文章编号：1000-3673（2016）06-1642-08 中图分类号：TM 72 文献标志码：A 学科代码：470·40

基于 Spark 平台和多变量 L_2-Boosting 回归模型的分布式能源系统短期负荷预测

马天男[1]，牛东晓[1]，黄雅莉[1]，杜振东[2]
（1．华北电力大学 经济与管理学院，北京市 昌平区 102206；
2．国网浙江省电力公司 经济技术研究院，浙江省 杭州市 310000）

Short-Term Load Forecasting for Distributed Energy System Based on Spark Platform and Multi-Variable L_2-Boosting Regression Model

MA Tiannan[1], NIU Dongxiao[1], HUANG Yali[1], DU Zhendong[2]
(1. Economics and Management School, North China Electric Power University, Changping District, Beijing 102206, China;
2. Economy Research Institute, State Grid Zhejiang Electric Power Company, Hangzhou 310000, Zhejiang Province, China)

ABSTRACT: Load forecasting for distributed energy system is precondition and basis of system planning and economic operation. Under background of current massive high-dimension data, effective online data processing platform and accurate load forecasting methods are current research focus. In this paper, considering characteristics of load data of distributed energy system, short-term load forecasting method for distributed energy system based on Spark platform and multi-variable L_2-boosting regression model was established after missing data processing, bad data classification and feature selection. Firstly, Spark platform was used to divide up all data to get multiple sub-data models by parallel computing to improve data processing efficiency and using feature extraction method to obtain input vector required by the model. Secondly, effective data information obtained was input to multivariable L_2-boosting regression model to train the model and obtain trained multivariable L_2-boosting regression model. Finally, test data was used to test the model. Results verified validity of the proposed model.

KEY WORDS: short-term load forecasting; multi-variable L_2-Boosting regression model; distributed energy system; Spark platform

摘要：分布式能源系统负荷预测是系统规划与经济运行的可靠前提和依据，在当前海量高维数据的背景下，有效的在线数据处理平台与精确的负荷预测方法是当前的研究重点。基于分布式能源系统负荷数据特点，在缺失数据处理、坏数据分类以及特征选择的基础上，建立了基于 Spark 平台与多变量 L_2-Boosting 回归模型的分布式能源系统短期负荷预测方法。首先，利用 Spark 平台分割全部数据得到多个子数据模型，通过并行计算提高数据处理效率，采用特征提取方法得出模型需要的输入向量；其次，将得出的有效数据信息输入到多变量 L_2-Boosting 回归模型进行训练学习，得到训练后的多变量 L_2-Boosting 回归模型；最后，利用测试数据测试模型。算例结果验证了所提模型的有效性。

关键词：短期负荷预测；多变量 L_2-Boosting 回归模型；分布式能源系统；Spark 平台

DOI：10.13335/j.1000-3673.pst.2016.06.006

0 引言

能源互联网概念的提出，将有效促进我国能源生产、消费以及体制变革，推动我国能源转型，从而实现能源的清洁、高效、安全、便捷、可持续利用[1]。分布式能源系统作为能源互联网的重要组成基础，其呈现出的安全可靠、能源效率高、环境友好等多方面出色的特点受到广泛关注[2]，但分布式能源的利用存在出力随机等弊端。因此针对分布式能源系统时空负荷精准预测具有十分重要的意义，它不仅能够帮助提高系统运行的整体稳定性，还能够促进分布式能源的有效利用。多年来，随着现代计算机技术的发展以及数学理论的不断提升，负荷预测的技术和方法也在不断的发展；诸如神经网络[3]、灰色预测[4]、支持向量机[5]、小波分析[6-7]、模糊理论[8]、时间序列[9]、区间预估[10]等技术为电力负荷预测提供了多种有效工具，并取得了良好的效果。在上述方法中，各模型的有效性建立在历史数据准确性好、规律强以及数据规模较小的基础上；当面对规模庞大、结构复杂程度高、具有高度

基金项目：国家自然科学基金项目(71471059)。
Project Supported by National Natural Science Foundation of China (NSFC) (71471059).

