DOI:10.20040/j.cnki.1000-7709.2023.20222030

基于多域特征分析与选择的电力数据识别方法

洪德华[1],刘翠玲[1],赵林燕[1],雷沁怡[1],王海鑫[2]

(1. 国网安徽省电力有限公司信息通信分公司,安徽 合肥 230061;2. 沈阳工业大学电气工程学院,辽宁 沈阳 110870)

摘要: 为解决电力数据特征挖掘不充分导致识别精度不高的问题,提出一种基于多域特征分析与选择的电力数据识别方法。首先针对现有电力数据特征提取方法存在的不足,提出一种基于经验模态分解(EMD)与Hilbert变换(EMD-Hilbert)的特征提取方法,并对电力数据的功率特征和V-I轨迹特征进行量化表征;然后基于随机森林与广义序列后向选择搜索策略相结合的特征选择算法(RF-GSBS)得到最优特征子集,并采用RF算法构建电力数据的识别模型;最后通过仿真算例验证所提方法的有效性和准确性。结果表明,该算法可利用不同特征互补性解决单一特征识别精度不高的问题,并通过特征选择进一步提高学习算法的性能。

关键词: 电力数据识别;多域特征提取;特征选择;随机森林;序列后向选择

中图分类号: TM714 **文献标志码:** A **文章编号:** 1000-7709(2023)09-0211-05

1 引言

用户电力数据的识别监测与准确分析为需求侧精细化管理提供数据支撑,是推进灵活互动智能用电的首要环节[1]。因此,对智能电表等终端设备的用能数据进行实时收集、监测识别、细粒度分析尤为重要。非侵入式负荷监测(NILM)技术为电力数据的监测识别提供了思路,但现有研究主要着眼于通过时域或频域表征电力数据,缺少对时频域特征的细粒度分析,且特征提取的有效性与冗余性还需进一步研究[2-3]。若能同时从时域、频域、时频域等多角度多电力数据进行表征量化,并利用特征选择抽取关键特征,则可为识别模型提供反映电力数据内在关联关系的深层特征。因此,本文首先通过经验模态分解(EMD)和Hilbert变换提取时频特征与频谱特征作为多域特征的一部分,并利用特征选择算法分析特征贡献度和相关性,结合不同特征组合的辨识效果,筛选出辨识结果最优的一组特征组合并作为电力数据识别的依据,最终通过算例验证所提方法的有效性与可行性,为电力数据的识别提供了新方法。

2 电力数据特征表示

针对稳态周期内电力数据,从多个角度描述和表征原始数据序列,增加负荷特征的细粒度信息量。

2.1 时频特征

EMD不需要预先定义任何基函数,即可实现对数据序列或信号的平稳化处理,是一种数据驱动的自适应方法,适合非线性、非平稳时间序列的处理,因此是一个非线性的过程。而有电器设备启动或者关闭时,其电力数据信号具有非线性和非平稳性。因此,可用EMD方法处理电力数据信号。利用EMD将目标数据分解为少量独立的、近似周期的IMF分量和1个趋势项,目标数据$x(t)$即可由IMF分量及1个趋势项表示[4],即:

$$x(t)=\sum_{s=1}^{S}c_s(t)+r(t) \tag{1}$$

式中,S为分解得到的IMF个数;$c_s(t)$为第s个IMF分量;$r(t)$为残差信号,即趋势项。

2.1.1 时域特征提取

每一个IMF分量代表一个特征尺度的数据序列,根据分解结果选取有效的IMF,进一步可提取任意阶次IMF的时域统计特征,分别为描述样本集中程度的平均值($\mu_{t,s}$)、描述样本散布程度的标准差($\sigma_{t,s}$)和描述样本分布偏离正态性的峰度($\beta_{t,s}$),其计算公式为:

$$\mu_{t,s}=\frac{1}{N}\sum_{i=1}^{N}c_s(i) \tag{2}$$

收稿日期: 2022-09-28,**修回日期:** 2022-11-18

基金项目: 国网安徽省电力有限公司科技项目(521207220002)

