DOI:10.19651/j.cnki.emt.2108711

基于特征加权 KNN 的非侵入式负荷识别方法

朱 浩[1] 曹 宁[1] 鹿 浩[1] 张正基[1] 柯 炜[2]
(1.河海大学计算机与信息学院 南京 211100; 2.江苏业力科技有限公司 南京 210061)

摘 要: 针对不同稳态特征对识别结果的影响程度不同,并考虑到不平衡数据集造成的少数类误判的问题,提出一种基于特征加权 KNN 的非侵入式负荷识别方法。首先,采用熵权法计算特征权重,利用特征权重改进特征距离的计算。其次,根据样本数量和对应算法 k 值计算得到表决权重,带入投票表决过程中,以此来增加少数类的分类准确性。实验结果表明,针对实测负荷数据集时,本文算法的平均识别准确率为 93.4%,与 KNN 算法相比提高了 2.8%;针对公开数据集时,本文算法的平均准确率和 F1 得分分别为 86.8%和 81.6%,要优于其他 4 种分类算法。

关键词: 非侵入式;负荷识别;稳态特征;KNN;特征权重;表决权重
中图分类号: TM714 **文献标识码:** A **国家标准学科分类代码:** 470.40

Non-intrusive load identification method based on feature weighted KNN

Zhu Hao[1] Cao Ning[1] Lu Hao[1] Zhang Zhengji[1] Ke Wei[2]
(1. School of Computer and Information, Hohai University, Nanjing 211100, China;
2. Jiangsu Yeli Technology Co., Ltd., Nanjing 210061, China)

Abstract: In view of the different influence of different steady-state features on the identification results, and considering the misjudgment of minority classes caused by unbalanced data sets, a non-invasive load identification method based on feature weighted KNN is proposed. Firstly, the feature weight is calculated by entropy weight method, and it is used to improved feature distance calculation. Secondly, the voting weight is calculated according to the number of samples and the k value of algorithm, which is brought into the voting process to increase the classification accuracy of minority classes. The experimental results show that the average recognition accuracy of algorithm in this paper is 93.4%, which is 2.8% higher than that of KNN algorithm. For public data sets, the average accuracy and F1 score of algorithm in this paper are 86.8% and 81.6%, which are better than the other four classification algorithms.

Keywords: non intrusive; load identification; steady characteristics; KNN; voting weight; feature weight

0 引 言

推进用电侧的智能化,其目的在于优化能源结构、改善用电效率以及提高用电安全。目前的电表只能获取到某段时间内的用电总量,并不能清晰全面地体现出用户的用电情况。因此对于电网而言,负荷监测就显得十分必要。通过负荷监测,电网可以及时地分析用户的用电行为和电器设备能耗情况,对引导用户节能用电和优化用电资源管理都有着重大意义[1]。

负荷监测目前的方式主要有两种。第 1 种是侵入式方法,主要通过配置在设备上的采集装置来获取用电信息,计量较为准确,缺点是装置安装和维护难度大。第 2 种是非侵入式方法(non-intrusive load monitoring, NILM)[1-2],直接在电力入口处加装采集装置,通过算法对原始负荷数据进行分析和挖掘。相对侵入式方法,NILM 能带来更大的经济效益,并且更易于实现和维护[3-4]。

随着智能电网的发展和普及,NILM 系统已然成为了国内外学者的一个研究热点,主要的工作体现在特征提取和负荷识别这两个方面。文献[5]通过多点均值和极值差量两种形式,优化了谐波特征量的提取方法,并验证了该方法可提升负荷辨识的精度。文献[6]通过贝叶斯滤波法构建负荷识别模型,提升了算法对不同类型负荷的自适应能力,并且优化了算法求解速度。文献[7]根据贝叶斯信息准则进行特征筛选,以加权皮尔逊距离作为判别准则,实现具名负荷的高精度识别。文献[8]基于电器的历史数据来确定其先验概率,并以此获取用电行为趋势,实验表明该方法精度高且成本低。文献[9]利用随机森林来降低数据的复

