基于Shapley值的电力负荷预测结果溯源分析方法

庞传军[1,2],刘金波[3],张 波[1,2],杨笑宇[3],余建明[1,2],刘 艳[1,2]
(1. 南瑞集团有限公司(国网电力科学研究院有限公司),江苏 南京 211106;
2. 北京科东电力控制系统有限责任公司,北京 100192;3. 国家电网公司国家电力调度控制中心,北京 100031)

摘要:针对由于机器学习的黑盒特性导致负荷预测结果不可溯源的问题,提出一种基于Shapley值的电力负荷预测结果溯源分析方法。阐述利用机器学习技术构建负荷预测模型的一般形式和基本过程;基于负荷预测模型,利用合作博弈论中的Shapley值计算各类负荷影响因素对负荷预测结果的影响;对利用梯度提升决策树算法训练的负荷预测模型的预测结果进行溯源分析。实验结果表明,利用所提方法可以洞察负荷预测过程,从而实现负荷预测结果的溯源分析以及考虑复杂非线性的负荷影响因素分析,也可以在构建负荷预测模型时指导特征选择提升模型的泛化能力。

关键词:负荷预测;负荷影响因素;溯源分析;Shapley值;梯度提升决策树;机器学习
中图分类号:TM 73 **文献标志码**:A DOI:10.16081/j.epae.202110001

0 引言

负荷预测是实现电力系统安全、经济运行的基础。随着风电、光伏等分布式清洁能源大规模接入电网以及电动汽车的广泛使用[1-2],影响电力负荷的因素愈加复杂。无论是电网调度运行人员还是电力市场成员,对负荷预测结果溯源分析的需求更加迫切,这是由于对负荷预测结果进行溯源分析不但可以帮助其理解负荷预测结果、洞察负荷预测过程,而且可以帮助其分析各类因素对负荷的影响,从而提升负荷预测的准确度。

目前,除了传统的基于时间序列分析的负荷预测方法[3]之外,越来越多的机器学习算法应用于电力负荷预测领域[4-5],大幅提升了负荷预测的准确度。但采用机器学习算法建立的负荷预测模型具备黑盒特性(不可解释性),无法对负荷预测结果进行溯源分析。

关于负荷预测结果溯源分析的研究大多是在负荷预测模型训练之前采用相关性分析的方法衡量各影响因素与负荷之间的关联关系,为特征选择提供参考。常用的相关性分析方法包括基于灰色关联度的方法和皮尔逊相关系数法。文献[6]提出一种改进的灰色关联分析模型,对影响电力负荷的因素进行量化分析。文献[7]引入灰色关联度理论对影响郑州电网负荷特性的因素进行量化分析,找出引起负荷特性变化的主要因素。文献[8]采用皮尔森相关系数获得气温和负荷之间的相关性特征,并结合格兰杰因果检验挖掘因素变化量之间的因果引导关系。此外,还有其他一些相关性分析方法,如基于费歇信息的方法[9]、层次分析法[10]、多项式回归法[11]等。

由于影响因素与用电负荷之间的关系具有高度复杂性、高度非线性等特点,现有方法并不能真实反映影响因素对负荷的复杂非线性影响,因此,本文基于先训练负荷预测模型、后分析相关性的思路,提出基于Shapley值的负荷预测结果溯源分析方法。基于实际负荷数据,对使用梯度提升决策树GBDT(Gradient Boosting Decision Tree)算法训练的负荷预测模型的预测结果进行溯源分析,利用Shapley值衡量时刻、气温等因素对负荷预测结果的影响。算例结果表明,利用该方法可以帮助电网调度运行人员和电力市场成员洞察负荷预测模型的预测过程,解释负荷预测结果产生的原因,分析各类因素与负荷之间的复杂非线性关系,也可以指导负荷预测模型构建之前的特征选择,从而提升负荷预测模型的泛化能力。

1 基于机器学习的负荷预测

影响电力负荷的各类因素与电力负荷之间具备高度复杂性、高度非线性等特点,导致无法采用数学分析的方法建立精准的电力负荷物理模型,基于"数据驱动"的思想,利用机器学习的方法构建负荷预测模型是可选的方法之一,且该方法已取得了较好的预测效果。从机器学习的角度,负荷预测是典型的有监督学习问题,是基于负荷影响因素历史数据X和历史负荷数据Y,以最小化损失$L\left(Y,\hat{F}(X)\right)$为目标,训练负荷预测模型$\hat{F}(X)$,然后采用模型预测未来负荷的过程,如式(1)[12]所示。

$$\hat{F}(X)=\underset{F}{\operatorname{argmin}}\ E_{X,Y}\left(L\left(Y,\hat{F}(X)\right)\right) \tag{1}$$