作者简介: 洪德华(1993-),女,硕士、工程师,研究方向为数据资产管理、数据可视化,E-mail:dehuahong0407@163.com

$$\sigma_{t,s}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}[c_s(i)-\mu_{t,s}]^2} \tag{3}$$

$$\beta_{t,s}=\frac{1}{N}\sum_{i=1}^{N}\left[\frac{c_s(i)-\mu_{t,s}}{\sigma_{t,s}}\right]^4 \tag{4}$$

式中，N 为样本点个数；$c_s(i)$为第 s 阶 IMF 的第 i 个样本点。

2.1.2 频谱特征提取

借用 Hilbert 变换的时频分析来进一步处理 EMD 分解后的信号，计算出各 IMF 分量的时频谱，以直观体现高低频信号。

(1)Hilbert 变换。对每个 IMF 分量 $c_s(t)$进行 Hilbert 变换，即：

$$H[c_s(t)]=\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{c_s(\tau)}{t-\tau}\mathrm{d}\tau \tag{5}$$

分别构造由 $c_s(t)$和 $H[c_s(t)]$共同组成的复解析信号 $Z_s(t)$[5]：

$$Z_s(t)=c_s(t)+\mathrm{j}H[c_s(t)] \tag{6}$$

计算每个 IMF 相应的瞬时幅值函数 $A_s(t)$、相位函数 $\Phi_s(t)$及瞬时频率函数 $w_s(t)$，为：

$$A_s(t)=\sqrt{c_s^2(t)+H^2[c_s(t)]} \tag{7}$$

$$\Phi_s(t)=\arctan[H[c_s(t)]/c_s(t)] \tag{8}$$

$$w_s(t)=\mathrm{d}\Phi_s(t)/\mathrm{d}t \tag{9}$$

忽略残差信号，分析信号的 Hilbert 谱可写为：

$$H(w,t)=\mathrm{Re}[\sum_{s=1}^{S}A_s(t)\mathrm{e}^{\mathrm{j}\int w_s(t)\mathrm{d}t}] \tag{10}$$

(2)特征提取。将 Hilbert 谱在时间轴上进行积分，使之从幅值—时间—频率三者间的关系转变为幅值—频率两者间的关系，即可定义信号序列的边际谱 $h(w)$为：

$$h(w)=\int_{-\infty}^{\infty}H(w,t)\mathrm{d}t \tag{11}$$

为突显信号中频率成分的频域特性，分别定义各模态分量的边际谱 $h_s(w)$的频谱中心($C_{f,s}$)、差异系数($\sigma_{f,s}^2$)及谱偏度($\beta_{f,s}$)作为频谱特征量，即：

$$C_{f,s}=\sum_w wh_s(w)\Big/\sum_w h_s(w) \tag{12}$$

$$\sigma_{f,s}^2=\sum_w(w-C_{f,s})^2h_s(w)\Big/\sum_w h_s(w) \tag{13}$$

$$\beta_{f,s}=\sum_w\left(\frac{w-C_{f,s}}{\sigma_{f,s}}\right)^3h_s(w)\Big/\sum_w h_s(w) \tag{14}$$

式中，w 为频率。

2.2 功率特征

经 FFT 变换后的 k 阶频域信号 $U(k)$、$I(k)$为：

$$\begin{cases}U(k)=U_{\mathrm{m}}(k)-\mathrm{j}U_{\mathrm{n}}(k)\\ I(k)=I_{\mathrm{m}}(k)+\mathrm{j}I_{\mathrm{n}}(k)\end{cases} \tag{15}$$

式中，$U_{\mathrm{m}}(k)$、$I_{\mathrm{m}}(k)$、$U_{\mathrm{n}}(k)$、$I_{\mathrm{n}}(k)$分别为复变量 $U(k)$、$I(k)$的实部和虚部。