实时性的多能源系统数据时，就会使上述模型出现过拟合、收敛速度慢、易陷入局部最优等问题，从而使负荷预测的精确性大幅度降低。

随着电网智能化进程的加快，大量智能电表和其他检测设备被部署在电网的各个重要位置，从而使得电力数据呈现“指数级”的增长，年数据量的增长也是从 GB(gigabyte)级增长到 TB(terabyte)级[11]。快速的数据增长使得电力负荷数据维度持续扩大，数据结构结构更复杂，使得负荷预测工作的难度大大增加。Boosting 方法[12]是一种通过提高弱分类算法准确度的方法，其主要通过对样本集的操作获取样本子集，进而利用弱分类算法将样本子集训练生成预测函数系列，将该系列通过组合后生成一个预测函数。Boosting 算法是一种任意逼近非线性曲线的学习算法，但其存在一个重大的缺陷，即需要事先知道弱分类算法分类正确率的下限；因此 Freund 和 Schapire 提出了 AdaBoost 算法[13]。目前，AdaBoost 算法已经得到了广泛的应用，其被证明具有较强的鲁棒性，易与其他算法结合，如 AdaBoost-SVM(support vector machine)[14]、AdaBoost-BP(back propagation)[15]等；尽管 AdaBoost 算法在单机预测工作中取得了相应的成果，但依然无法处理高维海量数据时的预测工作。因此，解决大数据背景下的分布式能源系统时空负荷精准预测显得十分重要。

针对上述问题，本文以分布式能源系统大数据为基础，提出一种基于 Spark 平台和多变量 L_2-Boosting 模型相结合的短期负荷预测方法。本文所提分布式能源系统短期负荷预测系统主要分为 3 个部分：首先，利用 Spark 平台将全部数据进行分割形成多个子数据模型，通过并行计算来提高数据处理效率，采用特征提取方法得出模型需要的输入向量；其次，将得出的有效数据信息输入到多变量 L_2-Boosting 回归模型进行训练学习，并得到训练后的多变量 L_2-Boosting 回归模型；最后，将测试数据带入到模型中进行预测。

1　基于 L_2-Boosting 模型的改进多元线性回归预测模型

假设给定 n 个样本观测数据$\{(X,Y)\}$，其中响应矩阵为 $\boldsymbol{Y}\in\boldsymbol{R}^{n\times q}$，协变量矩阵为 $\boldsymbol{X}\in\boldsymbol{R}^{n\times p}$，给定系数矩阵 $\boldsymbol{B}\in\boldsymbol{R}^{p\times q}$ 和误差矩阵 $\boldsymbol{E}\in\boldsymbol{R}^{n\times q}$，$p$ 为自变量维数，q 为响应维数，则多元线性回归模型可表示为

$$\boldsymbol{Y}=\boldsymbol{XB}+\boldsymbol{E} \tag{1}$$

通常，系数矩阵的估计值 $\hat{\boldsymbol{B}}$ 的求解是通过采用普通最小二乘算法(ordinary least square，OLS)得到[16]，其一般式为 $\hat{\boldsymbol{B}}=(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$。然而，随着数据量的增加、数据维度不断扩大，OLS 系数矩阵估计的准确性将会大大降低。因此，本文采用改进 Boosting 算法对协变量系数矩阵 $\hat{\boldsymbol{B}}$ 进行重构。

在改进 Boosting 算法的构建中，需要重新定义算法的损失函数和弱学习过程，本文采用负高斯对数似然函数(negative Gaussian likelihood function，NGLF)作为损失函数、分支线性最小二乘(component-wise linear least squares，CWLLS)作为弱学习过程[17]。

负高斯对数似然函数的方程公式为

$$-\boldsymbol{L}(\boldsymbol{B},\boldsymbol{Q})=-\ln((2\boldsymbol{\Pi})^{-\frac{nq}{2}}|\boldsymbol{Q}|^{-\frac{n}{2}})+\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{y}_{(i)}^{\mathrm{T}}-\boldsymbol{x}_{(i)}^{\mathrm{T}}\boldsymbol{B})\boldsymbol{Q}^{-1}(\boldsymbol{y}_{(i)}^{\mathrm{T}}-\boldsymbol{x}_{(i)}^{\mathrm{T}}\boldsymbol{B})^{\mathrm{T}} \tag{2}$$

式中：$\boldsymbol{\Pi}$ 代表高斯分布似然函数，具体见文献[18]；$\boldsymbol{y}_{(i)}$为第 i 个样本点的响应值(行)；$\boldsymbol{x}_{(i)}$为第 i 样本点对应的协变量矩阵(行)；$\boldsymbol{B}$ 的最大似然估计参数与普通最小二乘法的估计结果一致，与协方差矩阵 $\boldsymbol{Q}$ 独立。通常，协方差矩阵 $\boldsymbol{Q}$ 事先无法确定，因此，对式(2)进行调整得出本文所需损失函数为

$$\boldsymbol{L}(\boldsymbol{B})=\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{y}_{(i)}^{\mathrm{T}}-\boldsymbol{x}_{(i)}^{\mathrm{T}}\boldsymbol{B})\boldsymbol{\Phi}^{-1}(\boldsymbol{y}_{(i)}^{\mathrm{T}}-\boldsymbol{x}_{(i)}^{\mathrm{T}}\boldsymbol{B})^{\mathrm{T}} \tag{3}$$