收稿日期:2021-01-26;**修回日期**:2021-08-02
基金项目:国家电网公司科技项目(5700-202055368A-0-0-00)
Project supported by the Science and Technology Project of State Grid Corporation of China(5700-202055368A-0-0-00)

式中：argmin表示训练模型F使得损失函数$L(\cdot)$的期望最小，得到最终负荷预测模型$\hat{F}(X)$；$X=\{x_i\}_{i=1}^{n}$，$x_i=\{x_i^1,x_i^2,\cdots,x_i^j,\cdots,x_i^p\}$为影响因素数据集中的第$i$个样本，$x_i^j$为第$i$个样本中第$j$个影响因素，$p$为负荷影响因素的数量，$n$为训练数据中样本的数量；$Y=\{y_i\}_{i=1}^{n}$，$y_i$为负荷数据集中的第$i$个样本；$E(\cdot)$表示求数学期望。

目前，很多有监督机器学习算法被用于建立负荷预测模型[3-5]，并取得了较好的负荷预测效果。如果能够定量衡量机器学习方法训练的负荷预测模型中的负荷影响因素与负荷之间的复杂非线性关系，则可以对负荷预测结果进行溯源分析。

2　基于Shapley值的负荷预测结果溯源

2.1　负荷影响因素Shapley值的定义

Shapley值是合作博弈论中根据团队中的成员对总收益的贡献来为每个成员公平分配收益的方法[13]，其在电力系统领域已有应用[14]。电网负荷受多个因素的影响，如果将每个因素抽象为一个团队成员，将负荷预测的结果作为总收益，则每个因素对负荷预测结果的贡献可用Shapley值衡量。对于负荷影响因素历史数据X中的一个样本$x_i=\{x_i^1,x_i^2,\cdots,x_i^j,\cdots,x_i^p\}$，其第$j$个影响因素$x_i^j$的Shapley值$\phi_i^{(j)}$定义如下：

$$\phi_i^{(j)}=\sum_{S\subseteq x_i\setminus\{x_i^j\}}\frac{|S|!(p-|S|-1)!}{p!}\left(\mathrm{val}\left(S\cup\{x_i^j\}\right)-\mathrm{val}(S)\right)\quad(2)$$

式中：S为x_i中排除第j个影响因素x_i^j后剩余影响因素的子集；$|S|$为S中影响因素的数量；$\mathrm{val}(S)$为特征函数，表示S中的影响因素通过“协作”对负荷预测结果的影响程度，其值可通过S中的影响因素对不包含在S中的影响因素在负荷预测模型$\hat{F}$上的输出计算得到，如式(3)所示。

$$\mathrm{val}(S)=\int\hat{F}(x_i^1,\cdots,x,\cdots,x_i^p)\,\mathrm{d}x-\frac{1}{n}\sum_{i=1}^{n}\hat{F}(x_i)\quad x\notin S\quad(3)$$

式(3)中，对于样本x_i中不包含在S中的每个影响因素都要执行积分求和运算。

2.2　Shapley值的特性

Shapley值具有有效性、对称性、冗员性和可加性，并且满足这4个特性的Shapley值存在且唯一[13]。在负荷预测结果溯源分析问题中，Shapley值的4个特性可以表示为以下形式。

1）有效性。一个样本x_i中所有影响因素的Shapley值之和等于基于该样本的负荷预测结果与基于所有数据集负荷预测结果的平均值之差，如式(4)所示。

$$\sum_{j=1}^{p}\phi_i^{(j)}=\hat{F}(x_i)-\frac{1}{n}\sum_{k=1}^{n}\hat{F}(x_k)\quad(4)$$

2）对称性。如果一个样本x_i中有2个影响因素x_i^j和x_i^k，分别由这2个影响因素与任意其他相同影响因素组成的子集对负荷的影响均相同，则这2个影响因素的Shapley值也相同，即对于负荷影响因素的子集$S\subseteq x_i\setminus\{x_i^j,x_i^k\}$，如果满足式(5)，则$\phi_i^{(j)}=\phi_i^{(k)}$。

$$\mathrm{val}\left(S\cup\{x_i^j\}\right)=\mathrm{val}\left(S\cup\{x_i^k\}\right)\quad(5)$$