由此可求得有功功率 P 和无功功率 Q 分别为：

$$\begin{cases}P=\dfrac{1}{N_1^2}U_{\mathrm{m}}(0)I_{\mathrm{m}}(0)+\\ \qquad\dfrac{2}{N_1^2}\displaystyle\sum_{k=1}^{N_1/2-1}[U_{\mathrm{m}}(k)I_{\mathrm{m}}(k)+U_{\mathrm{n}}(k)I_{\mathrm{n}}(k)]\\ Q=\dfrac{2}{N_1^2}\displaystyle\sum_{k=1}^{N_1/2-1}[U_{\mathrm{m}}(k)I_{\mathrm{n}}(k)-U_{\mathrm{n}}(k)I_{\mathrm{m}}(k)]\end{cases} \tag{16}$$

式中，N_1 为一个稳态周期内电流、电压的采样点数。

2.3 V-I 轨迹特征

功率特征在低功率区域容易出现特征重叠现象，而时频特征只能表征非线性设备的电流信号特性。为深度挖掘负荷特性，提高电力数据识别精度，进一步利用电压、电流之间的耦合关系，基于其物理意义量化 10 个 V-I 轨迹特征，分别为电流跨度、区域面积、环路方向、不对称性、平均曲线率、自相交交点数、中间线峰值、中间线形状、左右段区域及瞬时导纳，其具体定义和量化过程详见文献[6]。

3 基于随机森林的特征选择算法

在多域组合特征的基础上，对特征的相关性和冗余性进行评估，筛选出辨识精度较优的特征组合，降低识别模型复杂度，增加模型的可解释性，对此提出一种基于随机森林(RF)特征重要性与广义序列后向选择(GSBS)的特征选择算法(RF-GSBS)，电力数据识别框架见图 1。

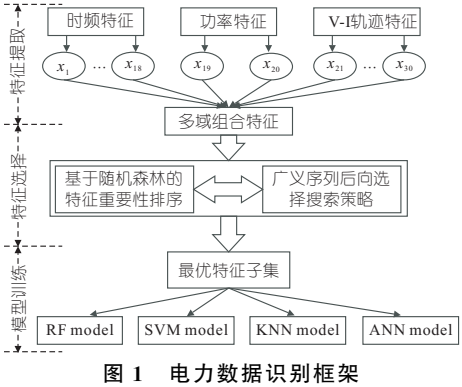


图 1 电力数据识别框架

Fig. 1 Framework of power data identification

3.1 特征重要性的计算

在 RF 算法训练过程中引入随机属性的选择，通过基尼(Gini)指数度量指标评估特征的重要性。假设样本有 K 个类别，在节点 m 处所有样本中，样本点属于第 k 类的概率为 $p_{m,k}$，则概率分布的基尼指数用 G_m 表示，其计算公式为：

$$G_m = \sum_{k=1}^{K} p_{m,k}(1-p_{m,k}) = 1 - \sum_{k=1}^{K} p_{m,k}^2 \quad (17)$$

假设有 c 个特征 $X_1 \sim X_c$,特征 X_i 在节点 m 分裂前后的 Gini 指数变化量 $V_{\mathrm{IM},im}$ 定义为:

$$V_{\mathrm{IM},im} = G_m - G_l - G_r \quad (18)$$

式中,G_l、G_r 均为分裂后新节点的 Gini 指数。

假设决策树 j 中包含特征 X_i 的节点集合为 M,特征 X_i 在第 j 棵树的重要性评分可定义为:

$$V_{\mathrm{IM},ij} = \sum_{m \in M} V_{\mathrm{IM},im} \quad (19)$$

假设 RF 共有 n 棵决策树,特征 X_i 在所有决策树节点分裂不纯度的平均改变量 $V_{\mathrm{IM},i}$ 可定义为:

$$V_{\mathrm{IM},i} = \sum_{j=1}^{n} V_{\mathrm{IM},ij} \quad (20)$$

最后对所求得的重要性评分进行归一化处理,即:

$$V'_{\mathrm{IM},i} = V_{\mathrm{IM},i} \Big/ \sum_{i=1}^{c} V_{\mathrm{IM},i} \quad (21)$$

