

基于大数据挖掘的用电异常特征自动化分析方法

董俐君, 岳恒

(国网汇通(北京)信息科技有限公司, 北京 100053)

摘要: 电异常数据中存在缺失值、异常值、噪声和数据不平衡等问题, 这些问题使得现有自动化分析方法的准确性和鲁棒性受到影响, 对此, 研究基于大数据挖掘的用电异常特征自动化分析方法。首先, 利用改进 K-means 聚类算法, 实施典型负荷特征曲线的聚类分析。通过获取大规模的负荷特征曲线数据, 可以减少数据缺失、异常值和噪声等问题, 提高数据质量。然后, 基于欧氏距离实施用户负荷曲线与典型负荷特征曲线的相似性度量, 实现用电异常用户筛选。最后, 基于模糊神经网络模型设计用电异常特征自动化分析模型, 实施用电异常特征自动化分析。实验结果表明, 设计方法的最大误判率为 1.32%, 最小误判率为 0.36%, 平均误判率为 0.52%, 整体误判率较低, 与期望输出的拟合情况较好, 与现实较为贴近, 具有一定的实用价值。

关键词: 自组织映射网络; 改进 K-means 聚类算法; 相似性度量; 用电异常特征分析

中图分类号: TP183

文献标志码: A

文章编号: 1003-7241(2025)10-0125-04

Automatic analysis method for abnormal characteristics of electricity consumption based on big data mining

DONG Lijun, YUE Heng

(State Grid Huitongjincai(Beijing) Information Technology Co., Ltd., Beijing 100053, China)

Abstract: There are problems such as missing values, outliers, noise, and imbalanced data in electrical anomaly data, which affect the accuracy and robustness of existing automatic analysis methods. Therefore, this paper studies an automated analysis method for electrical anomaly features based on big data mining. Firstly, it uses the improved K-means clustering algorithm to perform clustering analysis on typical load characteristic curves. By obtaining large-scale load characteristic curve data, problems such as missing data, outliers, and noise can be reduced, and data quality can be improved. Then, based on Euclidean distance, similarity measurement between user load curve and typical load characteristic curve is implemented to screen abnormal electricity users. Finally, based on the fuzzy neural network model, it designs an automatic analysis model for abnormal electrical characteristics and implement automatic analysis of abnormal electrical characteristics. The experimental results show that the maximum misjudgment rate of the design method is 1.32%, the minimum misjudgment rate is 0.36%, and the average misjudgment rate is 0.52%. The overall misjudgment rate is low, and it fits well with the expected output and is closer to reality, which has certain practical value.

Keywords: self-organizing mapping network; improve the K-means clustering algorithm; similarity measurement; analysis of abnormal electrical characteristics

0 引言

用电异常特征分析在保障电力系统安全稳定运行、提高用电管理智能化水平、促进能源合理利用和节能减排、推动电力行业技术创新和进步等方面具有重要意义^[1]。同时, 它还能够增强用户用电的透明度和信任度, 提升能源服务的用户体验, 促进跨学科的交流与合作。因此, 对用电异常特征分析的研究和应用具有十分重要的现实意义和价值^[2]。

国外在用电异常特征分析领域的研究起步较早, 且在某些方面具有领先优势。文献[3]中提出一种基于 GoogLeResNet3 网络模块的用户异常用电量行为检测分析方法, 模块包含全连接层、Inception 模块和残差模块。国内对用电异常特征分析的研究起步较晚, 但近年来发展迅速。

文献[4]提出了一种基于 K 均值聚类的居民用电行为特征分析方法, 该方法主要通过 K 均值聚类算法提取特征曲线实施用电行为分析。这一方法有助于提高电力服务的质量和效率。文献[5]提出了一种基于大数据的用电特征嫌疑用户分析方法, 以逆向思维为出发点, 通过提炼阈值规则搭建了用电异常行为识别模型。同时, 运用大数据技术实施了回归分析等计算, 构建了电力波动模型与行业特征模型。

以上方法的数据收集系统存在问题, 例如数据丢失、错误或不完整, 基于这些数据实施异常用电用户的判定可能会出现误判, 导致其用电异常特征分析结果容易出现错误, 因此设计一种基于大数据挖掘的用电异常特征自动化分析方法。通过使用改进的算法对典型负荷特征曲线进行聚类分析。并在指定的时间段内, 从用户负荷数据中提取相关信息。然后, 使用欧氏距离来度量每个用户负荷曲线与典型负荷特征曲线之间的相似程度。最后针对异常用