收稿日期:2021-12-28

杂度,接着结合烟花算法来优化随机森林模型的参数,提高了分类的准确度和效率。文献[10]提出基于自联想神经网络和多层感知神经网络的堆积神经网络模型,其中自联想神经网络用于提取特征,并增强特征的相关性;而多层感知神经网络则用于负荷分类。

传统的负荷识别算法很少会研究特征本身对算法的影响程度。本文以非侵入式负荷检测为背景,通过实测采集各种负载单独及组合工作时的稳定电流、电压和相位数据,利用 k 最近邻算法(k-nearest neighbor,KNN)算法[11]构建分类模型。由于本文所用的特征库是在实测数据的基础上建立的,因此特征库种各类样本的数量并不平衡。针对不同特征在特征距离计算时所占据的重要程度有所不同,并考虑 KNN 算法对少数类样本存在误判的情况,本文提出表决加权结合特征加权的 KNN(voting-weighted combined with feature weighted k-nearest neighbor, VFKNN)。首先,计算不同特征值的权重,将其代入到算法的距离公式之中,从距离计算上扩大相似数据之间的区别。其次,在投票判决选出 k 个近邻之前,为不同数量的样本分配权重,以此来均衡不同数量样本;最后利用实测数据来验证算法的优越性。

1　基本原理

KNN 算法的原理为距离匹配,其核心有 3 个步骤,分别为距离计算,k 个近邻选取,投票表决[11]。本文提出的负荷识别算法流程如图 1 所示。算法基于负荷稳定段数据。事件检测的主要作用是通过对原始数据中投切点的定位,来划分提取出稳定段数据,从而能从中提取出负荷特征,并建立起负荷稳态特征集;而对 KNN 的加权方式主要体现在两方面:1)通过增加特征权重来改进欧式距离计算公式,使得计算的距离更能体现负荷之间的差异;2)取各类样本数量的倒数,并且结合 KNN 算法的 k 值,对近邻进行表决加权,以均衡表决权的方式来改进算法投票表决的规则。

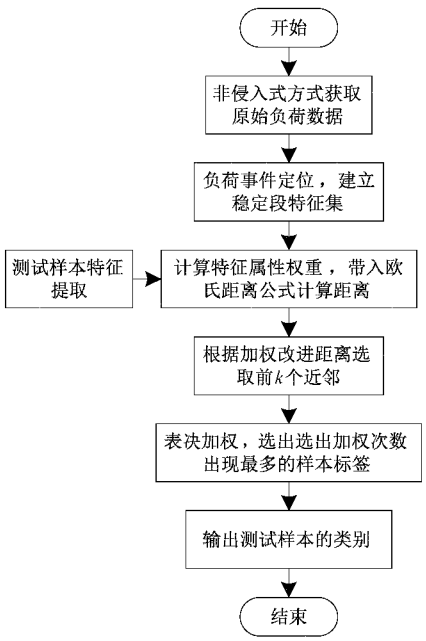

图 1　本文算法原理

2　稳态特征提取

所提取的特征都是建立在负荷稳定工作的基础上的,但实测的数据中包含了电器的投切动作即负荷事件[12],此外,因为噪声的影响以及部分电器工作时的不稳定性,原始数据需要经过筛选处理来提取出稳定工作数据段。文中采用改进的双边累计和(cumulative sum,CUSUM)算法来进行对负荷事件的检测并消除噪声的影响,以检测结果为依据提取出稳定段。

图 2(a)所示,为电饭煲工作时的电流波形,其中包含了稳定工作时段和负荷事件。因为电饭煲电流波动较大,CUSUM 算法无法识别出准确的负荷事件。因此在 CUSUM 识别之前,对电流数据进行平滑处理(数据水平处理)[13]。

平滑处理的具体步骤如下:

