研究与开发

基于加权随机森林算法的空巢电力用户识别方法

卢子萌[1]，陈佳怡[2]，李璟[1]，谢岳[1]，蒋欣利[2]，韩蕾[3]，郭倩[1]
（1. 中国计量大学机电工程学院，浙江 杭州 310018；2. 国网金华供电公司，浙江 金华 321000；
3. 浙江华云信息科技有限公司，浙江 杭州 310018）

摘　要：针对当前政府和社会对空巢老人的识别缺乏有效技术手段的问题，提出了一种基于加权随机森林算法的空巢电力用户识别方法。首先通过调查问卷获取部分准确空巢用户标签，并从用电水平、用电波动、用电趋势 3 个方面构建用户用电特征库，由于空巢与非空巢存在用户数据不平衡问题，采用加权随机森林算法改善机器学习对数据敏感的现象，将该算法模型在电力公司采集系统部署上线，并对 2 000 户未知类型用户进行空巢识别，其空巢识别准确率达到 74.2%。结果表明，从用电角度研究对空巢老人的识别，可以帮助电网公司了解空巢老人的个性化、差异化需求，从而为用户提供更精细的服务，也可以协助政府和社会开展帮扶工作。

关键词：空巢用户识别；加权随机森林算法；用户用电特征库；数据不平衡
中图分类号：TP181，TP311，F426
文献标识码：A
doi: 10.11959/j.issn.1000−0801.2020249

An empty-nest power user identification method based on weighted random forest algorithm

LU Zimeng[1], CHEN Jiayi[2], LI Jing[1], XIE Yue[1], JIANG Xinli[2], HAN Lei[3], GUO Qian[1]
1. School of Mechanical and Electrical Engineering,China Jiliang University, Hangzhou 310018, China
2. State Grid Jinhua Power Supply Company, Jinhua 321000, China
3. Zhejiang Huayun Information Technology Co., Ltd., Hangzhou 310018, China

Abstract: In view of the lack of effective technical means for the identification of empty-nesters by the government and the society, an empty-nesters prow user identification method based on weighted random forest algorithm was proposed. Firstly, some accurate labels of empty-nest users were obtained through questionnaires, and electricity characteristic library was drawn from three aspects: electricity consumption level, electricity consumption fluctuation and electricity consumption trend. Due to the data imbalance between empty-nest and non-empty-nest users, the weighted random forest algorithm was used to improve the data sensitivity phenomenon of machine learning. Finally, the algorithm model was put online in the power company's acquisition system. The 2 000 unknown users of various types were identified, among which the identification accuracy of empty-nest users was 74.2%. The results show that the identification of empty-nesters from the perspective of electricity consumption can help power grid companies to understand the personalized and differentiated needs of empty-nesters, so as to provide users with more sophisticated

收稿日期：2020−02−18；修回日期：2020−08−05
基金项目：浙江省基础公益研究计划项目（No.LGG20E070003）
Foundation Item: Basic Public Welfare Research Project of Zhejiang Province of China (No.LGG20E070003)

services, and also assist the government and society to carry out assistance work.

Key words: empty-nest user identification, weighted random forest algorithm, user electricity characteristic library, data imbalance

1　引言

随着人均寿命的增长、人口生育率的下降以及老龄化程度的加深，空巢老人数量不断增加[1]。据《2018 年社会服务发展统计公报》统计，截至 2018 年年底，全国 60 岁及以上老年人口为 24 949 万人，占总人口总数量的 17.9%，其中 65 岁及以上人口为 16 658 万人，占人口总数量的 11.9%，在如此庞大的人口数量现状下，政府和社会对空巢老人的识别缺乏有效技术手段。

现阶段部分学者对空巢用户识别问题展开了初步研究，参考文献[2-3]通过手机移动通信等数据开展空巢老人识别研究，利用手机用户的年龄信息、通话和短信清单、终端计费、上网套餐用量等数据建立神经网络算法模型，识别空巢用户，由于此类方法使用的通信数据容易引发用户隐私问题，因此局限性较大。

参考文献[4]提出了一种基于模糊聚类曲线相似度的养殖用户识别方法，该方法可准确将相似度高于阈值的用户识别为养殖户，但却没有进一步分析相似度低于阈值的用户类型。参考文献[5]采用基于“进化”的主成分分析法对用户用电行为进行分类，但并没有针对空巢用户识别问题展开分析，而且聚类一般属于无监督学习算法，并未赋予机器学习标签，无法实现对特定用户群体的识别。

但是，电力大数据作为日常生活的一项基本指标，并没有对空巢用户电力使用上的特征进行充分挖掘。例如城市居民由于经济水平较高，在冬夏两季的用电量可能会明显高于农村居民；而空巢老人的家庭人员数量可能随季节、节假日发生变化，从而导致空巢与非空巢用户具有差异化用电行为[6]。这一特点使得从电能消费角度构建居民用户的用电特征库，对空巢用户群体进行定义和划分具有一定的合理性、应用前景。

基于以上现状，本文提出了一种基于加权随机森林的空巢电力用户识别方法。首先提取部分空巢用户标签并构建用电特性指标库。然后采用加权随机森林算法从“构建加权决策树”和“加权多数投票决策”两方面自适应将分类策略倾向于空巢用户的识别，改善机器学习对不平衡数据的学习能力。最后通过加权随机森林分类器的空巢用户识别泛化性能曲线找出最优投票决策，在此基础上对未知用户群体进行验证。

