DOI:10. 3969/j. issn. 1674-1951. 2021. 08. 008

基于多参数代价敏感系数学习及数据驱动模型的电力能耗预测

Power consumption prediction based on multi-parameter cost-sensitive coefficient learning and data-driven model

施杰,张安勤*
SHI Jie,ZHANG Anqin*

(上海电力大学 计算机科学与技术学院,上海 200090)
(College of Computer Science and Technology,Shanghai University of Electric Power,Shanghai 200090,China)

摘 要:企业节能减排是我国社会发展的前沿问题和研究热点,对企业用户的电力能耗现状进行综合分析与评估是进行节能改造或节能设计的前提和基础。阐述了现阶段常用的电力能耗预测方法,分析了分类与回归树(CART)算法作为弱学习器构建预测模型的缺点,针对原始AdaBoost算法只关注预测误差率最小的不足,在算法实质基础上研究并提出一种基于多参数代价敏感系数学习的改进AdaBoost算法。建立基于改进AdaBoost算法的回归预测模型,通过真实数据进行短期电力能耗预测,验证了改进算法对模型性能的提升。

关键词:电力能耗预测;代价敏感系数;数据驱动;分类与回归树算法;AdaBoost算法
中图分类号:TM 74 **文献标志码**:A **文章编号**:1674-1951(2021)08 - 0054 - 07

Abstract: Energy conservation and emission reduction in enterprises are the frontier issues and research hot-spots in the way of China's development. Comprehensive analysis and evaluation on the current situation of power consumption in enterprises is the premise and foundation for energy-saving transformations or energy conservation designs. Having described the common forecasting methods for electricity consumption, the shortcomings of constructing a prediction model taking Classification and Regression Tree (CART) algorithm as the weak learner are analyzed. To deal with the deficiency of the traditional AdaBoost algorithm focusing on the minimum prediction error rate only, an improved AdaBoost algorithm based on multi-parameter cost-sensitive coefficient learning is studied and proposed based on the essence of the algorithm. A regression prediction model constructed based on the improved AdaBoost algorithm can make short-term power consumption prediction according to real data, which verifies the improvement of the model's performance.

Keywords: electricity consumption forecast; cost-sensitive coefficient; data-driven; Classification and Regression Tree algorithm; AdaBoost algorithm

0 引言

目前,我国正处在工业化与城镇化快速发展的阶段,居民消费水平不断提高,对能源的需求量越来越大。但目前我国经济增长方式仍呈现投入高、消耗高、污染高等粗放型特点,加剧了能源供求的矛盾[1]。与全球发达国家相比,我国现阶段电能利用效率还处在较低的水平[2-3],因此,我国政府在近年来的多个五年规划纲要中提出了单位国内生产总值能耗降低20%左右、主要污染物排放总量减少10%的约束性指标[4],以努力创造能源节约型企业模式。

电力能耗预测指以企业用户历史电力数据为基础,利用数学统计方法或智能模型预测方法对未来电力使用情况进行预测。准确合理的电力能耗预测对企业用户、供电公司等单位合理分配电力资源,避免电力能源浪费具有重要的现实意义[5-7]。近年来,随着我国智能电网的建设和发展,电力企业积累了海量的用电数据,为开展电力能耗精细化预测提供了丰富的数据基础。一方面,基于企业用户电力能耗数据分析结果,可对用户用电特征开展多维度分解,并建立有较强针对性的预测模型,以提高用户用电量预测的精确度;另一方面,智能电网

收稿日期:2020-09-11;**修回日期**:2021-05-20
基金项目:国家自然科学基金项目(61772327)

系统下企业用户积累的电力能耗数据体量大、类型多、维度高,使用原始的电力能耗预测方法存在一定的局限性,难以准确把握企业用户用电量变化规律及关联因素。因此,基于海量历史电力能耗数据开展短期内企业用户用电量预测,也是摆在国内外研究者面前的一个难题。从20世纪50年代开始,国外研究人员以及电力领域从业人员就已经开始了电力能耗预测工作,随着我国电力市场制度的逐渐完善,电力能耗预测逐渐成为研究热点并取得了一定的研究成果,涌现出多种新颖的预测方法,但绝大部分方法仍处于研究试验阶段,并未真正投入工程应用。电力能耗预测方法主要分为以下2类。