1)每种负荷取连续 30 个点的稳定电流数据,5 个数据为一小段,计算每一小段的平均值 $\bar{x}_i(i=1,\cdots,6)$。

2)根据步骤 1)中的平均值来计算数据偏离阈值 H,公式如下:

$$H=\max(|x_i-x_{i-1}|) \tag{1}$$

3)针对原始负荷电流数据,重复步骤 1),判断当前段数据均值与相邻数据段均值之间的差,若小于 H 则保留,如大于 H,则做如下处理:

$$\bar{x}_i=\frac{\bar{x}_{i-1}+\bar{x}_{i+1}}{2} \tag{2}$$

其中,初始的负荷数据都是稳定的,平滑处理的效果如图 2(b)所示。

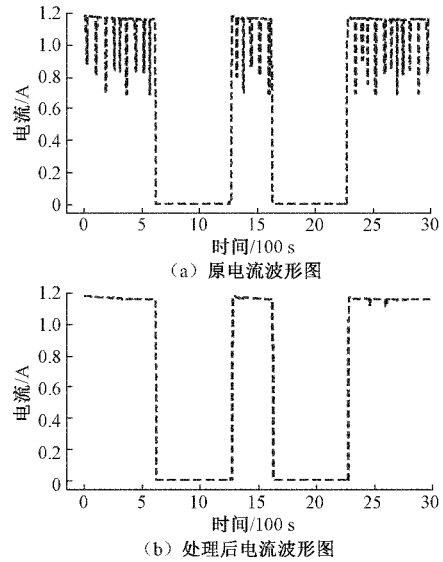

图 2　电饭煲电流波形图

分别以电饭煲稳定段的电流数据(序号 1)、包含非稳定段的电流数据(序号 2)以及经过平滑处理后的电流数据(序号 3)如表 1 所示。

表 1　电流数据对比

序号	电流有效值/A	相对误差/%
1	1.177	0
2	1.116	5.2
3	1.151	2.2

通过平滑处理后，数据更为准确。CUSUM 算法检测出的事件结果对应如表 2 所示。

表 2　事件定位结果

序号	事件定位点	是否符合实际
1	618	是
2	1 277	是
3	1 623	是
4	2 276	是

将改进 CUSUM 检测出的负荷事件发射点与负荷电流波形图对比可知，算法检测的结果能达到一定准确率，满足文中负荷检测的要求。

在 CUSUM 提取稳定工作段之后，构建电流有效值、伏安比系数、谐波总畸变系数、各次谐波(主要为 3 次和 5 次)与有功功率等特征。

3　特征加权 KNN 的负荷识别方法

3.1　特征权重计算

每一种特征值对 KNN 算法的重要程度不一样，需要计算每一种特征值的权重或占比。采用了熵权法，具体如下：

1)数据归一化处理。已有 N 类样本，每个样本包含了 d 个具体特征值，对每个特征 $x_{id}(i=1,2,\cdots,N)$ 进行正向化，其构成的矩阵记为 $\tilde{\boldsymbol{X}}$。

$$\tilde{\boldsymbol{X}}=\begin{bmatrix} |x_{11}| & |x_{12}| & \cdots & |x_{1d}| \\ |x_{21}| & |x_{22}| & \cdots & |x_{2d}| \\ \vdots & \vdots & \ddots & \vdots \\ |x_{N1}| & |x_{N2}| & \cdots & |x_{Nd}| \end{bmatrix} \tag{3}$$

对每一个特征 $x_{id}(i=1,2,\cdots,N)$ 进行标准化，标准化后记为 $\tilde{z}_{ij}$，公式如下：

$$\tilde{z}_{ij}=\frac{|x_{ij}|-\min\{|x_{1j}|,\cdots,|x_{Nj}|\}}{\max\{|x_{1j}|,\cdots,|x_{Nj}|\}-\min\{|x_{1j}|,\cdots,|x_{Nj}|\}} \tag{4}$$

$\max\{|x_{1j}|,\cdots,|x_{Nj}|\}$ 代表正向化 d 维特征值中的最大值，$\min\{|x_{1j}|,\cdots,|x_{Nj}|\}$ 对应为最小值。

2)获取每一个特征值所占比重。计算特征值的比重，以概率的形式呈现。每一项特征值的概率以 p_{ij} 表示。

$$p_{ij}=\frac{\tilde{z}_{ij}}{\sum_{i-1}^{N}\tilde{z}_{ij}},\ j=1,2,\cdots,d \tag{5}$$