2　用户用电特征指标库构建

由于不同居民群体在经济水平、生活习惯、生活地域等方面存在巨大差异，因此其电能消费的能力也往往有着很大的不同，其中电力用户用电数据中隐藏着用户用电行为习惯[7-9]，在对浙江某地 5 254 户年用电曲线进行聚类分析后，得到了图 1 所示的 3 种空巢用户和 3 种非空巢用户的典型用电曲线。

图 1（a）为空巢和非空巢典型用电曲线 I，空巢和非空巢用户在夏季用电高峰期都有用电趋势的上升，但空巢用户在其他时间内的用电波动较小，可见空巢用户家庭日常用电规律较为平稳，而非空巢用户用电行为较为随意，导致相邻两天用电波动较大。

图 1（b）为空巢和非空巢典型用电曲线 II，空巢和非空巢用户在夏季和冬季并未出现用电上升趋势，但空巢用户的用电水平较低，可见空巢用户家庭人数较少且生活较为规律，导致用电水平较低，而非空巢用户可能由于家庭人数较多等导致用电水平较高。

图 1（c）为空巢和非空巢典型用电曲线 III，

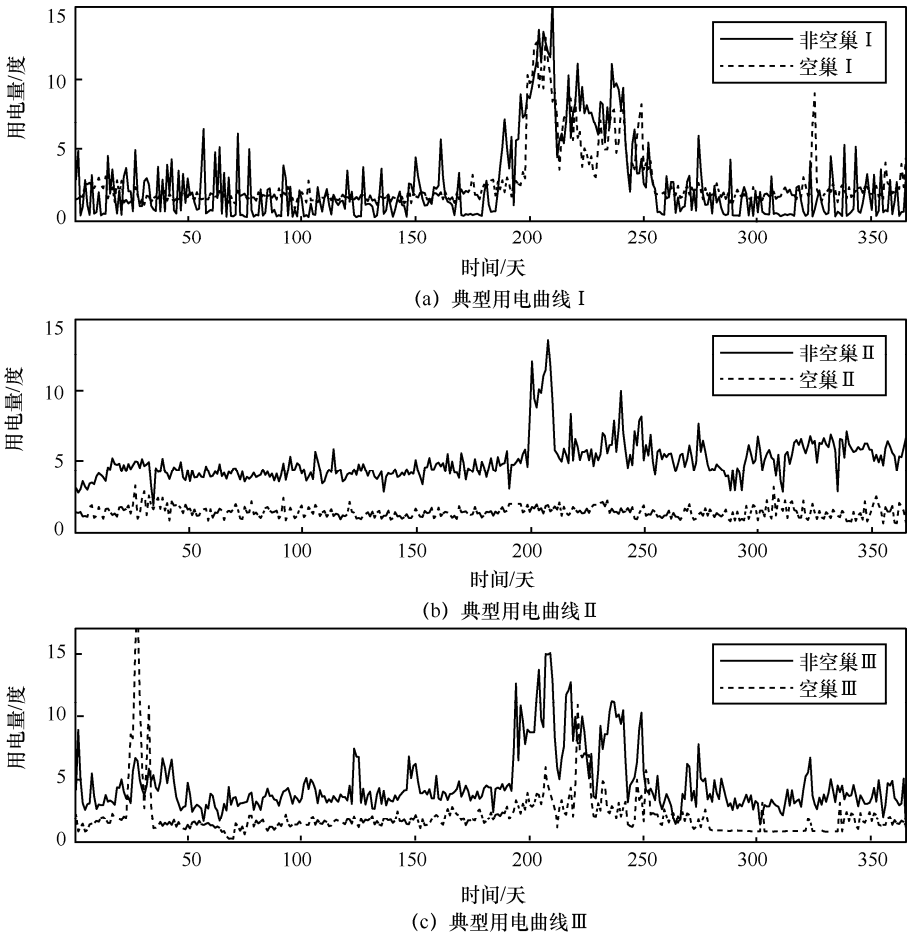

图 1　空巢和非空巢典型用电曲线

空巢和非空巢用户在春节期间（25~40 天）都出现用电上升趋势，但空巢用户在春节期间的用电上升幅度远高于非空巢用户，可见在春节期间，由于空巢家庭的子女、亲人回归，其家庭人口数量突然增加，导致空巢用户用电趋势明显不同于非空巢用户。

基于以上分析，空巢和非空巢用户由于生活行为轨迹、特点和生活质量的不同映射出的用电水平、用电波动、用电趋势也各不相同，其用电特征库也各不相同，所以本文将基于年、季度、月、日用电特征对用电水平、用电波动、用电趋势 3 个方面进行用电特征库构建。

设某用户日用电数据为 $x_d(d=1,\cdots,356)$；月用电数据为 $y_m(m=1,\cdots,12)$；季度用电数据为 $z_q(q=1,\cdots,4)$。在用电水平方面，引入年、季度、月用电量最大值、最小值及差值来表征用户用电水平。

在用电波动方面，采用概率统计中标准差 σ 反映用电数据集的离散和波动程度，但标准差 σ 易受用电水平影响，若用户平均日用电量较大，其标准差 $\sigma(d)$可能也较大[10]，因此引入归一化后的用电波动离散度 Cv 和标准差 σ 共同表征用户用电波动程度：

$$\begin{cases} u=\dfrac{\sum\limits_{d=k_1}^{k_2}x_d}{k_2-k_1+1} \\ \sigma(d)=\sqrt{\dfrac{\sum\limits_{d=k_1}^{k_2}(x_d-u)^2}{k_2-k_1+1}} \end{cases} (1\leqslant k_1<k_2\leqslant 365) \qquad (1)$$

$$Cv(d)=\frac{\sigma}{u} \tag{2}$$

其中，u 为用电时间长度为 k_2-k_1+1 的日用电量均值，$\sigma(d)$为用电时间长度为 k_2-k_1+1 的日用电量标准差，$Cv(d)$为用电时间长度为 k_2-k_1+1 的日用电量离散度。对于月用电量和季度用电量采用相同的处理方法。