式中 $\boldsymbol{\Phi}^{-1}$ 为加强的协方差矩阵，其用于协方差矩阵 $\boldsymbol{Q}$ 的估计，当响应矩阵维数 $q\to\infty$ 时，可取 $\boldsymbol{\Phi}=\boldsymbol{I}$ ($\boldsymbol{I}$ 为单位矩阵)。

采用弱学习过程可以不断地加强对协变量矩阵 $\boldsymbol{B}$ 的拟合与学习，从而达到参数估计的目的，假设给定矩阵 $\boldsymbol{X}$ 和虚拟响应矩阵 $\boldsymbol{R}\in\mathbf{R}^{n\times p}$ ($\boldsymbol{R}$ 并不等于 $\boldsymbol{Y}$)，采用分支最小二乘学习对 $\boldsymbol{X}$ 与 $\boldsymbol{R}$(列)之间进行最小二乘回归拟合，使得损失函数 $L(\boldsymbol{B})$最小，从而计算得出协变量系数的估计值，对于协变量系数矩阵 $\boldsymbol{B}$ 中元素的弱学习过程为

$$(\hat{s},\hat{t})=\arg\min\{\boldsymbol{L}(\boldsymbol{B}),\boldsymbol{B}_{jk}=\hat{\beta}_{jk},\boldsymbol{B}_{uv}=0(uv=jk)\}=\underset{1\le j\le p,1\le k\le q}{\arg\max}\frac{(\sum_{v=1}^{q}\boldsymbol{r}_v^{\mathrm{T}}\boldsymbol{x}_j\boldsymbol{\Phi}_{vk}^{-1})^2}{\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{x}_j\boldsymbol{\Phi}_{kk}^{-1}} \tag{4}$$

式中：arg 表示使目标函数取最小值时的变量值；$(\hat{s},\hat{t})$ 表示为使 $L(\boldsymbol{B})$取最小值时元素下标值；$\boldsymbol{B}_{jk}$ 为有 j 行、k 列元素的系数矩阵；$\hat{\boldsymbol{\beta}}_{jk}$ 为系数矩阵 $\boldsymbol{B}_{jk}$ 的估计矩阵；$\boldsymbol{B}_{uv}$ 为有 u 行、v 列元素的系数矩阵；$\boldsymbol{r}_v^{\mathrm{T}}$ 表示第 v 个误差矩阵的转置；$\boldsymbol{x}_j$ 为第 j 个协变量矩阵，$\boldsymbol{x}_j^{\mathrm{T}}$ 为 $\boldsymbol{x}_j$ 的转置矩阵；$\boldsymbol{\Phi}_{vk}^{-1}$ 为有 v 行、k 列元素单位矩阵的逆矩阵；$\boldsymbol{\Phi}_{kk}^{-1}$ 为 k 阶单位矩阵的逆矩阵。

对 $\hat{\beta}_{jk}$ 求导，使得 $\partial(\hat{s},\hat{t})/\partial\hat{\beta}_{jk}=0$ ，解得

$$\hat{\beta}_{jk}=(\sum_{v=1}^{q}r_v^T x_j \boldsymbol{\Phi}_{vk}^{-1})/(x_j^T x_j \boldsymbol{\Phi}_{kk}^{-1}) \tag{5}$$

令 $\hat{\boldsymbol{B}}_{\hat{s}\hat{t}}=\hat{\boldsymbol{\beta}}_{\hat{s}\hat{t}}$ ， $\hat{\boldsymbol{B}}_{jk}=0$ ， $(jk)\neq(\hat{s}\hat{t})$ ， $\hat{\boldsymbol{B}}_{\hat{s}\hat{t}}$ 为含 $\hat{s}$ 行 $\hat{t}$ 列元素的系数矩阵 $\boldsymbol{B}$ 的估计矩阵； $\hat{\beta}_{\hat{s}\hat{t}}$ 为 $\hat{\boldsymbol{B}}_{\hat{s}\hat{t}}$ 的估计矩阵； $\hat{\boldsymbol{B}}_{jk}$ 为拥有 j 行、k 列元素的系数矩阵；因此，多元线性回归方程可定义为

$$\hat{\boldsymbol{g}}_l=\begin{cases}\hat{\boldsymbol{\beta}}_{\hat{s}\hat{t}}\boldsymbol{x}_{\hat{s}}, & l=\hat{t},l=1,2,\mathrm{L}\ ,q\\ 0, & l\neq\hat{t},l=1,2,\mathrm{L}\ ,q\end{cases} \tag{6}$$