3）冗员性。如果一个影响因素在所有的影响因素子集中对负荷预测结果均没有贡献，则该影响因素的Shapley值为0，即对于除影响因素x_i^j外的负荷影响因素的子集$S\subseteq x_i\setminus\{x_i^j\}$，如果满足式(6)，则$\phi_i^{(j)}=0$。

$$\mathrm{val}\left(S\cup\{x_i^j\}\right)=\mathrm{val}(S)\quad(6)$$

4）可加性。如果采用相同的数据集训练多个负荷预测模型，则将所有模型预测结果的平均值作为最终的负荷预测结果。可加性保证了基于每个预测模型计算的Shapley平均值为影响因素对最终预测结果的贡献。

Shapley值的4个特性可以保证在各负荷影响因素之间公平地分配其对负荷预测结果的贡献，从而反映各影响因素对负荷预测结果的影响。

2.3　Shapley值的计算方法

由Shapley值的定义可知，对包含和不包含影响因素x_i^j的所有组合都要进行计算才能得到x_i^j的Shapley值。当影响因素较多时，组合数量会呈指数增长，导致计算Shapley值的性能降低，并且负荷预测模型训练完成后，模型输入影响因素的数量是固定的，无法真正地将影响因素x_i^j从数据集中排除计算Shapley值，因此，本文采用蒙特卡罗采样法近似计算负荷影响因素的Shapley值[15-16]。利用该方法计算Shapley值所需的输入包括训练完成的负荷预测模型$\hat{F}$、采样次数（迭代次数）M、需要计算的样本x_i及数据集$D=\{x_i,y_i\}_{i=1}^{n}$。计算过程如下。

1）设置初始迭代次数$m=1$。

2）在数据集D中随机选取一个数据样本z。

3）将x_i中的p个影响因素进行随机排列，生成新的排列顺序，如式(7)所示。

$$x_{i,o}=\{x_i^{(1)},x_i^{(2)},\cdots,x_i^{(j)},\cdots,x_i^{(p)}\}\quad(7)$$

式中：$x_{i,o}$为样本x_i中的影响因素按照新的排列顺序重新排序后的结果，$x_i^{(j)}$为其中的第j个影响因素。

4）根据新的排列顺序对随机选取的样本z中的影响因素进行排序：

$$z_o=\{z^{(1)},z^{(2)},\cdots,z^{(j)},\cdots,z^{(p)}\}\quad(8)$$

式中：z_o为随机选取的样本z中的影响因素按照新的排列顺序重新排序后的结果，$z^{(j)}$为其中的第j个影

响因素。

5)根据$x_{i,o}$和z_o构建包含及不包含$x_i^{(j)}$的2个新样本,分别如式(9)、(10)所示。

$$x_m^{+j}=\left\{x_i^{(1)},x_i^{(2)},\cdots,x_i^{(j-1)},x_i^{(j)},z^{(j+1)},\cdots,z^{(p-1)},z^{(p)}\right\} \quad (9)$$

$$x_m^{-j}=\left\{x_i^{(1)},x_i^{(2)},\cdots,x_i^{(j-1)},z^{(j)},z^{(j+1)},\cdots,z^{(p-1)},z^{(p)}\right\} \quad (10)$$

6)根据步骤5)生成的2个新样本和训练完成的负荷预测模型$\hat{F}$计算第m次迭代影响因素$x_i^{(j)}$对负荷预测结果的贡献$\phi_{i,m}^{(j)}$,如式(11)所示。

$$\phi_{i,m}^{(j)}=\hat{F}\left(x_m^{+j}\right)-\hat{F}\left(x_m^{-j}\right) \quad (11)$$

7)令$m=m+1$,若$m\leqslant M$,则循环执行步骤2)—6)。M次迭代完成后,计算M次迭代后影响因素$x_i^{(j)}$对负荷预测结果贡献的平均值,得到$x_i^{(j)}$的Shapley值,如式(12)所示。

$$\phi_i^{(j)}=\frac{1}{M}\sum_{m=1}^{M}\phi_{i,m}^{(j)} \quad (12)$$

负荷预测模型$\hat{F}$可以采用任何有监督机器学习算法进行训练,因此该Shapley值的计算方法具备通用性。采用该方法可以计算数据集D每个样本中各负荷影响因素的Shapley值。