(1)基于数学模型的电力能耗预测。基于数学模型的电力能耗预测主要是从历史用电数据中寻找规律,通过分析影响历史用电数据的主要因素,建立对应的数学统计模型,主要实现方法有基于时间序列、基于回归分析等2种。基于时间序列的方法可依据时间因果关系,将历史用电负荷依照时间先后顺序进行排序,最终在用户历史用电需求与未来用电需求之间建立预测模型;基于回归分析的方法可通过历史用电数据及影响用电的因素来确定多元线性回归模型的参数,对电力能耗做出预测。其中,基于回归分析的方法由于原理简单、易于理解而广泛应用于早期电力能耗预测中。

(2)基于数据驱动的电力能耗预测。随着复杂电力系统的发展和应用,电力能耗逐渐具有不确定性、非线性和时变性等特点,难以采用固定的数学方法来表达输入与输出之间的关系,基于回归分析的电力能耗预测方法已无法适应现阶段需求。另外,随着计算机技术以及人工智能、大数据等新兴技术的快速发展,研究基于数据驱动的电力能耗预测技术成为可能。数据驱动方法以海量历史电力能耗数据为基础,运用智能化手段构建数据模型,对未来短期内电力能耗进行预测。目前,机器学习、数据挖掘、专家系统和模糊逻辑系统等人工智能技术得到了广泛应用,涉及的算法包括提升(Boosting)、支持向量机(Support Vector Machine, SVM)、分类与回归树(Classification and Regression Tree, CART)、极限梯度提升(eXtreme Gradient Boosting, XGBoost)、人工神经网络等;同时,非线性系统理论、组合分析法、深度学习等其他技术方法也被逐渐提出。其中,神经网络具备学习能力较强、网络结构简单、容错能力强等特点,在构建预测模型过程中最受欢迎,但神经网络模型的自由度会随着模型复杂度的增加而降低,易导致预测模型出现过拟合或欠拟合等问题[8-9]。

本文根据企业用户原始用电能耗统计数据量大的特点,考虑影响电力能耗的时间因素,基于数据驱动进行电力能耗预测。首先分别采用CART算法、自适应增强(Adaptive Boosting, AdaBoost)算法构建电力能耗预测模型,对比分析CART算法以及AdaBoost算法的不足,在AdaBoost算法基础上提出一种基于多参数代价敏感系数学习的改进AdaBoost算法,构建基于数据驱动的电力能耗预测模型。以江苏省某高新技术企业近2年的用电量数据进行模型训练、迭代、优化,提升电力能耗预测的准确率,同时验证上述方法的科学性和有效性。可为未来短期内企业用户电力能耗预测提供参考,并为供电公司合理安排电力调配提供决策依据。

1　电力能耗数据描述及处理

本文采用的电力能耗数据主要来源于江苏省某高新技术企业近2年的用电量,由于原始数据集包含记录时间(record_data)和电力能耗数据(power_consumption),因此,对原始数据进行整合、处理,构建以天为单位的样本数据集,共涵盖300多万个样本数据。此外,针对样本数据集,通过增加时间特征、周末特征等信息,构造出一份时间连续、涵盖多特征的完整数据集并进行独热编码,形成符合本文分析建模的数据集。

原始数据集见表1,增加时间、周末特征等信息后形成的数据集见表2,进行独热编码后形成的数据集见表3。表中字段名称及含义:record_date,记录日期;power_consumption,电力能耗数据;dow,一个星期中的星期几(0代表星期一,1代表星期二,以此类推);dom,一个月中的第几天;month,一年中的第几个月份;year,今天是哪一年;weekend,是不是周末(0不是周末,1是周末);weekend_sat,是不是周六(0不是周六,1是周六);weekend_sun,是不是周日(0不是周日,1是周日);week_of_month,当月第几周;period_of_month,当月的上中下旬(1是上旬,2是中旬,3是下旬);period2_of_month,当月的上下半月(1是上半月,2是下半月);festival,是不是法定节假日(0不是法定节假日,1是法定节假日)。

表1　原始数据集

Tab. 1　Original data set

record_date	power_consumption/(kW·h)
2015-01-01	2 900 575
2015-01-02	3 158 211
2015-01-03	3 596 487
2015-01-04	3 939 672
2015-01-05	4 101 790