* 基金项目: 国家电网科技项目(717403219557)

收稿日期: 2024-01-03

电特征,设计了一个基于模糊神经网络的分析模型,实现了用电异常特征自动化分析。该方法通过使用大数据挖掘技术,可以处理庞大的用户负荷数据集,包括多个用户的用电行为和负荷曲线等信息。这种能力使得分析模型可以更全面、准确地捕捉异常用电特征,并提高分析结果的可靠性,这为电力服务的质量和效率提供了重要支持。

1 用电异常特征自动化分析方法设计

1.1 基于大数据挖掘的典型负荷特征曲线获取

利用大数据挖掘算法中的改进 K-means 聚类算法,实施典型负荷特征曲线的聚类分析。通过获取大规模的负荷特征曲线数据,可以减少数据缺失、异常值和噪声等问题,提高数据质量。采用自组织映射网络(SOM)对 K-means 算法的初始聚类中心选取过程实施优化,旨在提高算法的准确性。

1) 自组织映射网络(SOM)初始化

自组织映射网络(SOM)是模拟人脑信号处理特点的无监督系统人工神经网络,由输入层与输出层神经元构成数据处理阵列。输出节点与输入节点通过强度值连接,通过特定的规则调整强度值^[6]。当网络达到稳定状态时,每个领域的节点对特定输入产生相似的输出,从而实现聚类目的。将自组织映射网络输入节点连接到第 j 个输出节点的权值向量 $\tilde{\omega}_j(j=1,2,\dots,q)$ 初始化为随机数,并实施初始循环次数的设定^[7]。

2) 使用 SOM 进行初始聚类

对于各负荷特征曲线聚类对象 $\alpha_L(L=1,2,\dots,n)$,通过下式求取 α_L 与 $\tilde{\omega}_j$ 内的优胜者权值向量 $\tilde{\omega}_g$ 之间的距离为

$$\begin{cases} [\alpha_L - \tilde{\omega}_g] = \sum_{j=1}^q \|\alpha_L - \tilde{\omega}_j\| \\ \|\alpha_L - \tilde{\omega}_j\| = \sqrt{\sum_{j=1}^n (\alpha_L - \tilde{\omega}_j)^2} \end{cases} \quad (1)$$

将 $M_l(f)$ 定义为优胜者的邻域,将单元 l 定义为优胜者,使邻域内各单元的连接权值向量向负荷特征曲线聚类对象 α_i 靠拢^[8]。

通过迭代训练自组织映射网络,直至网络达到稳定状态。在该过程中网络的学习方程具体为

$$\begin{cases} \Delta \tilde{\omega}'_{ij} = \chi(f) |a_i^l - \tilde{\omega}'_{ij}| \\ \tilde{\omega}'_{ij}^* = \tilde{\omega}'_{ij} + \Delta \tilde{\omega}'_{ij} \end{cases} \quad (2)$$

式中, $\chi(f)$ 表示第 f 次的学习率,在训练次数增加时, $\chi(f)$ 会随之递减; $\Delta \tilde{\omega}'_{ij}$ 是指学习值; $\tilde{\omega}'_{ij}^*$ 指的是更新后的第 i 个输出节点与第 j 个输出节点的连接权值; ω'_{ij} 代表第 i 个输出节点与第 j 个输出节点的连接权值,其中 $j \in M_l(f)$; a_i^l 是第 L 个样本中第 i 个输入节点的输入^[9]。

根据网络输出节点的响应,对用电样本进行初始聚类,并获取聚类中心 $S = (S_1, S_2, \dots, S_l)$ 。

3) 将 SOM 的聚类中心应用于 K-means 聚类

计算所有用电样本数据与各自簇的聚类中心之间的

距离 $Z(B)$, 将每个样本分配到距离最近的类中心所在的簇。计算公式为

$$Z(B) = \min \sum_{n=1}^L \sum_{a \in b_n} v_{nm} D^2(b_n, a_m) \quad (3)$$

式中, v_{nm} 指的是二进制变量,当其值为 1 时表示第 m 个样本隶属于第 n 类,其值为 0 时表示第 m 个样本不属于第 n 类; a_m 是指类内其他样本数据; b_n 指的是样本聚类中心; $D(b_n, a_m)$ 是指样本与其所在类别聚类中心之间的距离^[10]。

