TP 表示真正例； TN 表示真负例； FP 表示假正例； FN 表示假负例。

召回率是模型正确预测的正例数占实际正例数的比例，计算公式为

$$\text{召回率} = \frac{TP}{TP + FN} \quad (18)$$

$F1$ 分数是准确率和召回率的调和平均数，计算公式为

$$F1\text{分数} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (19)$$

式中， Precision （精确率）的计算公式为

$$\text{Precision} = \frac{TP}{TP + FP} \quad (20)$$

此外，本文采用网格搜索、随机搜索和贝叶斯优化等优化方法，通过调整模型参数和特征选择来提高模型的性能模型。

最后，将训练好的模型通过本地部署、云部署及边缘计算等方式应用到实际环境中，实现自动化检测和预警。

通过模型评估、优化和部署应用，确保了基于大数据的用电检查异常分析模型的有效性和实用性。为电力系统的安全运行提供有力支持。

实验研究

为了验证本文基于大数据的用电检查异常分析方法与传统用电检查方法的有效性和优越性。通过实验，对比两种方法在实际应用中的表现，评估大数据分析在用电检查中对异常检测的准确性、效

率、响应时间以及发现潜在问题的能力等方面的优势。实验重点关注大数据方法在数据处理能力、预警精度和实时监测方面的提升，探索在实际用电检查中的应用前景和潜力。

实验中，本文选择了5个数据集D1、D2、D3、D4、D5作为测试对象，这些数据集包含了大数据方法和传统方法的各种数据。对两种不同检查方法在准确率与响应时间上的实验数据进行比对。实验结果见表1，表格列出了本文大数据方法F1与传统方法F2在准确率与响应时间上的不同结果。

表1 不同检查方法的数据

数据集	检查方法	准确率 (%)	响应时间/s
D1	F1	92	5
	F2	85	15
D2	F1	91	5
	F2	83	14
D3	F1	93	4
	F2	84	16
D4	F1	94	4
	F2	86	15
D5	F1	90	6
	F2	82	14

从表1可以看出，基于大数据的方法准确率范围为90%~94%，平均值为92.2%。而传统方法的准确率范围为82%~86%，平均值为84.3%。同样，在响应时间的比对中大数据方法也明显优于传统方法。实验结果验证了本文的大数据方法的有效性和优越性，为在用电检查中的推广应用提供了有力支持。

结束语

本文提出的基于大数据的用电检查异常分析方法，重点解决了传统方法在复杂用电环境中的低效

问题。通过实验验证，大数据方法在提高准确率、缩短响应时间和增强实时检测能力方面具有显著优势。这不仅为电力系统的安全运行提供了技术支持，也为大数据技术在电力领域的深入应用奠定了基础。未来，随着数据规模的持续增长和算法的不断优化，基于大数据的用电检查方法将进一步拓展应用范围，并为电力行业的智能化升级提供更强动力。EA

(收稿日期：2025-07-25)

参考文献

- [1] 崔亚洲，曹敬立，王玉君，等．基于电力营销大数据技术的反窃电检查应用分析[J]．自动化技术与应用，2024，43(5)：131-134．
- [2] 李培旺，张洪铭，陈伊敏，等．南水北调天津干线箱涵专用多功能作业车的研制[J]．工程机械，2024(5)：15-20．
- [3] 张军，张华宇，王伟博．某大型水电站集电装置打火原因分析及优化改进方法研究[J]．大电机技术，2023(S02)：34-37．
- [4] 罗浩．大型水电工程压力管道智能化检查方法及适用条件分析[J]．大电机技术，2023(S02)：74-78．
- [5] 杨晶晶，阮国恒，杨玲，等．基于电力大数据的分布式电网异常负荷动态检测方法[J]．电网与清洁能源，2023，39(3)：17-22．
- [6] 李厚恩．基于电力大数据的用户侧数据异常检测方法研究[J]．微型电脑应用，2023，39(11)：229-231．
- [7] 聂文芳，涂海亮，刘军．大数据环境下光纤通信网络异常流量动态检测方法[J]．激光杂志，2023，44(9)：172-176．