2.4 基于Shapley值的负荷影响因素重要性

Shapley值表示影响因素对负荷预测结果的影响程度,Shapley值的绝对值越大,影响因素对预测结果的影响越大。数据集D所有样本中每个影响因素Shapley值的平均绝对值可以衡量该影响因素对负荷预测结果的影响大小[17],如式(13)所示。

$$I_j=\frac{1}{n}\sum_{i=1}^{n}\left|\phi_i^{(j)}\right| \quad (13)$$

式中:I_j为负荷影响因素$x_i^{(j)}$对负荷预测结果的影响大小。

3 算例分析

3.1 实验数据及预测结果评估方法

采用Shapley值对某地区的负荷预测结果进行溯源分析。以该地区2002年至2018年每日24 h的历史负荷及负荷影响因素数据进行验证,时间分辨率为1 h。选取的负荷影响因素见表1。

表1 实验选取的负荷影响因素

Table 1 Influencing factors of load selected in experiment

负荷影响因素	说明
气温	单位为℃
年份	对应年份
月份	对应月份,取值范围为1~12
日期	当月第几日
星期	当前星期几,取值范围为1~7
时刻	当前时刻,取值范围为0~23
是否为工作日	非工作日取值0,工作日取值1,公休日为非工作日
第几天	当年第几天,闰年取值范围为1~366,否则取值范围为1~365

将2002年至2017年历史负荷和负荷影响因素数据作为训练集训练负荷预测模型,将2018年历史负荷和负荷影响因素数据作为测试集。采用模型在测试集上的绝对百分比误差APE(Absolute Percentage Error)、平均绝对误差MAE(Mean Absolute Error)、均方根误差RMSE(Root Mean Square Error)衡量模型的预测性能[18]。

3.2 负荷预测模型的训练

基于历史负荷和负荷影响因素数据,利用机器学习算法训练日前负荷直接预测模型,对未来一日每小时的负荷进行直接预测。负荷预测模型的准确程度影响Shapley值的计算效果,GBDT算法在负荷预测领域取得了较好的预测效果[5],因此,本文采用GBDT算法训练负荷预测模型$\hat{F}$,损失函数采用均方差损失[12],并利用网格搜索方法[19]寻找算法所需超参数(学习率、决策树最大深度和决策数的数量[5])的最优组合。

2018年各月负荷预测的绝对百分比误差分布的箱型图如图1所示。由图可知,1月至12月的绝对百分比误差大部分小于2.5%,全年平均绝对百分比误差为1.1%,平均绝对误差为114.72 kW,均方根误差为155.74 kW,模型取得了较好的预测效果。

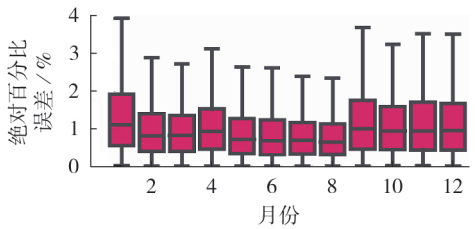

图1 2018年各月负荷预测的绝对百分比误差分布

Fig.1 Distribution of load forecasting APE for each month in 2018

3.3 影响因素Shapley值的计算

基于训练完成的负荷预测模型$\hat{F}$,利用2.3节Shapley值的计算方法计算2018年历史负荷数据组成的数据集每个样本中各影响因素的Shapley值。基于Shapley值对负荷预测结果和各因素对负荷的影响进行分析。

3.4 负荷预测结果溯源分析

采用每个影响因素的Shapley值可以对各时刻负荷预测结果进行溯源分析。以2018年1月1日09:00的预测结果为例,预测过程如图2所示,图中各条形图旁的数据为各因素的Shapley值,数据单位均为kW。

由图2可知,预测过程由预测模型在数据集上预测的平均负荷(10 682.344 kW)位置开始,受到各因素的影响得到该时刻的预测结果(12 788.065 kW),影响大小由各因素的Shapley值衡量。Shapley值为