根据上一步的分配结果,更新每个簇的聚类中心,直至满足收敛条件。以前一步的划分结果为依据,结合拉格朗日原理与最小二乘法实施 L 个典型负荷特征类的中心 b_n 的更新,直到满足收敛条件。更新公式为

$$b_n^* = \frac{\sum_{m=1}^N v_{nm} a_m}{\sum_{m=1}^N v_{nm}} \quad (4)$$

式中, N 指样本聚类中心的个数。

输出最终的典型负荷特征聚类结果曲线,即典型负荷特征曲线^[11]。

1.2 基于欧氏距离的用电异常用户筛选

基于欧氏距离实施用户负荷曲线与典型负荷特征曲线的相似性度量,实现用电异常用户筛选。具体步骤如下:

- 1) 实施检测时间段内的用户负荷数据抽取;
- 2) 通过式(5)实施检测时间段内用户负荷曲线和典型负荷特征曲线的相关系数为

$$\begin{cases} p = \frac{\sum_{i=1}^{96} (\zeta_i - \bar{\zeta})(\tau_i - \bar{\tau})}{\sqrt{\sum_{i=1}^{96} (\zeta_i - \bar{\zeta})^2 \sum_{i=1}^{96} (\tau_i - \bar{\tau})^2}} \\ \bar{\zeta} = \frac{1}{96} \sum_{i=1}^{96} \zeta_i \\ \bar{\tau} = \frac{1}{96} \sum_{i=1}^{96} \tau_i \end{cases} \quad (5)$$

式中, ζ_i 指的是考察日第 i 个用户负荷点; τ_i 是指典型负荷特征曲线中的第 i 个负荷点^[12];

3) 通过下式实施检测时间段内用户负荷曲线和典型负荷特征曲线之间的欧氏距离的计算公式为

$$D^* = \sqrt{\sum_{i=1}^{96} (\zeta_i - \tau_i)^2} \quad (6)$$

4) 通过下式实施匹配度计算公式为

$$\zeta = W_1 p + W_2 e^{-D^*} \quad (7)$$

式中, W_1, W_2 是指权重系数。

- 5) 观察匹配度是否大于正常用电时的警戒阈值;
- 6) 当匹配度比警戒阈值小,则该用户可以判断为用电异常用户。

1.3 基于模糊神经网络模型的用电异常特征自动化分析

基于模糊神经网络模型设计一种用电异常特征自动

化分析模型,实施用电异常特征自动化分析。该自动化分析模型的运行流程如下。

1) 构建异常用电评价指标体系

对数据实施重构,构建一套完整的异常用电评价指标体系。该体系由功率因数、电流三相不平衡率、电压三相不平衡率、合同容量比、告警指标、线损指标、负荷指标、用电量指标。

2) 样本选择

将样本数据划分为两部分,其中一部分用于训练网络,另一部分则作为测试样本,以评估经过训练模型的性能。

3) 构建模型

基于模糊神经网络模型构建用电异常特征自动化分析模型。

将功率因数、电流三相不平衡率、电压三相不平衡率、合同容量比、告警指标、线损指标、负荷指标、用电量指标作为模型输入特征。

将用户用电异常特征系数作为模型输出变量,其中用户用电异常特征系数的取值范围为[0,1]。

用电异常特征自动化分析模型的设计具体如下。

1) 第一层是输入层。

2) 第二层是模糊化层。各输入指标分别对应3个模糊度,也就是模糊集合 O_1 、 O_2 、 O_3 。则共有24个模糊度。

其中各节点的激活函数为各输入特征模糊度的隶属函数 $\gamma_o(\vartheta)$,具体为

$$\begin{cases} Q_i^2 = \gamma_o(\vartheta_1) (i = 1, 2, 3) \\ Q_j^2 = \gamma_o(\vartheta_2) (j = 4, 5, 6) \\ \vdots \\ Q_k^2 = \gamma_o(\vartheta_f) (f = 22, 23, 24) \end{cases} \quad (8)$$

式中, ϑ_1 到 ϑ_f 指的是输入特征; Q_i^2 到 Q_k^2 是指各节点的激活函数。

其中隶属函数选择高斯函数,具体为

$$\gamma_o(\vartheta) = f(\vartheta; \theta, \lambda) = e^{-\frac{1}{2}(\frac{\vartheta - \theta}{\lambda})^2} \quad (9)$$