表 2 增加时间、周末特征等信息后形成的数据集

Tab. 2 Data set formed by adding information such as time and weekend feature

record_date	power_consumption/(kW·h)	dow	dom	month	year	weekend	weekend_sat	weekend_sun	week_of_month	period_of_month	period2_of_month	festival
2015-01-01	2 900 575	3	1	1	2015	0	0	0	1	1	1	0
2015-01-02	3 158 211	4	2	1	2015	0	0	0	1	1	1	0
2015-01-03	3 596 487	5	3	1	2015	1	1	0	1	1	1	0
2015-01-04	3 939 672	6	4	1	2015	1	0	1	1	1	1	0
2015-01-05	4 101 790	0	5	1	2015	0	0	0	4	3	2	0

表 3 独热编码数据集

Tab. 3 One-hot encoding data set

festival	dow_0	dow_1	dow_2	dow_3	dow_4	dow_5	...	period_of_month_1	period_of_month_2	period2_of_month_1	period2_of_month_2
0	0	0	0	1	0	0	...	1	0	1	0
0	0	0	0	0	1	0	...	1	0	1	0
0	0	0	0	0	0	1	...	1	0	1	0
0	0	0	0	0	0	0	...	1	0	1	0
0	1	0	0	0	0	0	...	0	1	0	1

本文将数据集拆分为训练数据和测试数据，其中 2015-01-01—2019-08-31 的数据为训练数据，2019-09-01—30 的数据为测试数据。

2 基于 CART 算法的电力能耗预测

2.1 CART 算法原理

CART 算法是由 Breiman 等人在 1984 年提出的，其目的是构造一个有效的算法，让其能从观测的训练样本中给出分类器的分段常数估计或回归函数估计[10]。CART 算法基于最小二乘法，通过在训练数据集中递归地将每个区域划分成 2 个子区域并决定每个子区域上的输出值，构建形成二叉决策树。CART 算法描述如下。

第 1 步：算法输入训练集表示为 $D=\left(x_i,y_i\right)_{i=1}^{m}$，其中 $x_i\in x\subseteq \mathbf{R}^n$，选择数据集 D 中最优切分变量 x_i 与切分点 k，求解

$$\min_{i,k}=\left[\min_{c_1}\sum_{x_j\in R_1(i,k)}\left(y_j-c_1\right)^2+\min_{c_2}\sum_{x_j\in R_2(i,k)}\left(y_j-c_2\right)^2\right],\quad(1)$$

式中：R_m 为被二元划分的 2 个输入空间；c_m 为区域 R_m 上所有输入实例 x_i 对应输出 y_i 的均值。

遍历 x_i，在切分变量 x_i 上扫描切分点 k，找到使上式结果达到最小值的有序数对 (i,k)。

第 2 步：基于找出的有序数对 (i,k)，重新划分训练数据集区域并输出相应的值

$$R_1(i,k)=\left\{x|x^i\leqslant k\right\},R_2(i,k)=\left\{x|x^i>k\right\},\quad(2)$$

式中：R_n 为定义域。

解得 CART 回归生成树的枝结点、叶节点的因变量预测值 $\bar{c}_n$

$$\bar{c}_n=\frac{1}{M_n}\sum_{x_j\in R_n(i,k)}y_j,x\in R_n,n=1,2,\quad(3)$$

式中：M_n 为区域 R_n 上的所有样本数。

第 3 步：对形成的 2 个子区域中的枝结点，重复前 2 步操作并在满足条件时终止。

第 4 步：输出 CART 回归生成树 $f(x)$

$$f(x)=\sum_{n=1}^{N}\bar{c}_n\times I(x\in R_n),\quad(4)$$

式中：N 表示将输入空间划分为 N 个区域；$\bar{c}_n$ 为所在区域 R_n 的输出值的均值；$x\in R_n$ 用于判断输出值是否位于区域 R_n 中，当结果为真时，$I(x\in R_n)$ 取 1，否则取 0。