3)计算每一项特征值的熵值。利用概率 p_{ij} 来计算第 j 个特征值的信息熵 e_j。特征值的信息熵越大，则对算法的重要程度越低，反之亦然。

$$e_j=-\frac{1}{\ln n}\sum_{i=1}^{N}p_{ij}\ln(p_{ij}) \tag{6}$$

4)计算每一项特征值的权重。根据信息熵，可以得到每一项特征值指标的权重 ω_j 为：

$$\omega_j=\frac{d_j}{\sum_{j=1}^{N}d_j} \tag{7}$$

其中，$d_j=1-e_j$ 表示信息效用值。利用式(5)得到的特征权重改进 KNN 的距离公式(欧氏距离)有：

$$d(X_t,X_i)=\sum_{j-1}^{j-d}\sqrt{\omega_j*(x_{tj}-x_{ij})^2} \tag{8}$$

其中，X_t 和 X_i 分别表示测试样本集和稳态特征集，而 x_{tj} 和 x_{ij} 则表示单个测试负荷特征以及单个稳态特征。

3.2　表决加权

由于负荷数据都是实测采集的，而各电器在实测中的工作时长和投切状态不一致，因此最后得到的各类负荷的样本数量也有差异。KNN 在投票表决时，会选出与测试样本最相近得前 k 个近邻，样本数量的差异会诱导近邻的选择。常用的解决方法是，在投票表决的过程中，为不同类别分配对应的权重[14-15]，以此均衡各类样本的表决权。比如文献[16]取各类样本数量的倒数，将其作为表决权重，并代入到投票表决的过程中。然而，这一做法没有考虑到 k 值的变化。随着 k 值的增加，k 个近邻里面出现多数类样本的数量也会增加，而多数类样本的权重应当也被逐渐减少。因此为了均衡不同数量的各类样本在投票表决时的表决权，并考虑到不同 k 值下，多数类样本的在投票表决中的权重也应当有所变化，引入表决权重。具体改进权重计算方法为：

$$weight(T_j)=\left(\frac{1}{num(T_j)}\right)^{\lg(k+\alpha)} \tag{9}$$

其中，$num(T_j)$ 表示 j 类样本 T_j 所包含的特征总数量，k 即为 KNN 算法中的参数 k，α 为修正因子，随着训练集数量和构成的变化而不同，主要用于控制多数类的权重变化，本文 α 值取 3。将式(9)代入到 KNN 算法投票表决的过程中，则待测样本 c 与类别 c_i 的相似度计算公式为：

$$\begin{aligned} c_t&=\operatorname{argmax}_{c}\sum_{c_i\in T_k}weight(T_j)*\delta(c=c_i) \\ \delta(c=c_i)&=\begin{cases}1, & c=c_i\\ 0, & c\neq c_i\end{cases}\end{aligned} \tag{10}$$

其中，T_k 表示待测样本 c 的前 k 个近邻的集合；c_i 表示集合 T_k 中的任意一类样本。

3.3　实验测试与分析

根据历史获取的负荷数据建立特征库,并按 8∶2 的比例划分训练集和数据集,数量分别为 855 和 241。

首先利用熵权法获取训练集中各特征值指标的权重,结果如表 3 所示。

表 3　特征值权重

特征指标	权重
电流有效值	0.054
伏安比	0.486
谐波总畸变	0.212
三次谐波	0.104
五次谐波	0.086
有功功率	0.059

从特征权重中可以看出,伏安比系数、谐波总畸变、三次谐波对算法的重要程度较大,其中伏安比系数占比最大。

取各类样本的倒数作为标签次数的权重,并带入到算法投票表决前,进行表决加权。

通过负荷稳定段特征训练集和测试集分别对如下 4 种算法进行验证。算法 1 为原 KNN 算法;算法 2 为文献[16]提出的加权方式改进的 KNN 算法(VKNN);算法 3 基于特征加权欧氏距离改进的 KNN 算法(FKNN);算法 4 结合特征加权和加权次数改进的 KNN 算法(VFKNN)。k 值取值范围 1 ～ $\sqrt{M}$(M 为训练集样本数量)。