在用电趋势方面，引入相邻两月用电量平均用电量差值 T_d 和比值 T_s 来反映用户用电量趋势：

$$\begin{cases} T_d=\left|\bar{y}_m-\bar{y}_{m+1}\right| \\ T_s=\dfrac{\bar{y}_m}{\bar{y}_{m+1}} \end{cases} \tag{3}$$

季度用电量采用相同的处理方法。

通过上述方法，共采用 36 个用电指标对用户用电特征库进行构建，建立用电特征指标库 D，见表 1，其中 D_1~D_{10}、D_{11}~D_{28}、D_{29}~D_{36} 分别为用电水平、用电波动和用电趋势的用电特征库构建指标。

表 1　用户用电特征指标库

用电特征指标	维数
全年用电量最大值、最小值及差值	D_1~D_3
季度用电最大值、最小值及差值	D_4~D_6
月用电最大值、最小值及差值	D_7~D_9
年用电平均值	D_{10}
每季度用电平均值	D_{11}~D_{14}
全年用电标准差、离散系数	D_{15}~D_{16}
每季度用电标准差最大值、最小值及比值	D_{17}~D_{19}
每季度用电离散系数最大值、最小值及比值	D_{20}~D_{22}
每月用电标准差最大值、最小值及比值	D_{23}~D_{25}
每月用电离散系数最大值、最小值及比值	D_{26}~D_{28}
季度用电之差均值和比值的最大值、最小值	D_{29}~D_{32}
月用电之差均值和比值的最大值、最小值	D_{33}~D_{36}

3　加权随机森林算法及评价指标

3.1　传统随机森算法

随机森林（random forest，RF）算法是以 CART 决策树为基础分类器的集成学习模型[10-12]，利用 Bagging 算法有放回地随机抽样生成多个训练集，针对每个训练集构建决策树模型，将多个决策树预测结果进行投票选出最终结果[13-14]。随机森林算法的“随机”主要体现在两方面：采用 Boostrap 自助采样法有放回地从原始数据抽取训练数据和采用随机化方法选取决策树特征，降低决策树之间的相关性，减少随机森林过拟合的风险。

随机森林采用自助采样法和随机化特征子集增强了集成分类的多样性，其相比于单个决策树具有更强的准确性泛化能力和稳健性[15-18]。据调查统计，非空巢与空巢用户比例接近 10:1，两者的比例存在严重数据失衡，数据集中样本的混杂程度较低，造成随机森林容易产生数据敏感化问题[19-20]，其主要原因如下：

- 在基于 Gini（基尼）指数的单棵决策树分裂过程中，决策树输出节点倾向于识别非空巢方向生长，而识别空巢用户方向出现不完全生长或逆生长问题；
- 在随机森林最终投票环节中，赋以每棵决策树的权重相同，忽略了不同决策树的性能差异，导致随机森林无法有效精确地识别出空巢用户。

3.2　加权随机森林算法

为克服随机森林对数据不平衡分布的敏感问题，采用加权随机森林（weighted random forest，WRF）算法来改善随机森林对数据分布的学习能力，其本质是通过增加少数样本的权重，将随机森林的分类策略自适应提高到少数样本中，从而提高机器算法对少数样本的学习和识别能力[21-22]。

本文将权值应用于决策树构建和最终投票两个过程[23-24]，在决策树的构建中，采用加权 Gini 指标寻找最优分裂，构建基于加权 Gini 指标的决策树，在最终预测结果中采用“加权多数投票”决策，将每棵加权决策树投票结果赋以权重，权重为当前加权决策树的袋外数据准确率，最终在所有投票结果中选取投票多数的类别为最终分类输出。其加权随机森林模型结构如图 2 所示。