式中： $\boldsymbol{x}_{\hat{s}}$ 为第 $\hat{s}$ 维响应的元素矩阵；通过弱学习过程对不同虚拟响应 $\boldsymbol{R}$ 和回归方程估计进行多次拟合和学习，从而可以逐步建立基于 Boosting 算法加强的多元线性回归方程 $\hat{f}:\boldsymbol{R}^p\to\boldsymbol{R}^q$ ，其中 $\hat{f}(\boldsymbol{x})=\hat{\boldsymbol{B}}^{\mathrm{T}}\boldsymbol{x}$ ， $\boldsymbol{R}^p$、$\boldsymbol{R}^q$ 分别为第 p 维和 q 维响应矩阵； $\hat{\boldsymbol{B}}$ 为估计的系数矩阵； $\boldsymbol{x}$ 为输入元素矩阵。基于分支最小二乘算法的多变量 L_2-boosting 回归预测模型的步骤如下：

1）初始化。令 $\hat{f}_k^0(\cdot)=0$ ， k=1,⋯,q，初始化迭代次数为 m=1。

2）计算当前矩阵中元素的偏差 $\boldsymbol{r}_i^m=\boldsymbol{y}_i-\hat{f}^{(m-1)}(\boldsymbol{x}_{(i)})\ (i=1,2,\mathrm{L}\ ,n)$ ，并利用弱学习过程计算参数 $\hat{\boldsymbol{\beta}}_{st}^{(m)}$ ，进而得出估计函数 $\hat{g}^{(m)}(\cdot)$ ； $\boldsymbol{r}_i^m$ 为第 m 代、第 i 个元素偏差矩阵；$\boldsymbol{y}_i$ 为第 i 个响应矩阵； $\hat{\boldsymbol{\beta}}_{st}^{(m)}$ 为第 m 代系数矩阵 $\hat{\boldsymbol{B}}_{\hat{s}\hat{t}}$ 的估计矩阵。

3）方程更新， $\hat{f}^{(m)}(\cdot)=\hat{f}^{(m-1)}(\cdot)+v\cdot\hat{g}^{(m)}(\cdot)$ ，其中 v≤1。

4）令 m=m+1，调整步骤 2）进行循环计算，直到满足程序终止条件。

上述过程中，多变量 L_2-Boosting 算法在每次迭代中是对 $\boldsymbol{B}$ 中某一列元素的拟合，每次迭代均得到一组 $\boldsymbol{B}$ 的估计值 $\hat{\boldsymbol{B}}(m)$ ，使得 $\hat{f}^{(m)}(x)=(\hat{\boldsymbol{B}}^{(m)})^{\mathrm{T}}x$ 。当程序满足终止条件时， $\hat{f}_{\mathrm{stop}}^{(m)}$ 即为所求的回归预测模型 $\boldsymbol{E}[y|x=\cdot]$ 的估计方程。当然，当协变量 $\boldsymbol{B}_{jk}$ 对响应矩阵 $\boldsymbol{R}$ 中全部元素均有影响或者均无影响时，上述计算过程可能只是得到次优值，即算法拟合效果将达不到理想值。此时，可以将 $\boldsymbol{B}$ 中全部列在同次迭代中进行拟合，从而在每代中选出最优拟合加入到估计方程中，直到达到预期精度。然而，利用并行迭代计算使得计算过程异常复杂，常规方法已不能满足如此大规模计算。因此，本文将多变量 L_2-Boosting 回归模型与 Spark 平台相结合，使得大数据下短期负荷预测工作的适用性更强，预测精度更高。

2 基于 Spark 平台的 L_2-Boosting 负荷预测模型

2.1 Spark 技术简介

Spark 平台是在 Hadoop MapReduce 的基础上提出的新一代大数据分析框架[19]，拥有 Hadoop MapReduce 所具备的的全部优点，并且 Spark 是将计算结果直接存储在内存中，使得运算效率更高。此外，Spark 平台可以实现线上的机器学习、数据交互式分析等，各线程任务可直接从内存中调取所需数据，实现数据高度共享，从而提高了运算速度。随着智能电网的发展，第一时间线上大数据的高效率挖掘和应用显得尤为重要，基于 Spark 基础特有的性质，本文将 L_2-Boosting 模型与 Spark 数据分析框架相结合，实现分布式能源系统的短期负荷预测。