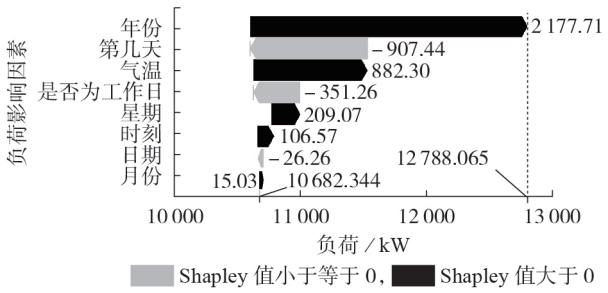

图2　2018年1月1日09:00负荷预测过程

Fig.2　Load forecasting process at 9 am on January 1, 2018

正时，升高负荷预测的结果，为负时则降低负荷预测的结果。预测结果与预测的平均负荷的偏差为12 788.065−10 682.344≈2 105.72（kW），该偏差等于各影响因素Shapley值之和（2 177.71−907.44+882.3−351.26+209.07+106.57−26.26+15.03=2 105.72（kW）），满足Shapley值的有效性。

3.5　影响因素与负荷相关性分析

采用皮尔逊相关系数法计算负荷与各影响因素之间的相关关系，并根据式（13）计算各影响因素Shapley值的平均绝对值，对比两者衡量各影响因素对负荷影响的大小，将结果按照由大到小排序如图3所示。

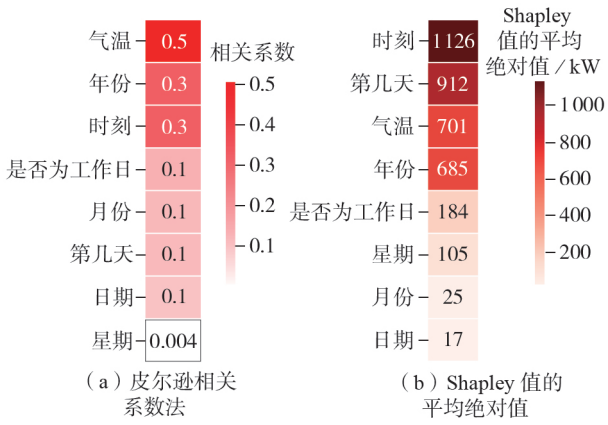

图3　利用皮尔逊相关系数法和Shapley值的平均绝对值衡量各影响因素与负荷之间的相关关系

Fig.3　Correlation between each influencing factor and load measured by Pearson correlation coefficient method and average absolute value of Shapley value

由图3可知：Shapley值的平均绝对值与皮尔逊相关系数法得到的结果排序不同；采用皮尔逊相关系数衡量各影响因素对负荷的影响时，是否为工作日、月份、第几天、日期与负荷的相关系数相同，因此不能区分这4个影响因素对负荷影响的大小，而采用Shapley值的平均绝对值可以进行区分；月份和日期与第几天相关，这2个因素的Shapley值的平均绝对值较小，对负荷的影响较小，这说明即使影响因素之间存在关联关系，Shapley值的平均绝对值仍能够有效地衡量影响因素对负荷的影响。

皮尔逊相关系数仅体现了各影响因素与负荷的线性关系，而没有体现非线性关系。例如，时刻与负荷在不同时段的相关性并不同，某些时段呈正相关，某些时段呈负相关，采用相关系数会导致正、负相互抵消，不能正确反映时刻对负荷的影响程度。而采用Shapley值的平均绝对值可考虑影响因素对负荷的非线性影响，正确反映影响因素对负荷影响的大小，更好地指导负荷预测模型建立过程中的特征选择。

3.6　特征选择对负荷预测结果的影响

在训练负荷预测模型时，如果选取的影响因素之间存在较强的线性或者非线性关联关系，则会导致模型过拟合，影响模型的泛化能力以及模型训练时间。因此，训练模型之前进行特征选择和降维是提升负荷预测准确性的重要步骤之一。由图3可知，星期、月份、日期3个影响因素Shapley值的平均绝对值较小，对负荷影响较小，因此在训练GBDT负荷预测模型时不考虑这3个影响因素。利用2018年全年负荷作为测试集对影响因素选择前、后的预测性能进行对比，2018年各月负荷预测的绝对百分比误差箱型图如图4所示。