$f(x)$ 将训练数据集空间划分为 n 个区域 $R_j(j=1,2,\cdots,n)$，包含 n 个叶子节点（n 个类）。

2.2 基于 CART 算法的电力能耗预测

首先基于前文构建的数据集，建立基于 CART 算法的电力能耗预测模型，主要调用 python 的 sklearn. tree 库实现 CART 回归预测。sklearn 是一个机器学习的第三方库，整个库中提供了多类型算法，回归生成树（Tree）模块就是其中之一。本文构建基于 CART 算法的电力能耗预测模型，并对 2019-09-01—30 的电力能耗进行预测，同时与实际数据进行对比，如图 1 所示。

由图 1 可知，基于 CART 算法模型预测的电力能耗数值范围为 3 400 000~4 400 000 kW·h，其中 9 月中旬的预测值与实际值相差较大，整体决定系数 R^2 为 0. 376 1（R^2 越接近 1 说明模型拟合效果越好，越接近 0 说明模型拟合效果越差），所以基于 CART 算法构建的电力能耗预测模型整体拟合效果较差。

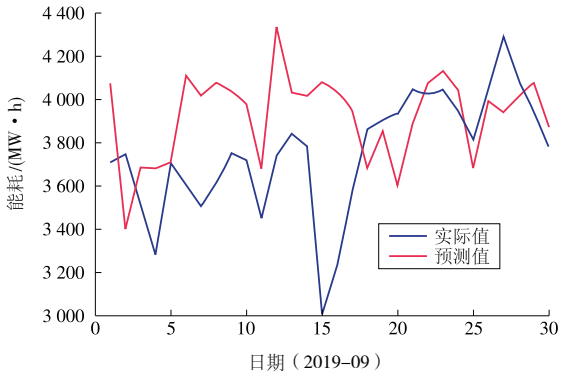

图 1　CART 算法预测的电力能耗与实际值对比

Fig. 1　Comparison of the power consumption predicted by CART algorithm with its actual value

通过分析,CART 算法主要存在以下不足。

(1)CART 算法易过拟合,导致泛化能力不强。

(2)CART 算法易因样本发生改动,导致树结构剧烈改变。

(3)寻找最优的决策树是一个非确定性多项式(NP)难问题,CART 算法易陷入局部最优。

(4)某些特征样本比例过大时,生成的决策树易偏向于这些特征。

3　基于 Boosting 算法的电力能耗预测

Boosting 算法是集成学习(Ensemble Learning, EL)算法的一类,Friedman 在 2000 年左右发表论文时提出 Boosting 算法,并成功地实现应用[11-12]。Boosting 算法可以将弱学习器算法提升为强学习器算法,即先用初始训练集训练出一个弱学习器,再根据弱学习器的表现对训练样本分布进行调整。对于弱学习器判断错误的样本,在之后的训练过程中将赋予更大的关注度(权值),并基于调整后的样本来训练下一个弱学习器,经过反复迭代训练,直至达到理想效果。本文算法的改进主要基于 Boosting 中常用的 AdaBoost 算法。

3.1　AdaBoost 算法

上文分析结果表明,基于 CART 算法构建的电力能耗预测模型拟合效果较差且存在多处不足。为解决以上问题,本文基于 AdaBoost 算法来构建数据驱动的电力能耗预测模型。目前,AdaBoost 回归问题有许多变种,本文主要采用 Adaboost R2 回归算法。AdaBoost 算法描述如下。

算法输入训练集表示为 $D=\left(x_i,y_i\right)_{i=1}^{m}$,其中 $x_i\in x\subseteq \mathbf{R}^n$,学习算法表示为 ℓ,弱学习器个数为 T。

第 1 步:初始化训练样本的权值分布 D_1

$$D_1=\left(w_{11},w_{12},\cdots,w_{1i},\cdots,w_{1M}\right), \tag{5}$$

$$w_{1i}=1/M,i=1,2,\cdots,M, \tag{6}$$

式中:M 为训练样本数量;w_{1i} 为训练样本的初始权重。

第 2 步:模型迭代,其迭代轮次为

$$k=1,2,\cdots,K, \tag{7}$$

式中:K 为模型迭代轮次。

第 3 步:使用具有当前分布 D_k(D_k 为第 k 轮迭代训练样本的权值分布)的训练数据集训练弱学习器,弱学习器表示为

$$G_k=\ell\left(D,D_k\right)。 \tag{8}$$

(1)针对误差率问题,第 k 个弱学习器在训练集上样本最大误差表示为

$$E_k=\max\left|y_i-G_k\left(x_i\right)\right|,i=1,2,\cdots,N。 \tag{9}$$

计算每个样本的相对误差,并将误差数值规范到[0,1]之间。

线性误差,相对误差表示为

$$e_{ki}=\frac{\left|y_i-G_k\left(x_i\right)\right|}{E_k}。 \tag{10}$$

平方误差,相对误差表示为

$$e_{ki}=\left(y_i-G_k\left(x_i\right)\right)^2E_k^2。 \tag{11}$$

指数误差,相对误差表示为

$$e_{ki}=1-\exp\left\{-\frac{\left|y_i-G_k\left(x_i\right)\right|}{E_k}\right\}。 \tag{12}$$