以 k 值为横坐标,准确率 P(预测正确的样本数与实际样本总数之比)为纵坐标作图,4 种算法准确率对比结果如图 3 所示。

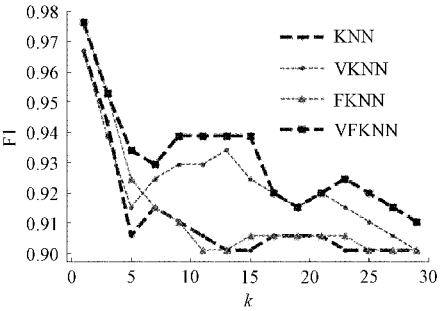

图 3　算法准确率实验对比

随着 k 值的增加,VKNN 和 FKNN 表现效果不佳,甚至出现准确率低于 KNN 算法的情形。其中本文提出的方法,即 VFKNN 算法识别准确率最高,始终在 4 种方法中最优。实验证明了本文提出的识别方法提高了识别准确率。

已知精确率为 Pe(实际中正类样本数 S_t 与预测为正类样本数 S_a 之比)和召回率 R(实际中正类样本数 S_t 与应有样本数 $\tilde{S}_a$ 之比),则 F1 得分可以表示为:

$$\mathrm{F1}=\frac{Pe*R*2}{Pe+R},Pe=\frac{S_t}{S_a},R=\frac{S_t}{\tilde{S}_a} \tag{11}$$

为了更好地证明本文方法在性能上的优越性,采用召回率 R 和 F1 分数,对 4 种算法的识别表现进行实验对比,如图 4 所示。通过图 4(a)、(b)两幅图,可以看到 VFKNN 在召回率和 F1 得分两项指标上都要优于其他 3 种算法。

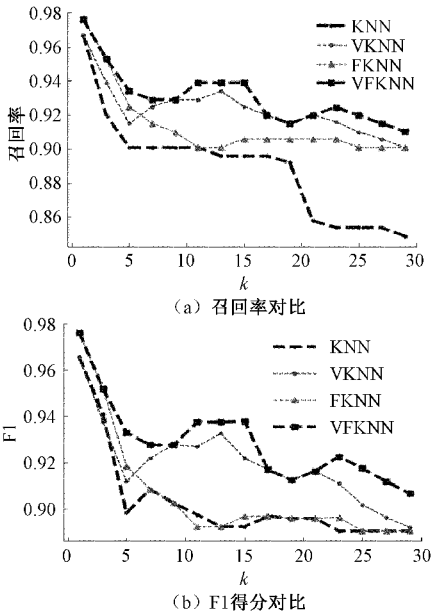

图 4　算法表现对比

经过分析,训练集中的“风扇+锂电池”为少数类(样本数为 3),记为类别 1;而“电饭煲”和“空调+锂电池+冰箱”、“冰箱+空调”和“空调+锂电池”属于相似类,前者记为类别 2,后者记为类别 3;在 $k=5$ 时,改进过后的 VFKNN 算法对比原 KNN 算法,负荷整体识别率由 90.6%提升到 93.4%,相较于 VKNN 和 FKNN 算法也分别提高了 2%和 1%。4 种算法对这两种特殊类别的识别情况如表 4 所示。在相似类中,只针对“电饭煲”与“冰箱+空调”的单向识别情况。

表 4　4 种算法对特殊类别的识别情况

类别	识别数量	识别错误率/%			
		KNN	VKNN	FKNN	VFKNN
类别 1	3	66.7	0	66.7	0
类别 2	18	27.8	27.8	22.2	16.7
类别 3	27	29.6	29.6	22.2	22.2

对比 4 种算法在特殊类别下的识别错误率,可以看出 VFKNN 算法提高了了相似类样本电饭煲与空调+锂电池+冰箱,以及冰箱+空调与空调+锂电池的区分度;此外,还增加了少数类样本风扇+锂电池和电饭煲的识别准确率。