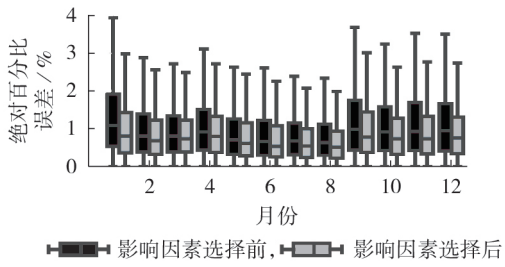

图4　影响因素选择前、后2018年各月负荷预测的绝对百分比误差分布

Fig.4　Distribution of load forecasting APE for each month in 2018 before and after selection of influencing factors

影响因素选择后，全年的平均绝对误差由114.72 kW下降至99.32 kW，均方根误差由155.74 kW下降至136.66 kW。结合图4可以看出，不考虑冗余的影响因素后，负荷预测的准确性并没有降低，反而有所提升，这表明依据Shapley值排除冗余影响因素后可以提升预测模型的泛化能力。

3.7　影响因素变化对负荷的非线性影响

将影响因素的值和相应的Shapley值以散点图的形式展示，可以衡量影响因素的变化对负荷的非线性影响，不同季节中气温与Shapley值的散点图如图5所示。由图可知：当气温在0以下或者25 ℃以上时，Shapley值为正，会升高负荷预测的结果；当气温在0~25 ℃之间时，Shapley值为负，会降低负荷预测

的结果；在春季、秋季和冬季这3个季节，气温和负荷之间呈复杂的非线性关系，当气温高于12 ℃时，呈正相关关系，当气温低于12 ℃时，呈负相关关系，但是在夏季，两者基本呈正相关关系。综上可知，该地区的电力负荷是对气温较为敏感的负荷。