最终,第 k 个弱学习器的误差率表示为 ε_k

$$\varepsilon_k=\sum_{i=1}^{N}w_{ki}e_{ki}, \tag{13}$$

式中:w_{ki} 为训练样本第 k 轮迭代的样本权重。

(2)算法中弱学习器权重系数表示为 α_k。

$$\alpha_k=\frac{\varepsilon_k}{1-\varepsilon_k}。 \tag{14}$$

同时,算法将更新训练集样本分布 D_{k+1}

$$D_{k+1}=\left(w_{k+1,1},w_{k+1,2},\cdots,w_{k+1,i},\cdots,w_{k+1,m}\right), \tag{15}$$

$$w_{k+1,i}=\frac{w_{k,i}}{Z_{\mathrm{K}}}\alpha_k^{\,1-e_{ki}}, \tag{16}$$

$$Z_{\mathrm{K}}=\sum_{i=1}^{N}w_{k,i}\alpha_k^{\,1-e_{ki}}, \tag{17}$$

式中:Z_{K} 为规范因子。

(3)最后,算法将采取对加权弱学习器取中位数的策略构建弱学习器线性组合,得到并输出最终的强学习器 $f(x)$

$$f(x)=\sum_{k}^{N}\alpha_kG_k(x), \tag{18}$$

$$f(x)=\sum_{k}^{N}\left(\ln\frac{1}{\alpha_k}\right)G_k(x)。 \tag{19}$$

3.2 AdaBoost算法加权方式分析

针对AdaBoost算法中各弱学习器权值的调整是提高算法模型性能的重点和难点，原始AdaBoost算法权值调整依据主要参考当前弱学习器产生的预测误差率。主要过程如下。

(1)算法赋予训练数据集D中每个实例相同的权值，并应用弱学习算法，基于加权的训练集合学习出一个弱学习器。

(2)根据弱学习器的预测结果对每个样本重新赋权，赋权的原则是增加上次预测误差率大的样本权值，减小预测误差率小的样本权值。

(3)选择下一轮训练样本时，权值越高的样本被重复选择的概率越高。一旦新选取的训练集与原始的训练集大小相同，便在新训练集上学习一个新的弱学习器。如此循环迭代直至预测误差率大于等于一定的阈值(通常为0.5)。

由于训练生成的弱学习器都是建立在重新加权的数据集上的，并且更专注于对那些权值高的样本进行预测，因此AdaBoost算法实质上提供了一种方法来生成一系列互补的弱学习器，最后通过加权组合构成集成学习算法。

3.3 改进AdaBoost算法

原始AdaBoost算法基于当前弱学习器在训练数据集上预测误差率最低来选择弱学习器，存在因预测误差率较大而造成此后更新样本权值带来较大损失的问题，因此本文在调整权重分配时针对不同样本引入多参数代价敏感系数，通过设置代价敏感系数，使原始算法由关注分类误差率最小转而关注误差代价最小[13]。本文在回归预测中引入多参数代价敏感系数C_i和G_i，以提升模型的性能。

AdaBoost算法在抽取训练子集时，样本权重与被抽中的概率成正比，且在$0<\varepsilon_k<0.5$的情况下，a_k取值范围为$0<\alpha_k<1.0$，因此，以a_k为底的指数函数为单调递减函数。本文在引入不同相对误差代价系数C_i和G_i的基础上，按照被上一轮弱学习器对样本预测相对误差代价较小、样本预测相对误差代价较大、样本预测相对误差代价大等情况进行加权，选中正确预测损失大样本的弱学习器，并最终组合形成强学习器。