3.2　特征选择搜索策略及算法设计

为降低特征选择的随机性和人为因素的干扰,在 RF 特征选择过程中引入 GSBS 算法。GSBS 为序列后向选择的加速算法,根据评价函数每次从特征集合中剔除 L(默认 $L=1$)个重要性得分最低的特征,并重新计算分类的准确率,经过多次迭代,选取分类准确率最高且特征个数最少的候选特征集作为最终特征选择结果。为增强选择结果的鲁棒性和可靠性,算法采用 10 折交叉验证评估分类结果。

特征选择算法的具体流程见图 2。首先设置全局最大识别准确率 $\max A_{\mathrm{cc}}$ 及对应的最优特征子集 S_{best};其次采用 10 折交叉验证训练评估随机森林识别模型,取 10 次迭代中模型平均识别准确率作为该特征子集的识别准确率并赋值于 $\max A_{\mathrm{cc}}$,取 10 次迭代中识别准确率最高一次迭代产生的变量重要性分数作为特征排序的依据;然后,根据特征排序结果,删除数据集中特征重要性分数最低的 L 个特征得到最新的训练集。重复执行此步骤,直到特征子集中的特征数目为 0。最后,输出特征选择过程中的最大平均识别率,对应的特征子集即为最优特征子集。

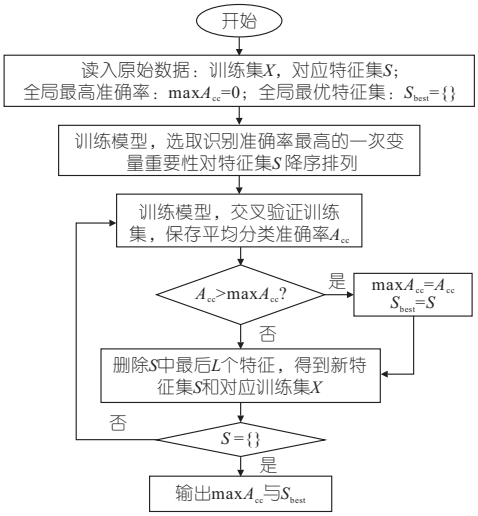


图 2　特征选择算法 GSBS 流程

Fig. 2　Flowchart of GSBS feature selection algorithm

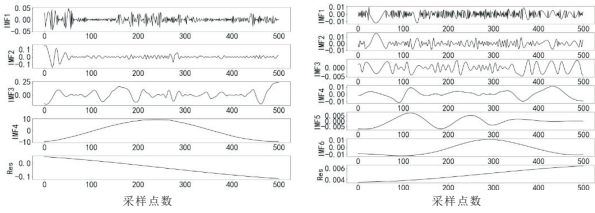
4　算例验证

4.1　数据集描述

基于高频公开数据集 PLAID[7] 对所提方法的有效性进行验证。将 PLAID 数据集中每种设备类型的 80%的数据作为训练集,剩余 20%的数据作为测试集。

4.2　数据集特征提取

以吹风机和空调为例,首先对吹风机和空调数据进行 EMD 分解,分解结果见图 3。由图 3 可知,吹风机电流序列经 EMD 后得到 4 阶 IMF 及 1 个残差分量,而空调电流序列得到 6 阶 IMF 及 1 个残差分量,且对应的 IMF 幅值和波形具有明显区别。为保持特征向量的统一性,本文只保留前 3 阶 IMF 分量,对于 IMF 阶数少于 3 的数据序列,采用零向量补足特征信号。将计算出的响应阶次的时频域特征与功率特征及 V-I 轨迹特征的特征向量组合,维数为 30,经 RF-GSBS 筛选出特征子集,最后将其输入分类模型进行训练与测试。