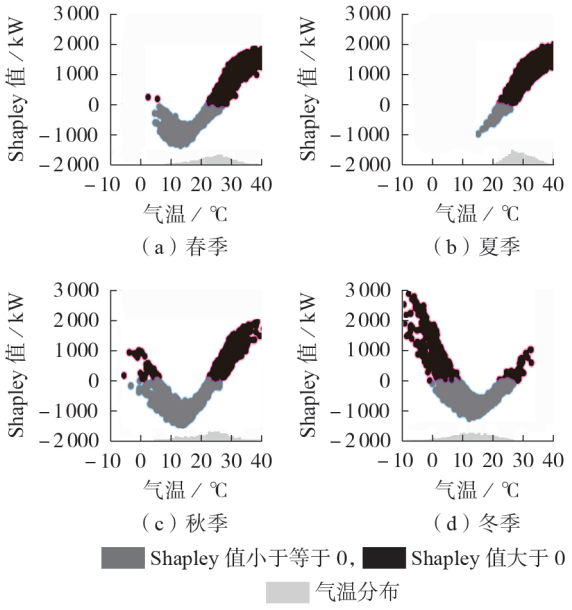

图5 气温与Shapley值的散点图

Fig.5 Scatter plot of temperature and Shapley values

由图5还可看出，在不同季节，当气温相同时，其Shapley值并不相同，这说明负荷还受其他因素的影响。图6为春季时刻与其Shapley值的散点图。

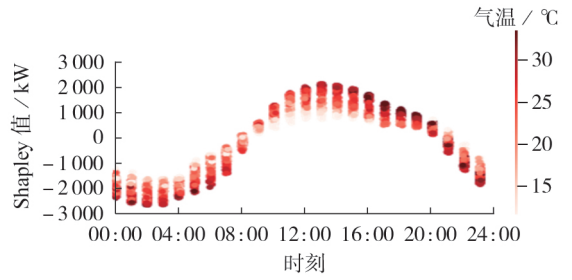

图6 时刻与Shapley值的散点图

Fig.6 Scatter plot of time and Shapley values

由图6可以看出，在一天中的不同时刻，气温的高低对负荷的影响不同，白天气温升高会升高负荷预测的结果，晚上则相反。

4 结论

本文提出基于Shapley值的负荷预测结果溯源分析方法，可对基于机器学习算法的负荷预测结果进行溯源分析。采用实际的负荷数据进行验证，实验结果表明：

1）采用本文方法计算的Shapley值满足有效性的要求，能够在负荷影响因素之间公平分配其对负荷预测结果的贡献；

2）采用Shapley值解决了负荷预测模型不可解释的问题，可以洞察负荷预测模型的预测过程，分析负荷预测结果产生的原因；

3）与皮尔逊相关系数法相比，采用Shapley值可以计及各影响因素与负荷之间的复杂非线性关系，辨识影响负荷的重要因素，指导特征选择，提升负荷预测模型的泛化能力；

4）采用Shapley值可以衡量影响因素变化对负荷的非线性影响。

Shapley值的计算以训练完成的负荷预测模型为基础，预测模型可以采用其他机器学习算法进行训练，在使用本文方法前需要选择准确性高的负荷预测模型。另外，不同负荷的影响因素不同，采用该方法分析不同负荷的影响因素可能会得到不同的结果。后续笔者将采用Shapley值对基于机器学习的负荷预测模型中的损失函数进行分析，以衡量造成负荷预测误差的原因。

参考文献：

[1] 崔杨，周慧娟，仲悟之，等. 考虑源荷两侧不确定性的含风电电力系统低碳调度[J]. 电力自动化设备，2020，40(11)：85-93.
CUI Yang，ZHOU Huijuan，ZHONG Wuzhi，et al. Low-carbon scheduling of power system with wind power considering uncertainty of both source and load sides[J]. Electric Power Automation Equipment，2020，40(11)：85-93.

[2] 陈丽丹，张尧，Antonio Figueiredo. 融合多源信息的电动汽车充电负荷预测及其对配电网的影响[J]. 电力自动化设备，2018，38(12)：1-10.
CHEN Lidan，ZHANG Yao，FIGUEIREDO A. Charging load forecasting of electric vehicles based on multi-source information fusion and its influence on distribution network[J]. Electric Power Automation Equipment，2018，38(12)：1-10.

[3] 庞昊，高金峰，杜耀恒. 基于多神经网络融合的短期负荷预测方法[J]. 电力自动化设备，2020，40(6)：37-43.
PANG Hao，GAO Jinfeng，DU Yaoheng. Short-term load forecasting method based on fusion of multiple neural networks[J]. Electric Power Automation Equipment，2020，40(6)：37-43.

[4] 郑瑞骁，张姝，肖先勇，等. 考虑温度模糊化的多层长短时记忆神经网络短期负荷预测[J]. 电力自动化设备，2020，40(10)：181-186.
ZHENG Ruixiao，ZHANG Shu，XIAO Xianyong，et al. Short-term load forecasting of multi-layer long short-term memory neural network considering temperature fuzziness[J]. Electric Power Automation Equipment，2020，40(10)：181-186.

[5] 刘波，秦川，鞠平，等. 基于XGBoost与Stacking模型融合的短期母线负荷预测[J]. 电力自动化设备，2020，40(3)：147-153.
LIU Bo，QIN Chuan，JU Ping，et al. Short-term bus load forecasting based on XGBoost and Stacking model fusion[J]. Electric Power Automation Equipment，2020，40(3)：147-153.

[6] 王雁凌，吴梦凯，周子青，等. 基于改进灰色关联度的电力负荷影响因素量化分析模型[J]. 电网技术，2017，41(6)：1772-1778.
WANG Yanling，WU Mengkai，ZHOU Ziqing，et al. Quantitative analysis model of power load influencing factors based on improved grey relational degree[J]. Power System Technology，2017，41(6)：1772-1778.

[7] 康田园，尹淑萍，王现法，等. 大型城市电网负荷特性及其影响因素分析[J]. 电测与仪表，2016，53(6)：51-56.