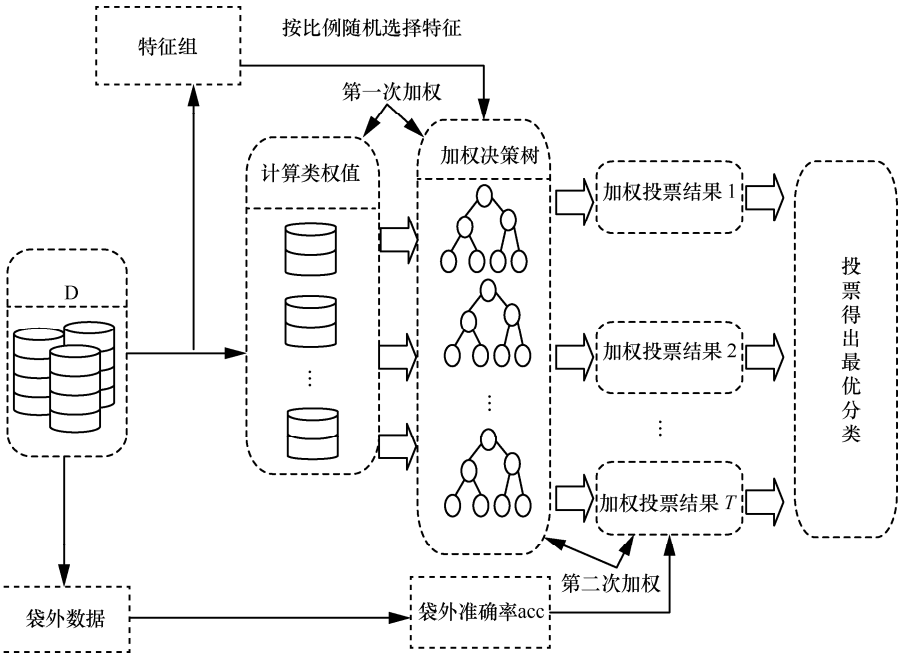

图 2　加权随机森林模型结构

（1）加权 Gini 决策树构建

采用 Boostrap 自助采样法从原始数据中抽取数据集 $D_{N\times M}$，数据集 D 由 N 个训练样本和 M 个特征构成，数据集 D 中每类样本的权重 W_k 与该分类在样本集中出现的频率 P_k（k=1,2,⋯,C）成反比，C 为样本类别个数。

$$W_k=\frac{1}{P_k}(k=1,2,\cdots,C) \tag{4}$$

则样本数据集 D 的加权 Gini 指数 GW(D)为：

$$\begin{cases}\mathrm{GW}(D)=1-\sum_{k=1}^{C}S_k{}^2\\ S_k=\dfrac{I_k\times W_k}{\sum_{k=1}^{C}I_k\times W_k}\end{cases} \tag{5}$$

式（5）中，I_k为样本 D 中类别 k 的个数，S_k为第 k 类样本的加权比例，GW(D)为数据集 D 加权后的基尼值。

设数据集 D 根据特征 a 可被分割成 D_1 和 D_2 两部分，求取 GW(D,a)最小值获得最优划分，构建加权决策树。

$$\mathrm{GW}(D,a)=\frac{D_1}{D}\mathrm{GW}(D_1)+\frac{D_2}{D}\mathrm{GW}(D_1) \tag{6}$$

GW(D,a)最小值即该节点的最优特征，a($a\in M$)为分裂特征。

（2）基于“加权多数投票”决策

由于随机森林在自助采样法中有放回地抽取数据，则每个样本未被抽取到的概率为：

$$P=(1-\frac{1}{N})^N \tag{7}$$

当 $N\to\infty$时，$P\approx0.37$，所以每棵决策树接近37%的数据并未抽取，该数据称为袋外（OBB）数据，则单棵决策树 t（t=1,2,⋯,T）的权重为当前决策树模型在 OOB 数据集上的预测准确率 acc_t，最终随机森林的预测结果为：

$$\begin{cases}t_k=\dfrac{W_k\times\mathrm{leaf}(k)}{\sum_{k=1}^{C}W_k\times\mathrm{leaf}(k)}\\ f_{\mathrm{WRF}}(k)=\dfrac{\sum_{t=1}^{T}\mathrm{acc}_t\times t_k}{\sum_{t=1}^{T}\mathrm{acc}_t}\end{cases} \tag{8}$$

式（5）中 t_k为决策子树 t 预测为类别 k 的概率，leaf（k）为在当前决策子树 t 输出节点上类别 k 的样本数，f_{WRF}（k）为加权投票后预测为类别 k

的概率。

在对测试集用户预测时，若测试样本识别结果为空巢的概率 f_{WRF}（空巢）大于阈值 $\alpha\in(0,1)$ 则判定该用户为空巢用户，否则为非空巢用户。

$$\begin{cases} f_{\mathrm{WRF}}(\text{空巢})>\alpha，\text{判定为空巢用户} \\ f_{\mathrm{WRF}}(\text{空巢})\leqslant\alpha，\text{判定为非空巢用户} \end{cases} \tag{9}$$

3.3　评价指标

对于预测结果，使用二分类评定指标进行精确度计算，表 2 为分类结果混淆矩阵。

表 2　分类结果混淆矩阵

预测	空巢	非空巢
空巢	TP	FP
非空巢	FN	TN

精确率 Pre 表示预测为空巢用户的样本中有多少是真正的空巢用户。

$$\mathrm{Pre}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{10}$$

召回率 Rec 表示样本中的空巢用户有多少被预测正确。

$$\mathrm{Rec}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{11}$$

ROC 曲线是一种检验数据不平衡的评估方法，纵轴是“真正例率（true positive rate，TPR）”，横轴是“假正例率（false positive rate，FPR）”。

$$\mathrm{TPR}=\mathrm{Rec} \tag{12}$$

$$\mathrm{FPR}=\frac{\mathrm{FP}}{\mathrm{TN}+\mathrm{FP}} \tag{13}$$

通过空巢用户的 ROC 曲线可以查出任意阈值 α 对分类器的泛化性能，ROC 曲线下的面积 AUC 用来衡量分类器模型效果的好坏，AUC 值越大，则分类器对空巢用户的精准识别能力越强。

4　实例分析

4.1　数据来源及预处理

利用调查问卷方式对浙江某地区部分电力用户的社会信息进行统计，其社会信息包括用户家庭结构、人数、年龄、家用电器、缴费方式等信息，通过调查问卷获取部分准确空巢老人的标签。对于填写不符合逻辑要求的调查问卷进行剔除，处理后共收集 6 000 份调查问卷。