(a) 吹风机电流序列　　(b) 空调电流序列

图 3　电流信号的 EMD 分解图

Fig. 3　EMD decomposition diagram of current signal

4.3　评估指标

$F_{1\text{-score}}$ 作为一项综合评价指标,是精确率 P_{rec} 和查全率 R_{ec} 的加权调和平均。其定义为:

$$F_{1\text{-score}} = 2P_{\mathrm{re}}R_{\mathrm{ec}}/(P_{\mathrm{re}}+R_{\mathrm{ec}}) \quad (22)$$

其中

$$P_{\mathrm{re}} = T_{\mathrm{TP}}/(T_{\mathrm{TP}}+F_{\mathrm{FN}}) \quad (23)$$

$$R_{\mathrm{ec}} = T_{\mathrm{TP}}/(T_{\mathrm{TP}}+F_{\mathrm{FP}}) \quad (24)$$

式中,T_{TP} 为真实类别为正类,预测类别为正类;F_{FP} 为真实类别为负类,预测类别为正类;F_{FN} 为真实类别为正类,预测类别为负类。

识别准确率 A_{cc} 定义为:

$$A_{cc}=\sum_{i=1}^{K}T_{TP_i}\Big/\sum_{i=1}^{K}(T_{TP_i}+F_{FP_i}+F_{FN_i}) \quad (25)$$

式中，K 为电力数据的类别种类；T_{TP_i}、F_{FP_i}、F_{FN_i} 分别为第 i 类的 T_{TP}、F_{FP}、F_{FN}。

为满足多分类评估需要，$F_{1\text{-score}}$ 指标需通过算术平均每类的度量值加以扩展，具体定义为：

$$F_{1_{macro}}=\frac{1}{K}\sum_{i=1}^{K}F_{1\text{-score}_i} \quad (26)$$

式中，$F_{1_{macro}}$ 为 $F_{1\text{-score}}$ 的宏平均值；$F_{1\text{-score}_i}$ 为第 i 类数据的 $F_{1\text{-score}}$ 值。

4.4 仿真结果与分析

4.4.1 不同特征辨识效果对比

为分析多域特征提取的有效性，将组合特征与单一特征进行比较分析，见表 1。利用所提方法提取特征，引入时频特征（FS1），功率特征（FS2），V-I 轨迹形状特征（FS3），组合特征集（PFS：FS1、FS2、FS3 的融合）4 个特征集。由表 1 可知，组合特征识别性能远远优于单一特征识别性能，与单一特征中识别性能最优的时频特征的 $F_{1_{macro}}$ 相比，提高了 2.98%。另外，对比 3 个单一特征与组合特征的每类设备的识别性能，除了空调和笔记本电脑外，其他类别电力数据的识别率均得到明显提升。可见，识别模型可利用时频特征、功率特征和 V-I 轨迹形状特征之间的互补性，克服单一特征的缺点，深度挖掘数据中蕴藏的差异性特征信息，从而提高模型的辨识能力。

表 1 不同特征下的识别结果的对比

Tab. 1 Comparison of recognition results of different feature sets

设备类别	$F_{1\text{-score}}$/%			
	FS1	FS2	FS3	PFS
空调	69.90	71.79	66.32	70.50
荧光灯	92.57	80.37	93.62	94.62
风扇	81.21	63.77	82.35	88.47
冰箱	60.58	47.62	54.63	66.45
吹风机	92.58	92.47	79.76	95.32
热水器	90.10	86.96	87.12	92.96
白炽灯	86.89	84.44	88.89	92.44
笔记本电脑	96.03	80.81	92.31	93.59
微波炉	90.74	86.49	89.74	92.56
吸尘器	96.65	100.00	100.00	100.00
洗衣机	80.21	36.36	72.14	82.21
$F_{1_{macro}}$	85.22	75.55	82.44	88.10