3.4　算法有效性分析

根据第 3.3 节的实验及分析,证明了本文算法对于 KNN 算法的改善。为了验证本文方法的有效性以及对比

于其他分类算法的优势，选取 9 种公开的不平衡数据集进行测试，数据集的具体信息如表 5 所示。用于对比的算法分别为 KNN、朴素贝叶斯(naive bayes classifier，NBC)、支持向量机(support vector machines，SVM)、逻辑回归(logistic regression，LR)以及 VFKNN。各算法在 9 种数据集上的准确率和 F1 得分对比情况分别如表 6 和 7(其中 KNN 和 VFKNN 取最优 k 值，为 5)所示。

表 5　数据集信息

名称	简称	特征数量	样本量	不平衡度
hayes-roth1	hr1	4	132	1.70
glass1	g1	9	214	1.82
haberman	hm	3	305	2.78
new-thyroid	nt	5	215	4.84
ecoil1	e1	5	336	5.46
glass6	g6	9	214	6.38
yeast3	y3	8	1 484	8.10
glass	g0	9	214	8.44
cleveland-0-vs-4	cov4	13	177	12.62

表 6　准确率对比　%

数据集	KNN	NBC	SVM	LR	VFKNN
hr1	45.3	33.3	77.8	44.4	74.1
g1	76.4	67.4	67.4	65.1	79.9
hm	71.6	67.2	72.1	72.1	72.5
nt	86.4	95.4	81.8	95.5	95.5
e1	89.7	85.3	92.6	88.2	92.7
g6	85.4	93.0	81.4	90.7	96.3
y3	92.8	87.5	94.3	90.6	93.0
g0	65.8	48.8	32.6	62.8	80.1
cov4	97.4	91.4	97.1	94.3	97.5

表 7　F1 得分对比

数据集	KNN	NBC	SVM	LR	VFKNN
hr1	52.5	27.1	77.2	22.4	68.9
g1	57.8	55.1	54.3	59.2	70.8
hm	53.6	66.5	67.1	65.3	67.9
nt	85.8	95.0	81.2	95.3	95.3
e1	88.4	78.5	91.9	87.2	92.2
g6	82.7	92.8	73.0	90.7	91.7
y3	88.5	81.7	94.1	88.2	88.9
g0	63.2	44.7	16.0	58.4	69.4
cov4	78.9	94.9	95.7	94.3	89.0

根据实验测试可知，在准确率表现上，VFKNN 算法在 7 个数据集上性能最优；在 F1 得分表现上，VFKNN 算法在其中 5 个数据集上性能最优。综合而言，VFKNN 在 9 个数据集上的平均准确率和 F1 得分分别为 86.8%和 81.6%，在 5 种算法中均为最优。因此可以得出结论，VFKNN 算法在处理不平衡数据集分类时具有一定的有效性。

3.5　利用实测数据进行测试

实测环境下，控制间隔为 200 ms，通过采集设备获取电饭煲、冰箱＋空调、锂电池＋空调、风扇＋锂电池等 4 种负荷(第 3.3 节中的特殊类别)的实测数据。共采集 130 组负荷稳定工作时的电流、电压和相位数据，并提取算法对应的负荷特征，将其作为测试样本。将第 3.3 节中的数据集作为训练样本。表 8 为本文算法 VFKNN 的性能表现(其中 $k=5$)。整体识别准确率为 0.854，针对实测环境下的电器，本文算法均能达到较高的识别准确率。

表 8　具体负荷识别情况

种类	各类数量	准确率	召回率	F1 得分
电饭煲	20	1.000	0.800	0.889
冰箱＋空调	60	0.881	0.867	0.874
风扇＋锂电池	20	1.000	1.000	1.000
锂电池＋空调	30	0.742	0.800	0.770