按调查问卷中户号提取用户 2018 年 1 月 1 日—12 月 31 日的日用电量数据，发现不符合要求的用电数据为残缺数据、重复数据[25]。本文对残缺电量数据大于全年电量数据 30%的用户进行剔除，对残缺电量数据小于全年电量数据 30%的用户采用相邻缺失值线性插值法进行数据补全，利用时间排列法去除重复数据。经过数据预处理后的用户共 5 254 户，按照 2:1 的比例划分为训练集和测试集，见表 3。

表 3　测试集和测试集用户数据

	非空巢	空巢	总和
训练集	3 055	465	3 520
测试集	1 521	213	1 734
总和	4 576	678	5 254

4.2　加权随机森林特征对比及参数选择

将用户用电数据按照式（1）~式（3）构建用电特性识别库 $D_{5\ 254\times36}$，训练集数据为 $D_{3\ 520\times36}$，测试集数据为 $D_{1\ 734\times36}$。本文采用人工选取特征的方法从用电水平、用电波动、用电趋势 3 个方面构建用电特性指标库，为证明该用电特性指标的可行性，引入了基于主成分分析（principal component analysis，PCA）的特征学习方法作为对比，PCA 是一种基于机器学习的数据分析方法，在保留原始数据的信息量的前提下，通过构造一组新的潜隐变量来表征原始数据，常用于提取数据的主要特征分量，为增强对比试验的说服力，设置 PCA 提取的特征维数与用电特征指标库维数相同。通过 PCA 提取后的 36 个特征累计信息贡献率为 85.96%，即新特征保留了原始数据 85.96%的信息量。

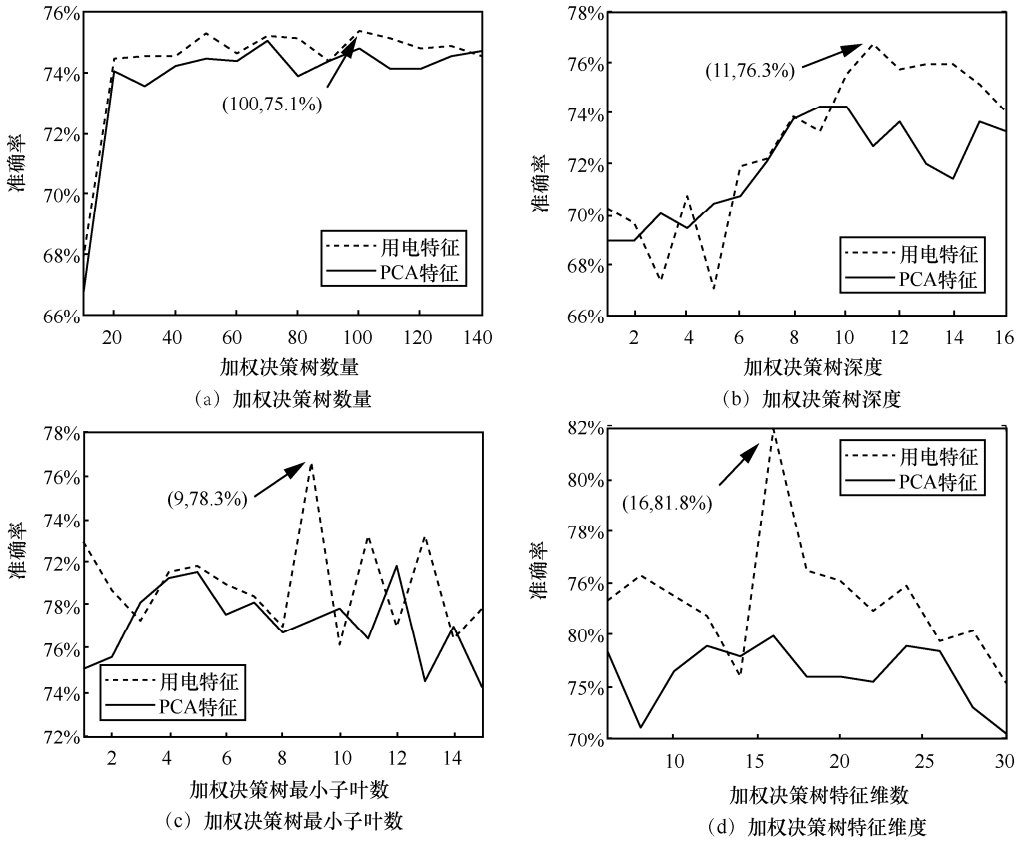

图 3　加权随机森林不同特征在不同参数下的性能

对于加权随机森林的最优模型参数，将训练集数据采取 10 折交叉验证训练，交叉验证下的空巢识别准确率平均值作为评价指标，其加权随机森林不同特征在不同模型参数下的性能如图 3 所示。

图 3 中加权随机森林参数寻优顺序为(a→b→c→d)，即首先通过图 3（a）将最优加权决策树数量设置为 100，再通过图 3（b）将最优加权决策树深度设置为 11，然后通过图 3（c）将最优加权决策树最小子叶数设置为 9，最后通过图 3（d）将最优加权决策树特征维数设置为 16，此时加权决策树对空巢用户的识别准确率最高为 81.8%。并且从图 3 中可见，本文构建的用电特征指标要优于基于 PCA 的机器学习特征指标，其不同用电特征对加权随机森林分类器性能贡献率如图 4 所示。

由图 4 分析可得，不同特征指标的贡献率较为相似，每个特征对分类器都有贡献作用，特征间的相关性较低，并未出现某类或某几类特征贡献率较高而导致其余特征为无用特征的情况，从侧面说明了本文选取用电特征的可行性。