4.4.2 特征选择过程分析

随着不相关和干扰特征的删除，RF 模型的识别性能在早期有明显提高，然后趋于稳定。当特征子集的特征个数达到 12 时，RF 模型的识别准确率达到最大值 92.41%。若再进一步特征搜索，识别准确率开始逐渐下降。另外，识别模型的准确率在特征选择初期变化不大，这是因为 RF 算法本身具有一定的特征选择功能。

随着 L 值的增加（1、2、3、4、5），模型的识别性能有所下降，这是因为数据集中特征维数较小，尽管 L 值增加可加快算法搜索和收敛的速度，但也容易误删重要变量导致识别性能下降。综合考虑算法的识别精度和时间复杂度，将参数 L 设置为 2。

4.4.3 不同特征选择方法识别性能对比

为进一步验证 RF-GSBS 特征选择算法的有效性，基于特征数据集与 RF 分类模型，与 ReliefF、mRMR 及 RF-RFE 特征选择算法进行对比，见表 2。由表 2 可知，4 种算法筛选出的特征个数相差不大，但 RF-GSBS 算法选择出的特征子集的识别性能明显优于其他 3 种算法，能够获得更好的识别效果。

表 2 不同特征选择算法的选择结果与识别精度

Tab. 2 Feature selection result and recognition accuracy for different feature selection algorithms

特征选择方法	A_{cc}/%	最优特征子集特征数	特征选择方法	A_{cc}/%	最优特征子集特征数
所有特征	89.10	30	RF-RFE	92.26	15
ReliefF	91.56	12	RF-GSBS	92.41	14
mRMR	90.97	15			

4.4.4 不同分类模型在最优特征集上识别性能

为进一步分析特征选择算法的泛化能力，除 RF 模型外，使用 K 最近邻算法（KNN）、支持向量机（SVM）、人工神经网络（ANN）作为分类模型。试验结果见图 4，对比特征选择前后不同分类模型的识别率可知，基于 RF-GSBS 选出的特征集进行识别的分类精度均有所提升，分别达到了 3.7%、9.2%、5.9%、3.6%，表明所筛选出的特征集具有较强的泛化能力和鲁棒性，能有效提高模型的识别性能。此外，与 RF 模型相比，其他 3 种模型对无关和冗余特征更敏感，导致特征选择前的准确率较低。相比于其他 3 种识别模型，RF 模型具有更强的抗干扰能力与泛化能力，能获得良好的识别效果。

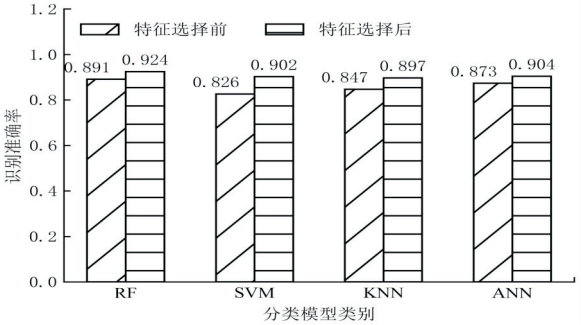


图 4 不同分类模型在最优特征集上分类准确率图

Fig. 4 The classification accuracy of different classifiers based on the optimal feature sets

基于 4 种识别模型,测试了特征选择前后模型从训练到测试的总耗时。如图 5 所示,模型的运行时间均在 ms 级别,同模型特征选择后运行时间较特征选择前相对减少 117、154、15、34 ms,表明特征选择可加快模型的计算速度。在电力数据识别过程中,RF 模型的耗时并非最少,识别精度却最高,但 RF 模型的计算速度仍控制在 ms 级,因此即使在普通嵌入式设备中,模型计算速度也完全能够达到实时分类预测要求,对于部署环境算力的要求较低。KNN 算法相比于其他 3 种算法,运行速度最快,但识别准确率最低,对比 RF 模型,准确率下降了 2.7%。因此综合权衡算法的识别精度和时间复杂度,选择 RF 模型作为最终的识别模型。