改进的算法思路如下。

第1步：初始化训练样本的权值分布，同式(5)。

第2步：迭代轮次k=1,2,…,K(弱学习器数量)，同式(7)；使用具有当前分布D_k的训练数据集训练弱学习器，同式(8)。

计算训练集上的样本最大误差E_k，同式(9)；计算每个样本的相对误差e_{ki}，同式(10)—(12)。

计算弱学习器G_k在训练数据集上的预测误差率ε_k，同式(13)。

第3步：当弱学习器预测错误率大于0且小于0.5时，按照式(15)—(17)更新样本权值D_{k+1}，否则退出循环。本文主要在原始样本权值分布$w_{k+1,i}$上引入代价敏感系数C_i和G_i(其中C_i为指数型代价敏感系数，G_i为线性代价敏感系数)，实现原始样本权值分布更新，改进后的样本权值分布更新如下

$$w_{k+1,i}=w_{k,i}\times\begin{cases}G_1\alpha_k^{C_1(1-e_{ki})}, 0<e_{ki}\leqslant 0.1\\ G_2\alpha_k^{C_2(1-e_{ki})}, 0.1<e_{ki}\leqslant 0.3\\ G_3\alpha_k^{C_3(1-e_{ki})}, 0.3<e_{ki}<0.5\end{cases}。\quad(20)$$

构建弱学习器的线性组合，组合、输出强学习器$f(x)$，同式(18)—(19)。

3.4 算法试验结果及其性能分析

基于前文构建的数据集，采用试验方式来确定合适的相对误差代价系数，将原始AdaBoost算法改进成代价敏感的学习器。首先，引入相对误差代价系数C_i和G_i时采用2种方式进行比较。第1种是按照被上一轮学习器对样本预测相对误差代价较小、样本预测相对误差代价较大、样本预测相对误差代价大3种情况进行加权(式(20))，其中C_i采用递减方式，取值(1.00,0.75,0.50)，G_i采用递增方式，取值(1.00,2.00,3.00)，此种改进AdaBoost算法称为参数组合1。第2种是采取对被上一轮学习器样本预测相对误差代价大的样本分配大权重，其他样本分配小权重的加权方式(式(20))，其中C_i取值(2.00,1.00,0.25)，G_i取值(1.00,1.00,3.00)，此种改进AdaBoost算法称为参数组合2。

试验首先使用Python的sklearn. ensemble库实现AdaBoostRegressor回归预测，建立基于AdaBoost原始算法的回归预测模型；同时，采用不同参数构建基于AdaBoost多参数代价敏感系数改进算法的回归预测模型，称为参数组合1和多参数组合2，并记录在不同弱学习器数量、学习率、损失函数条件下，AdaBoost原始算法和AdaBoost多参数代价敏感系数改进算法在训练集和测试集上的分数(score)。试验中，AdaBoost算法所使用的弱学习器为决策树。

表4数据表明：在训练集上，不同数量弱学习器条件下，基于AdaBoost原始算法的回归预测模型(以下简称原始模型)和基于AdaBoost多参数代价敏感系数改进算法的回归预测模型(以下简称参数组合模型)的分数大致相同，且随着弱学习器数量的增加，预测模型的性能有一定幅度的降低；在测试集上，参数组合模型分数大幅高于原始模型且参

数组合2模型的性能更优,弱学习器数量为200时,分数相较于原始算法提高0.17左右。

表5数据表明:在训练集上,不同学习率条件下,原始模型的分数整体低于参数组合模型,且随着学习率的增加,原始模型的分数变化不大,而参数组合模型的分数呈下降趋势;在测试集上,参数组合模型的分数明显高于原始模型,且参数组合1模型的性能低于参数组合2模型,并在学习率为0.03时,分数相较于原始模型提高了0.17左右。

表6数据表明:在训练集上,不同损失函数条件下,原始模型的分数低于参数组合模型;在测试集上,不同损失函数条件下,参数组合2模型的分数明显高于原始模型和参数组合1模型,且在弱学习器数量为700、损失函数为指数函数时性能更优。

上述试验结果表明:与原始模型相比,参数组合模型的性能有较大幅度的提升;在对AdaBoost多参数代价敏感系数算法进行改进时,通过采取给预测相对误差代价大的样本分配大权重、其他样本分配小权重的加权方式,构建的回归预测模型性能更优。