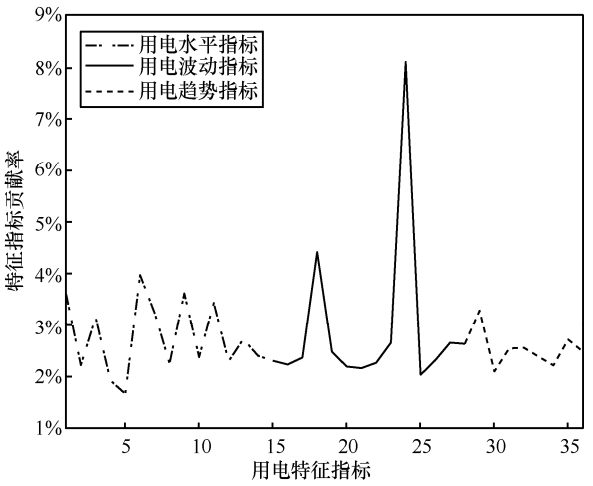

图 4　不同用电特征对加权随机森林分类器性能贡献率

4.3　加权随机森林与其他算法性能比较

本文采用支持向量机（support vector machine，SVM）、分类回归树（classification and regression tree，CART）、随机森林与加权随机森林算法进行对比，首先基于训练集数据构建分类器模型，在 10 折交叉验证下寻找各算法的最优参数，然后对测试集进行预测，最后采用分类结果的 ROC 曲线和 AUC 作为分类器性能评价指标，其对比结果分别如图 5 和图 6 所示。

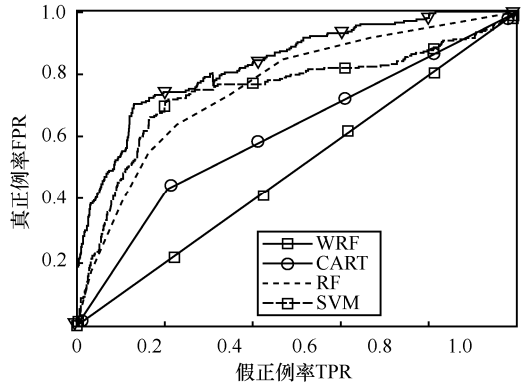

图 5　不同算法下空巢用户识别 ROC 曲线

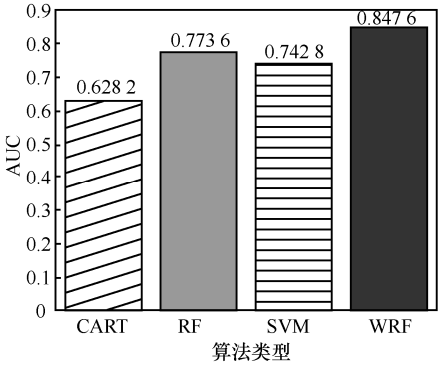

图 6　不同算法下空巢用户识别 AUC 值

图 5 中 WRF 的 ROC 曲线位于其他算法的上方，图 6 中 WRF 的 AUC 值为 0.847 6，均大于其他算法的 AUC 值。由于 WRF 对非平衡数据的处理性能优于 CART、SVM、RF 算法，对空巢用户的识别能力更强，因此下文只给出了加权随机森林的结果参数。

其多数投票阈值 α 与 WRF 分类器对空巢用户识别泛化性能曲线如图 7 所示。

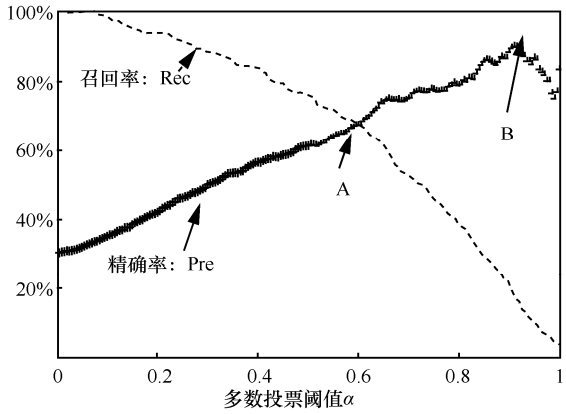

图 7　WRF 分类器对空巢用户识别泛化性能曲线

由图 7 可见，对空巢用户进行预测的精确率 Pre 随阈值 α 的增大而增大，召回率 Rec 随 α 增大而减小，其 A、B 两点的指标见表 4。

在 A 点 α=0.595 时，Pre=Rec=67.1%，FN 与 FP 用户数量相等，即 WRF 预测出空巢用户数等于测试集样本中空巢用户数量，预测空巢用户共 213 户，预测正确为 142 户，预测错误为 71 户。

表 4　WRF 分类器空巢用户预测结果

	α	精确率	召回率
A	0.595	67.1%	67.1%
B	0.912	93.4%	14.2%

在 B 点 α=0.91 时，对空巢用户预测精确率 Pre 达到最高，为 93.4%，预测空巢用户共 30 户，预测正确为 28 户，预测错误为 2 户。

4.4　采用加权随机森林算法对未知类型用户识别结果

基于以上分析，本文采用 WRF 算法建立空巢用户识别模型，最优多数投票阈值 α 设为 0.595，并将该模型在浙江省某电力公司的用户采集系统中部署上线。利用已上线模型对空巢用户识别时，需要先确定识别区域，然后将用电信息采集系统中该区域待识别用户整年用电数据输入 WRF 空巢用户识别模型中，最后输出疑似空巢用户名单，再由工作人员对疑似空巢用户名单进行现场查证，从而排除名单中的非空巢用户。上线的识别模型需要通过定期更新用户数据，调整模型参数，