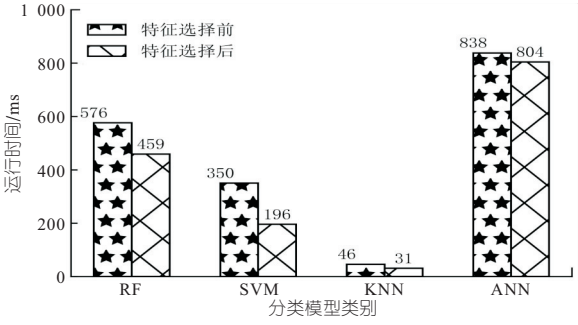


图 5　不同分类模型在最优特征集上运行时间

Fig. 5　The running time of different classifiers based on the optimal feature sets

5　结论

a. 引入基于 EMD 与 Hilbert 变换的时频域特征,能够实现电力数据表征量化的多样性,同时增强异域特征互补优势,有效提高电力数据的识别精度。

b. 考虑特征选择可有效删除冗余变量,降低模型计算复杂度,进一步提高模型精度与计算速度,有利于实现电力数据状态变化在线监测。

参考文献:

[1] 卢瑞瑞,于海阳,杨震,等. 基于能源分解的用户用电行为模式分析[J]. 北京航空航天大学学报, 2022, 48(2): 311-323.

[2] 邓晓平,张桂青,魏庆来,等. 非侵入式负荷监测综述[J]. 自动化学报,2022,48(3):644-663.

[3] 殷煌凯,许仪勋,李宁,等. 基于复现性和熵权区分度的非侵入式负荷识别方法[J]. 水电能源科学, 2019, 37(10): 163-167.

[4] 刘兵,郑承利. 基于 EMD 特征提取的高频面板数据自适应聚类方法[J]. 统计与决策,2022, 38(10): 16-20.

[5] 朱永利,贾亚飞,王刘旺,等. 基于改进变分模态分解和 Hilbert 变换的变压器局部放电信号特征提取及分类[J]. 电工技术学报, 2017, 32(9): 221-235.

[6] WANG A L, CHEN B X, WANG C G, et al. Non-intrusive load monitoring algorithm based on features of V-I trajectory[J]. Electric power systems research, 2018, 157:134-144.

[7] GAO JINGKUN, GIRI S, KARA E C, et al. PLAID: A public dataset of high-resolution electrical appliance measurements for load identification research:demo abstract[C]//1st ACM Conference on Embedded Systems for Energy-Efficient Buildings, ACM. 2014: 198-199.

Power Data Identification Method Based on Multi-Domain Feature Analysis and Selection

HONG De-hua[1], LIU Cui-ling[1], ZHAO Lin-yan[1], LEI Qin-yi[1], WANG Hai-xin[2]

(1. Information and Communication Branch of State Grid Anhui Electric Power Co., Ltd., Hefei 230061, China;
2. School of Electrical Engineering, Shenyang University of Technology, Shenyang 110870, China)

Abstract: To solve the problem of low recognition accuracy caused by insufficient power data feature mining, this paper proposed a novel power data identification method based on multi-domain feature analysis and feature selection. Firstly, aiming at the shortcomings of existing power data feature extraction methods, a feature extraction method based on empirical mode decomposition (EMD) and Hilbert transform (EMD-Hilbert) was proposed, and the power features and V-I trajectory features of power data were quantified. Secondly, based on random forest and generalized sequence backward selection search strategy, the optimal feature subset was obtained. The random forest was employed to build a recognition model for the power data. Finally, the experimental results verified the effectiveness and identification accuracy of the proposed method. The results show that the proposed method can utilize the complementarity of different features to overcome the problem of low accuracy by single feature, and further improve the model recognition performance through feature selection.

Key words: power data identification; multi-domain feature extraction; feature selection; random forest; generalized sequence backward selection