式中, $f(\cdot)$ 指的是高斯函数; θ 是指高斯函数的中心; λ 指的是高斯函数的宽度。

在模型构建初期,利用随机函数对 θ 与 λ 实施初始化,在模型的训练中反复对其实施调整修正,直至 θ 与 λ 达到稳定状态。

3) 第三层为模糊推理层。

模糊推理层中的各节点只与网络的第二层中的一个节点相连,总共包含38个节点,从而产生了38条规则。

4) 第四层是去模糊化层。

在该层中,输出变量的模糊度划分数量等于该层的节点数。在用电异常特征自动化分析模型中,最终的评价结果只有一个,即用电异常程度。因此,模糊度划分个数为1。

5) 第五层是输出层。

这一层的作用是将去模糊化层的每个节点输出转换为输出变量的确定值。这一层的节点数只有一个,为用户用电异常特征系数。

2 实验测试

2.1 实验环境与实验过程

利用设计的基于大数据挖掘的用电异常特征自动化分析方法实施某地电网部分用户的用电异常特征自动化分析,测试设计方法的性能。

采用该地供电局近一年的数据背景下,以小时为粒度的电力负荷系数变化情况。作为实验数据集实施设计方法性能的分析与验证。实验环境如图1所示。

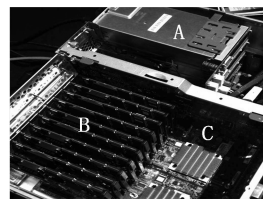


图1 实验环境

图1中,A为处理器,B为存储卡,C为显卡,其中硬件环境具体如下:

处理器: Intel Core i7-8700K;

存储卡: 512 GB SSD;

显卡: NVIDIA GeForce GTX 1080。

软件环境具体如下:

操作系统: Linux Ubuntu 18.04;

编程语言: Python 3.6;

数据处理工具: NumPy, Pandas, Scikit-learn;

神经网络库: TensorFlow, PyTorch;

数据库: MySQL。

首先使用改进K-means聚类算法实施典型负荷特征曲线挖掘。挖掘结果如图2所示。

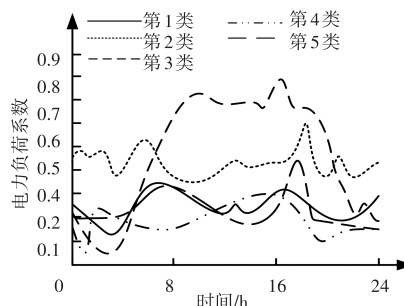


图2 挖掘结果

其中各类别的占比情况如下:

第1类: 25.36%;

第2类: 18.69%;

第3类: 8.51%;

第4类: 7.62%;

第5类: 38.91%。

接着基于欧氏距离实施用户负荷曲线与典型负荷特征曲线的相似性度量,筛选用电异常用户。在该步骤中,参数设置情况如下:

W_1 :0.5;

W_2 :0.5;

警戒阈值:0.8。

最后利用设计的自动化分析模型实施用电异常用户的用电异常特征分析。在该步骤中,将实验数据集的80%作为训练样本,其余作为验证样本。

参数的设置情况具体如下:

学习速度:0.5;

惯性系数:0.6;

单个样本最大误差:0.001;

系统最大误差:0.1;

每个网络的迭代次数:1 000次。

测试设计方法的误判率以及用电异常特征分析的样本输出拟合度。在测试中,将基于虚拟回路阻抗的低压台区异常工况分析方法与基于大数据的用电特征嫌疑用户分析方法作为对比测试方法,以获得更加明确的测试结果。在后续测试中,分别用方法1、方法2来表示。

2.2 测试结果

首先测试三种方法的误判率,也就是将正常用户判定为异常用电用户的占比,测试结果见表1。

表1 误判率测试结果

误判率	设计方法	方法1	方法2
最大误判率/%	1.32	4.36	6.25
最小误判率/%	0.36	1.69	2.36
平均误判率/%	0.52	2.65	3.64