从而保证识别模型的实时准确性。空巢用户精准识别系统界面如图 8 所示。

图 8　空巢用户精准识别系统界面

利用空巢用户精准识别系统对浙江省金华市某地区 2 000 户未知类型用户进行识别，识别出 140 户疑似空巢老人群体，经电力公司有关部门实地调查，其中准确识别的空巢用户为 104 户，错误识别为 36 户，其精确率 Pre 为 74.2%。如果按照空巢用户比例为 10%的原则，其空巢召回率 Rec 为 52%。

识别错误的原因主要分为两个方面：一是非空巢用户与空巢用户用电规律基本相同，导致将该用户预测成空巢用户；二是用电特征提取并未全面或只从用电角度并不能完整构建空巢用户特征库，从而导致部分非空巢用户被预测为空巢用户。

5　结束语

由实例分析可见，本文提出的基于用电特征库的空巢老人识别方法是可行的。电力公司通过利用用电信息采集系统中的数据实现对空巢用户的精准识别，一方面可协助政府对空巢老人用户进行帮扶，能够从电能使用角度对其日常生活和用电行为进行实时监测，更好地规避用电安全风险；另一方面该方法也可以完善现有空巢用户的识别方法，与实地调查、居委统计等方式相比，可节省人力物力的投入。

本文通过用电特征库识别模型对空巢用户精准识别，其效率更高、范围更广，但只从用电角度并不能完整构建空巢用户特征库，若加入电力部门的营销数据可更加全面地构建用户用电特征库，例如用户电费的缴费渠道、欠费记录、电表注册信息、户主信息等，这些较为丰富的用户个人信息有利于增强群体分类依据的可解释性，提高空巢用户识别的精准度。

参考文献：

[1] 翟振武, 陈佳鞠, 李龙. 2015—2100 年中国人口与老龄化变动趋势[J]. 人口研究, 2017, 41(4): 60-71.
ZHAI Z W, CHEN J J, LI L. Future trends of China's population and aging: 2015—2100[J]. Population Research, 2017, 41(4): 60-71.

[2] 冯先成, 李寒, 周密, 等. 基于前馈神经网络的智慧城市空巢老人识别[J]. 武汉工程大学学报, 2015, 37(10): 33-39.
FENG X C, LI H, ZHOU M, et al. Recognition of empty-nest elders in intelligent city based on feedforward neural network[J]. Journal of Wuhan Institute of Technology, 2015, 37(10): 33-39.

[3] 崔瀚文, 栾石圳南, 李远帆, 等. 空巢老人手机用户的精准识别[J]. 数学建模及其应用, 2014, 3(1): 49-62.
CUI H W, LUAN S Z N, LI Y F, et al. Accurate identification approach to empty-nester mobile-phone users[J]. Mathematical Modeling and its Applications, 2014, 3(1): 49-62.

[4] 吴郅君, 殷新博, 陈中, 等. 基于模糊聚类曲线相似度的负荷用户识别方法[J]. 电力工程技术, 2019, 38(3): 151-156.
WU Z J, YIN X B, CHEN Z, et al. Identification method of load customers based on similarity of fuzzy clustering curves[J]. Electric Power Engineering Technology, 2019, 38(3): 151-156.

[5] 和敬涵, 卢育梓, 陆金耀, 等. 基于“进化”主成分分析法的用户分类及其应用[J]. 电力建设, 2017, 38(3): 101-107.
HE J H, LU Y Z, LU J Y, et al. User classification method based on 'evolution' PCA and its application[J]. Electric Power Construction, 2017, 38(3): 101-107.

[6] 辛苗苗, 张延迟, 解大. 基于电力大数据的用户用电行为分析研究综述[J]. 电气自动化, 2019, 41(1): 1-4, 27.
XIN M M, ZHANG Y C, XIE D. Summary of researches on consumer behavior analysis based on big power data[J]. Electrical Automation, 2019, 41(1): 1-4, 27.

[7] 李如初, 沈名龙, 彭海棠. 大数据、云计算在电力工业中的应用[J]. 电信科学, 2018, 34(4): 151-155.
LI R C, SHEN M L, PENG H T. Application of big data and cloud computing in power industry[J]. Telecommunications Science, 2018, 34(4): 151-155.

[8] 谷红勋, 杨珂. 基于大数据的移动用户行为分析系统与应用案例[J]. 电信科学, 2016, 32(3): 139-146.
GU H X, YANG K. Mobile user behavior analysis system and applications based on big data[J]. Telecommunications Science, 2016, 32(3): 139-146.

[9] 王志宏, 杨震. 人工智能技术研究及未来智能化信息服务体系的思考[J]. 电信科学, 2017, 33(5): 1-11.
WANG Z H, YANG Z. Research on artificial intelligence technology and the future intelligent information service architecture[J]. Telecommunications Science, 2017, 33(5): 1-11.