表4　采用不同数量弱学习器的算法模型分数

Tab. 4　Algorithm model scores with different numbers of weak learners

预测模型		弱学习器数量									
		100	200	300	400	500	600	700	800	900	1 000
原始算法	训练集	0.414 3	0.414 3	0.453 7	0.476 1	0.476 1	0.474 1	0.491 2	0.497 4	0.476 0	0.492 0
	测试集	0.581 4	0.681 4	0.661 4	0.663 2	0.563 2	0.514 8	0.661 1	0.561 1	0.544 1	0.514 1
参数组合1	训练集	0.460 5	0.460 5	0.468 1	0.488 1	0.492 7	0.499 6	0.460 5	0.493 6	0.470 9	0.501 2
	测试集	0.673 4	0.773 4	0.753 4	0.755 2	0.655 2	0.606 8	0.753 1	0.653 1	0.636 1	0.606 1
参数组合2	训练集	0.493 6	0.483 6	0.474 2	0.482 2	0.496 1	0.519 4	0.501 1	0.446 8	0.514 2	0.521 6
	测试集	0.755 4	0.855 4	0.835 4	0.837 2	0.737 2	0.688 8	0.835 1	0.735 1	0.718 1	0.688 1

表5　采用不同学习率的算法模型分数

Tab. 5　Algorithm model scores with different learning rates

预测模型		学习率									
		0.001	0.003	0.010	0.030	0.100	0.300	0.500	1.000	3.000	10.000
原始算法	训练集	0.478 7	0.465 7	0.464 5	0.490 7	0.480 0	0.479 1	0.480 9	0.476 5	0.516 3	0.489 4
	测试集	0.478 2	0.552 8	0.590 2	0.616 6	0.510 0	0.452 2	0.437 3	0.572 7	0.480 3	0.435 1
参数组合1	训练集	0.513 8	0.495 2	0.508 7	0.469 6	0.475 5	0.470 1	0.475 6	0.473 0	0.523 7	0.488 3
	测试集	0.568 2	0.642 8	0.680 2	0.706 6	0.600 0	0.542 2	0.527 3	0.662 7	0.570 3	0.525 1
参数组合2	训练集	0.505 4	0.558 6	0.555 3	0.569 7	0.552 8	0.581 2	0.482 9	0.528 6	0.461 6	0.525 1
	测试集	0.649 2	0.723 8	0.761 2	0.787 6	0.681 0	0.623 2	0.608 3	0.743 7	0.651 3	0.606 1

表6　采用不同损失函数的算法模型分数

Tab. 6　Algorithm model scores with different loss function

学习器数量	损失函数	原始算法		参数组合1		参数组合2	
		训练集	测试集	测试集	测试集	测试集	测试集
100	线性	0.459 0	0.590 6	0.511 5	0.679 6	0.543 1	0.770 6
	平方	0.425 3	0.514 0	0.488 5	0.603 0	0.491 0	0.694 0
	指数	0.435 5	0.496 1	0.496 3	0.585 1	0.504 3	0.676 1
300	线性	0.466 9	0.495 8	0.525 4	0.584 8	0.565 4	0.675 8
	平方	0.453 5	0.597 7	0.468 2	0.686 7	0.519 8	0.777 7
	指数	0.443 0	0.193 1	0.545 9	0.282 1	0.509 2	0.373 1
500	线性	0.486 0	0.495 8	0.494 1	0.584 8	0.551 9	0.675 8
	平方	0.481 7	0.497 7	0.531 3	0.586 7	0.530 9	0.677 7
	指数	0.454 2	0.507 5	0.493 3	0.596 5	0.508 9	0.687 5
700	线性	0.485 5	0.492 7	0.560 8	0.581 7	0.565 5	0.672 7
	平方	0.487 0	0.451 0	0.535 4	0.540 0	0.584 4	0.631 0
	指数	0.447 2	0.693 4	0.546 0	0.782 4	0.513 7	0.873 4