根据上表,设计方法的最大误判率为1.32%,最小误判率为0.36%,平均误判率为0.52%,整体误判率远低于方法1、方法2,实现了异常用电用户判定上的突破。

接着测试用电异常特征分析的样本输出拟合度,测试结果如图3所示。

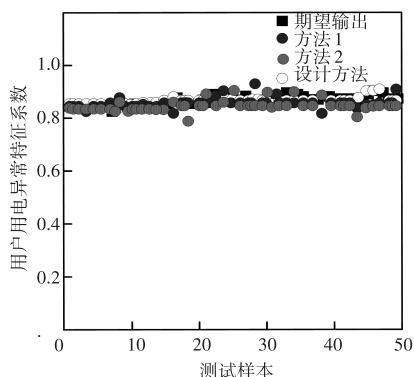


图3 样本输出拟合度

观察设计方法、方法1、方法2与期望输出的拟合情况,可以发现设计方法与期望输出的拟合情况是最佳的,而方法1、方法2与期望输出的拟合出入均较大,说明设计

方法的用电异常特征分析是最准确,与实际贴近。

3 结束语

随着电力系统的复杂性和用户需求的多样化,用电异常行为检测和分析显得尤为重要。这不仅关乎电力系统的安全稳定运行,也直接影响着电力服务的品质和效率。经过深入研究和实践,构建了一种基于大数据挖掘的用电异常特征自动化分析方法。该方法结合了改进K-means聚类算法与相似性度量实现了用电异常用户筛选,并对用电异常用户特征实施了判定分析。在实际应用中,该方法取得了显著的效果,提高了用电异常用户筛选的准确性与异常判定分析的准确性,为电力系统的管理和维护提供了有力支持。未来,将继续深入研究用电异常特征的自动化分析方法。进一步优化算法,提高异常检测的精度和速度。同时,也将关注与其他领域的交叉研究,以寻求更广泛的应用和价值。此外,将注重隐私保护和数据安全,确保在分析和利用用电数据的过程中用户隐私得到充分保护。总之,基于大数据挖掘的用电异常特征自动化分析方法为电力系统的管理和维护提供了新的思路和技术手段。在未来的研究中,将不断探索和创新,为电力行业的发展和进步做出更大的贡献。

参考文献

- [1] 孔祥玉, 马玉莹, 艾芊, 等. 新型电力系统多元用户的用电特征建模与用电负荷预测综述[J]. 电力系统自动化, 2023, 47(13): 2-17.
- [2] 赵明明, 司红星, 刘潮. 基于数据挖掘与关联分析的工控设备异常运行状态自动化检测方法分析[J]. 信息安全与通信保密, 2022(4): 1-10.
- [3] YIN L, ZHONG Q. GoogLeResNet3 network for detecting the abnormal electricity consumption behavior of users[J]. International Journal of Electrical Power & Energy Systems, 2023(145): 108733.
- [4] 葛得辉, 华定忠, 王齐, 等. 基于虚拟回路阻抗的低压台区异常工况分析方法[J]. 电力系统自动化, 2022, 46(13): 104-111.
- [5] 陈明, 张丽文, 王璐, 等. 基于大数据的用电特征嫌疑用户分析[J]. 电力大数据, 2022, 25(6): 24-35.
- [6] 郭上华, 王钢. 基于多粒度聚类和多元特征统计的低压配电网拓扑识别与监测[J]. 电力自动化设备, 2023, 43(6): 86-93.
- [7] 安磊, 童国峰, 梁皓, 等. 造纸工业电力设备节能控制方法研究[J]. 造纸科学与技术, 2022, 41(5): 18-22.
- [8] 刘昌, 何正磊, 朱小林, 等. 基于机器学习的造纸用能负荷特征日获取模型[J]. 造纸科学与技术, 2023, 42(2): 6-12.
- [9] 杨毅, 高传薪. 基于数据挖掘的电缆隧道附属设施故障自动化识别方法[J]. 自动化技术与应用, 2022, 41(12): 39-4263.
- [10] 殷浩然, 苗世洪, 韩信, 等. 基于三维卷积神经网络的配电网联网异常辨识方法[J]. 电力系统自动化, 2022, 46(1): 42-50.
- [11] 肖庆追, 李捷, 陈鹤峰, 等. 基于组合模型的电力用户用电行为分层分类方法[J]. 电力系统及其自动化学报, 2023, 35(5): 82-88.
- [12] 廖建锐. 基于特征挖掘的低压配电段用户异常用电数据挖掘方法[J]. 电气技术与经济, 2023(7): 214-216.

作者简介:董俐君(1983—),女,硕士,高级工程师,研究方向:智能用电技术、电力交费系统建设等。