[10] 蔡一鸣. 几种方差概念的比较[J]. 统计与信息论坛, 2008(4): 20-22.
CAI Y M. Comparison on diverse concepts of variances[J]. Statistics & Information Forum, 2008(4): 20-22.
[11] 张棪, 曹健. 面向大数据分析的决策树算法[J]. 计算机科学, 2016, 43(S1): 374-379, 383.
ZHANG Y, CAO J. Decision tree algorithms for big data analysis[J]. Computer Science. 2016, 43(S1): 374-379, 383.
[12] 孙梦婷, 魏海平, 李星澄, 等. 利用 CART 分类树分类检测交通拥堵点[J]. 武汉大学学报(信息科学版), 2019(11): 1-10.
SUN M T, WEI H X, LI X Y, el at. Detection and classification of traffic congestion points using CART classification tree[J]. Geomatics and Information Science of Wuhan University, 2019(11): 1-10.
[13] 刘玉茹, 赵成萍, 臧军, 等. CART 分析及其在故障趋势预测中的应用[J]. 计算机应用, 2017, 37(S2): 57-59, 73.
LIU Y R, ZHAO C P, ZANG J, et al. Analysis of CART regression tree and its application in fault trend forecasting[J]. Journal of Computer Applications, 2017, 37(S2): 57-59, 73.
[14] XU Z Y, KANG Y, CAO Y, et al. Man-machine verification of mouse trajectory based on the random forest model[J]. Frontiers of Information Technology & Electronic Engineering, 2019, 20(7): 925-929.
[15] 刘亚丽, 李国栋, 刘云, 等. 基于随机森林的电动汽车充电行为聚类技术研究[J]. 电力工程技术, 2019, 38(6): 115-121.
LIU Y L, LI G D, LIU Y, et al. Clustering technology of electric vehicle charging behavior based on random forest[J]. Electric Power Engineering Technology, 2019, 38(6): 115-121.
[16] 吴涛, 刘韬, 王斌. 安徽联通企业级大数据平台构建及应用实践[J]. 电信科学, 2018, 34(1): 135-147.
WU T, LIU T, WANG B. Construction and application of Anhui Unicom enterprise big data platform[J]. Telecommunications Science, 2018, 34(1): 135-147.
[17] 王铮, 任华, 方燕萍. 随机森林在运营商大数据补全中的应用[J]. 电信科学, 2016, 32(12): 7-12.
WANG Z, REN H, FANG Y P. Application of random forest in big data completion[J]. Telecommunications Science, 2016, 32(11): 7-12.
[18] 王德文, 孙志伟. 电力用户侧大数据分析与并行负荷预测[J]. 中国电机工程学报, 2015, 35(3): 527-537.
WANG D W, SUN Z W. Big data analysis and parallel load forecasting of electric power user side[J]. Proceedings of the CSEE, 2015, 35(3): 527-537.
[19] 李艳霞, 柴毅, 胡友强, 等. 不平衡数据分类方法综述[J]. 控制与决策, 2019, 34(4): 673-688.
LI Y X, CHAI Y, HU Y Q, et al. Review of imbalanced data classification methods[J]. Control and Decision, 2019, 34(4): 673-688.
[20] 张宏莉, 鲁刚. 分类不平衡协议流的机器学习算法评估与比较[J]. 软件学报, 2012, 23(6): 1500-1516.
ZHANG H L, LU G. Machine learning algorithms for classifying the imbalanced protocol flows: evaluation and comparison[J]. Journal of Software, 2012, 23(6): 1500-1516.
[21] 吴成东, 卢紫微, 于晓升. 基于加权随机森林的图像超分辨率算法研究[J]. 控制与决策, 2019, 34(10): 2243-2248.
WU C D, LU Z W, YU X S. Image super resolution reconstruction algorithm based on weighted random forest[J]. Control and Decision, 2019, 34(10): 2243-2248.
[22] 魏正韬, 杨有龙, 白婧. 基于非平衡数据的随机森林分类算法改进[J]. 重庆大学学报, 2018, 41(4): 54-62.
WEI Z T, YANG Y L, BAI J. An improved random forest algorithm based on unbalanced data[J]. Journal of Chongqing University, 2018, 41(4): 54-62.
[23] 彭微, 王灵矫, 郭华. 基于随机森林的文本分类并行化[J]. 计算机科学, 2018, 45(12): 148-152.
PENG Z, WANG L J, GUO H. Parallel text categorization of random forest[J]. Computer Science, 2018, 45(12): 148-152.
[24] 常玉清, 孙雪婷, 钟林生, 等. 基于改进随机森林算法的工业过程运行状态评价[J]. 自动化学报, 2019(11): 1-10.
CHANG Y Q, SUN X T, ZHONG L S, et al. Industrial operation performance evaluation of industrial processes based on modified random forest[J]. Acta Automatica Sinica, 2019(11): 1-10.
[25] 姜红红, 张涛, 赵新建, 等. 基于大数据的电力信息网络流量异常检测机制[J]. 电信科学, 2017, 33(3): 134-141.
JIANG H H, ZHANG T, ZHAO X J, et al. A big data based flow anomaly detection mechanism of electric power information network[J]. Telecommunications Science, 2017, 33(3): 134-141.

[作者简介]

卢子萌（1995– ），男，中国计量大学机电工程学院硕士生，主要从事电力系统大数据分析研究工作。

陈佳怡（1993– ），女，国网金华供电公司工程师，主要从事营销服务与综合能源创新优化工作。

李璟（1977– ），女，博士，中国计量大学机电工程学院副教授，主要从事数值分析、数据处理数据挖掘等算法研究工作。

谢岳（1964– ），男，博士，中国计量大学机电工程学院教授，主要研究方向为检测技术与自动化装置。

蒋欣利（1993– ），男，国网金华供电公司助理工程师，主要从事营销计量与大数据分析工作。

韩蕾（1979– ），女，浙江华云信息科技有限公司中级工程师，主要从事电力营销信息化工作。

郭倩（1987– ），女，博士，中国计量大学机电工程学院讲师，主要从事电力新能源分布式发电及控制技术与电力大数据分析工作。