4 结束语

采用本文研究成果对江苏省某高新技术企业用户电力能耗进行建模分析，有利于企业用户预测未来短期内电力能耗，达到合理安排生产计划、绿色节能的目的，可为供电公司合理安排供电计划提供重要依据。从试验结果及分析结论看，基于多参数代价敏感学习Boosting算法构建数据驱动的电力能耗预测模型具有较强学习能力，用户根据企业当前及过往的生产日期即可快速预测未来短期内电力能耗情况。但由于本研究用于分析建模数据的维度较少，且样本量只有3 579 180个，因此获得的数据信息有限。本文仅对此做初步探讨，后期计划采集大样本量、多维度用电数据，以进一步分析不同条件下模型性能的优劣，逐步提升基于数据驱动的预测模型的准确度并应用于工业生产。

参考文献：

[1]梁玉红，聂淑花.产业升级背景下我国战略性新兴产业发展研究[J].中国党政干部论坛，2017(1)：22-24.
LIAGN Yuhong，NIE Shuhua.Research on the development of strategic emerging industries in China under the background of industrial upgrading [J]. Chinese Cadres Tribune，2017(1)：22-24.

[2]郑琪文，程方晓.电力系统继电保护不稳定原因及解决办法研究[J].科技风，2020(18)：201-202.
ZHENG Qiwen，CHENG Fangxiao.Research on the reasons and solutions of the unstable relay protection of power system [J].Technology Wind ，2020(18)：201-202.

[3]岳峰，史志伟，董金才，等.智能电网继电保护控制设备硬件可靠性设计及测试[J].华电技术，2020，42(2)：50-57.
YUE Feng，SHI Zhiwei，DONG Jincai，et al. Reliability design and test for relay protection devices applied in smart grid[J]. Huadian Technology，2020，42(2)：50-57.

[4]吴源梓.厉行节约倡导低碳生活——基层央行节能减排工作思考[J].时代金融，2016(23)：315-316.
WU Yuanzi.Strict economy and advocate low-carbon life：Thinking about the work of energy saving and emission reduction of the basic central bank[J].Times Finance ，2016(23)：315-316.

[5]胡倩，孙志达，江坷滕，等 基于短期负荷打捆预测的售电公司偏差考核控制方法[J].华电技术，2021，43(4)：47-55.
HU Qian，SUN Zhida，JIANG Keteng，et al. Deviation assessment and control method for electricity sales companies based on short-term load bundling forecast[J]. Huadian Technology，2021，43(4)：47-55.

[6]卢芸.短期电力负荷预测关键问题与方法的研究[D].沈阳：沈阳工业大学，2007.

[7]陈涛，吕松，任廷林，等.基于最小二乘支持向量机的周用电量预测方法[J].华电技术，2020，42(1)：35-40.
CHEN Tao，LYU Song，REN Tinglin，et al. Prediction method for weekly electricity consumption based on LSSVM algorithm[J]. Huadian Technology，2020，42(1)：35-40.

[8]JIN Y，RIVARD H，ZMEUREANU R. On-line building energy prediction using adaptive artificial neural networks [J]. Energy & Buildings，2005，37(12)：1250-1259.

[9]NETO A H，FIORELLI F A S. Comparison between detailed model simulation and artificial neural network for forecasting building energy consumption [J]. Energy & Buildings，2008，40(12)：2169-2176.

[10]李正方，杜景林，周芸.基于改进CART算法的降雨量预测模型[J].现代电子技术，2020(2)：133-137.
LI Zhengfang，DU Jinglin，ZHOU Yun.Rainfall prediction model based on improved CART algorithm [J]. Modern Electronics Technique，2020(2)：133-137.

[11]FRIEDMAN J H，HASTIE T，TIBSHIRANI R. Additive logistic regression：A statistical view of boosting[J].Annals of Statistics，2000，28(2)：337-407.

[12]FRIEDMAN J H. Greedy function approxi-mation：A gradient boosting machine[J]. Annals of Statistics，2001，29(5)：1189-1232.

[13]王学玲，王建林.基于代价敏感的AdaBoost算法改进[J].计算机应用与软件，2013(10)：123-125.
WANG Xueling，WANG Jianlin. Improving AdaBoost algorithm based on cost-sensitive [J]. Computer Applications and Software，2013(10)：123-125.

（本文责编：刘芳）

作者简介：

施杰(1992—)，女，河南驻马店人，在读硕士研究生，从事数据挖掘研究(E-mail：1181540647@qq.com)。

张安勤*(1974—)，女，上海人，副教授，工学博士，从事数据挖掘和普适计算等方面的研究(E-mail：aqz612@sina.com)。