KANG Tianyuan, YIN Shuping, WANG Xianfa, et al. Load characteristics of large city power gridand analysis on its influencing factors[J]. Electrical Measurement & Instrumentation, 2016, 53(6): 51-56.
[8] 马瑞，周谢，彭舟，等. 考虑气温因素的负荷特性统计指标关联特征数据挖掘[J]. 中国电机工程学报，2015，35(1)：43-51.
MA Rui, ZHOU Xie, PENG Zhou, et al. Data mining on correlation feature of load characteristics statistical indexes considering temperature[J]. Proceedings of the CSEE, 2015, 35(1): 43-51.
[9] 蔡舒平，张保会，汤大海，等. 短期负荷预测中气象因素处理的费歇信息方法[J]. 电力自动化设备，2020，40(3)：141-146.
CAI Shuping, ZHANG Baohui, TANG Dahai, et al. Fisher information method for processing weather factors in short-term load forecasting[J]. Electric Power Automation Equipment, 2020, 40(3): 141-146.
[10] 李亦言，严正，冯冬涵. 考虑城市化因素的中长期负荷预测模型[J]. 电力自动化设备，2016，36(4)：54-61.
LI Yiyan, YAN Zheng, FENG Donghan. Mid / long-term load forecasting model considering urbanization characteristics[J]. Electric Power Automation Equipment, 2016, 36(4): 54-61.
[11] 张贲，史沛然，蒋超. 气象因素对京津唐电网夏季负荷特性影响分析[J]. 电力自动化设备，2013，33(12)：140-144.
ZHANG Ben, SHI Peiran, JIANG Chao. Impact of meteorological factors on summer load characteristics of Beijing-Tianjin-Tangshan Power Grid[J]. Electric Power Automation Equipment, 2013, 33(12): 140-144.
[12] 李航. 统计学习方法[M]. 北京：清华大学出版社，2012：3-21.
[13] SHAPLEY L S. A value for n-person games[J]. Contributions to the Theory of Games, 1953, 2(28): 307-317.
[14] 陈星莺，郁清云，谢俊，等. 基于合作博弈论的电能替代效益分摊方法[J]. 电力自动化设备，2019，39(3)：30-35，44.
CHEN Xingying, YU Qingyun, XIE Jun, et al. Benefit allocation method for electric energy substitution based on cooperative game theory[J]. Electric Power Automation Equipment, 2019, 39(3): 30-35, 44.
[15] LUNDBERG S, LEE S I. A unified approach to interpreting model predictions[C]// NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems. Chicago, USA: Advances in Neural Information Processing Systems, 2017: 4768-4777.
[16] LUNDBERG S M, ERION G, CHEN H, et al. From local explanations to global understanding with explainable AI for trees[J]. Nature Machine Intelligence, 2020, 2(1): 56-67.
[17] CHRISTOPH M. Interpretable machine learning[EB/OL]. (2020-03-24)[2020-09-08]. https://christophm.github.io/interpretable-ml-book.
[18] 国家电网公司. 电网短期超短期负荷预测技术规范：QGDW552—2010[S]. 北京：国家电网公司，2010.
[19] BERGSTRA J, BENGIO Y. Random search for hyper-parameter optimization[J]. Journal of Machine Learning Research, 2012, 13: 281-305.

作者简介：

庞传军

庞传军（1984—），男，山东聊城人，高级工程师，硕士，主要研究方向为电力系统及自动化、人工智能技术在电力系统中的应用（**E-mail**：pangchuanjun@sgepri.sgcc.com.cn）；

刘金波（1975—），男，辽宁大连人，教授级高级工程师，硕士，主要研究方向为电力系统调度自动化（**E-mail**：liu-jinbo@sgcc.com.cn）；

张　波（1978—），男，辽宁辽中人，高级工程师，硕士，主要研究方向为电力系统调度自动化（**E-mail**：zhangbo7@sgepri.sgcc.com.cn）。

（编辑　王锦秀）

Traceability analysis method of power load forecasting results based on Shapley value

PANG Chuanjun[1,2], LIU Jinbo[3], ZHANG Bo[1,2], YANG Xiaoyu[3], YU Jianming[1,2], LIU Yan[1,2]

(1. NARI Group Corporation Co., Ltd.(State Grid Electric Power Research Institute Co., Ltd.), Nanjing 211106, China;
2. Beijing Kedong Electric Power Control System Co., Ltd., Beijing 100192, China;
3. National Power Dispatching and Control Center of State Grid Corporation of China, Beijing 100031, China)

Abstract: Aiming at the problem that load forecasting results cannot be traced due to the black-box characteristic of machine learning, a traceability analysis method of power load forecasting results based on Shapley value is proposed. The general form and basic process of building load forecasting model with machine learning technology are explained. On the basis of load forecasting model, the Shapley value in cooperative game theory is used to calculate the impact of various influencing factors of load on the load forecasting results. Traceability analysis of the forecasting results of load forecasting model trained by the gradient boosting decision tree algorithm is carried out. The experimental results show that the proposed method can gain insight into the load forecasting process, so as to realize the traceability analysis of load forecasting results and analysis of influence factors of load considering complex nonlinearity, and can guide feature selection to improve the generalization ability of model when building load forecasting model.

Key words: load forecasting; influencing factors of load; traceability analysis; Shapley value; gradient boosting decision tree; machine learning