4　结　　论

针对不同负荷特征在算法中重要程度不同的情况，并考虑到负荷识别时，不平衡数据集中少数类存在误判的问题，本文提出一种基于特征加权 KNN 的非侵入式负荷识别方法。将特征权重代入到欧氏距离公式中，改善样本数据与测试数据之间的距离计算；为了提高少数类在投票表决时的权重，引入表决权重，来改进投票表决的过程。实验结果表明，本文所提算法对实测数据集的平均识别准确率为 93.4%，较原 KNN 算法提高了 2.8%，且提高了对少数类样本的识别准确率；而在公开数据集上，本文算法的平均识别准确率和 F1 得分分别为 86.8%和 81.6%，优于传统的朴素贝叶斯、支持向量机和逻辑回归分类算法。综上表明本文算法在处理负荷不平衡数据集分类时，具有一定有效性。后续工作重点在于，提高本文算法对平衡数据集的适用性，并优化特征提取和构建，进一步提高算法识别准确率。

参考文献

[1]　陈家瑞，陈忠孝，秦刚，等. 基于 PSO 算法与 SVR 算法在企业直流配电网短期负荷预测的研究[J]. 国外电子测量技术，2020，39(12)：70-73.

[2]　程祥，李林芝，吴浩，等. 非侵入式负荷监测与分解研究综述[J]. 电网技术，2016，40(10)：3108-3117.

[3]　KIM J，LE T T，KIM H. Nonintrusive load monitoring based on advanced deep learning and novel signature [J]. Computational Intelligence and

Neuroscience, 2017, 2017: 4216281.

[4] 陈军锋，王雪，张效天. 非侵入式负荷识别边缘计算颜色编码研究[J]. 仪器仪表学报，2020,41(9): 12-19.

[5] 吕志宁，赵少，饶竹一，等. 非侵入负荷辨识的谐波特征量提取改进方法研究[J]. 电子测量技术，2019，42(7): 29-34.

[6] 陈中，方国权，赵家庆，等. 基于贝叶斯迭代的非侵入式负荷事件检测方法[J]. 电测与仪表，2021，58(4): 1-8.

[7] 夏飞，张洁，张浩，等. 基于 BIC 准则和加权皮尔逊距离的居民负荷模式精细识别及预测[J]. 电子测量与仪器学报，2020，34(11): 33-42.

[8] WELIKALA S, DINESH C, EKANAYAKE M, et al. Incorporating appliance usage patterns for non-intrusive load monitoring and load forecasting [J]. IEEE Transactions on Smart Grid, 2017, 10 (1): 448-461.

[9] TAVEIRA P R Z, MORAES C H V D, LAMBERT-TORRES G. Non-intrusive identification of loads by random forest and fireworks optimization[J]. IEEE Access, 2020, 8: 75060-75072.

[10] CARMO S D J D, CASTRO A R G. Automated non-intrusive load monitoring system using stacked neural networks and numerical integration[J]. IEEE Access, 2020, 8: 210566-210581.

[11] HE Q P, WANG J. Fault detection using the k-nearest neighbor rule for semiconductor manufacturing processes[J]. IEEE Transactions on Semiconductor Manufacturing, 2007, 20(4): 345-354.

[12] 周东国，张恒，周洪，等. 基于状态特征聚类的非侵入式负荷事件检测方法[J]. 电工技术学报，2020，35(21): 4565-4575.

[13] 周倩，韩璞，翟永杰. 电力负荷预测中的数据处理及实验研究[J]. 计算机工程与应用，2010，46(15): 193-195.

[14] TAN S. Neighbor-weighted K-nearest neighbor for unbalanced text corpus [J]. Expert Systems with Application, 2005, 28(4): 667-671.

[15] 刘鹏，杜佳芝，吕伟刚，等. 面向不平衡数据集的一种改进的 k-近邻分类器[J]. 东北大学学报(自然科学版)，2019，40(7): 932-936.

[16] 延菲，张瑞祥，孙耀杰，等. 基于改进 kNN 算法的非侵入式负荷识别方法[J]. 复旦学报(自然科学版)，2021，60(2): 182-188.

作者简介

朱浩，硕士，研究方向为智能电网数据挖掘研究。

E-mail:18262622315@163.com

曹宁，硕士，教授，主要研究方向为计算机测控与通信技术。

E-mail:caoning@vip.163.com

鹿浩，博士，讲师，主要研究方向为信号与信息处理。

E-mail:toluhao@hotmail.com