为支持多次迭代运算，Spark 平台提供了 2 个重要概念[20]。1）弹性分布式数据集(resilient distributed datasets，RDD)。2）共享内存。每个 RDD 是只读的，可以通过其他 RDD 上的批量操作创建；在并行操作中，任务之间的变量、任务与驱动程序之间变量是相互共享的。RDD 提供了 4 种重要算子：1）输入算子。数据从外部空间输入 Spark 平台，进而转化为运行数据块，并通过 Block Manager 进行管理。2）变换算子。可通过变换算子(如 fliter 等)对数据进行操作，并将 RDD 转化为新的 RDD 数据集。3）缓存算子。可通过 Cache 算子将数据缓存于内存，避免重新调用数据的麻烦，提高运算效率。4）行动算子。通过数据管理中心将任务指派在各 RDD 区域，通过并行计算执行各区域任务。

2.2 基于 Spark 平台 L_2-Boosting 回归模型的实现

本文所提分布式能源系统短期负荷预测系统流程如图 1 所示。

图 2 展示了本文设计的分布式能源系统负荷预测并行算法流程。

如图 2 所示，负荷相关数据搜集后通过 RDD 转化将其储存于内存空间中，通过聚类管理中心(Cluster manager)将数据进行分块，进而将各分块任

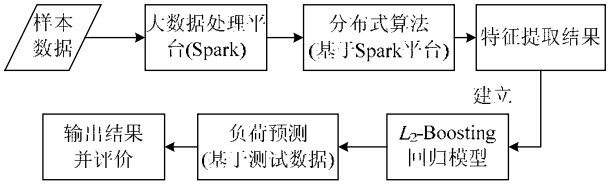

图 1 分布式能源系统短期负荷预测系统流程图

Fig. 1 Work flow of short term load forecasting system for distributed energy system

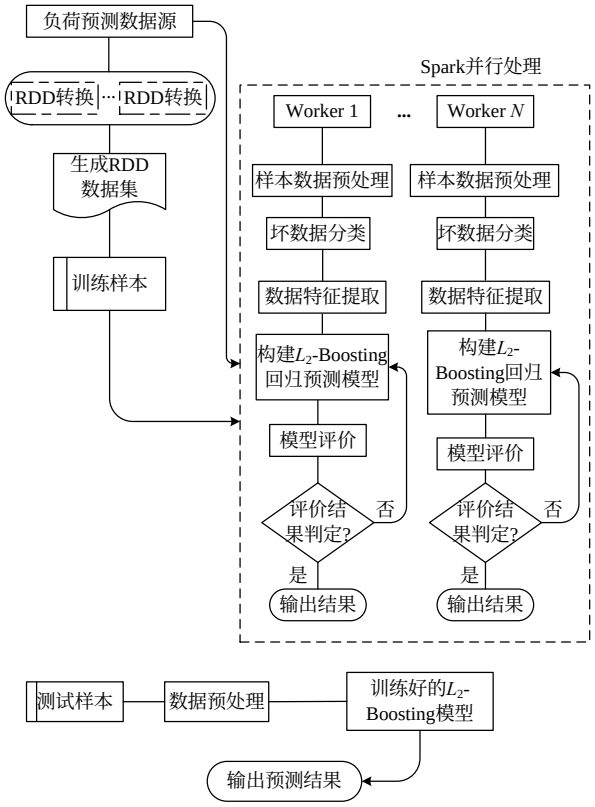

图 2　分布式能源系统负荷预测并行算法流程

Fig. 2　Parallel algorithm flow for load forecasting of distributed energy system

务指派在 N 个工作节点(Workers)进行并行处理。在指派的数据处理任务中，本文采用 Neville 法进行缺失数据处理、利用改进减法聚类进行坏数据分类以及通过计算不一致率进行特征选择。具体为：

1）基于 Neville 法[21]的缺失数据处理。

令 $\boldsymbol{Z}=\{z_1,z_2,\text{L}\ ,z_n\}$ 为负荷样本数据集，n 为节点数，z 为插值点，令 p=1、q=1，其中 $p\in[1,\ n]$，$q\in[1,p]$，则计算拉格朗日插值多项式为

$$L_{p,q}=[(z-z_{p,q})L_{p,q-1}-(z-z_p)L_{p-1,q-1}]/(z_p-z_{p-q}) \tag{7}$$

判断 $|L_{p,q}-L_{p-1,q-1}|<\eta$ 是否成立，若成立，则输出 $L_{p,q}$；若不成立，则继续计算，$z_{p,q}$ 表示第 p 行、q 列插值点，η 为期望精度。

2）基于改进减法聚类的坏数据分类。

监控设备由于外部故障等原因会造成观测数据反常，以至于大多数观测值不一致；此外，外部特殊事件的发生也会引发数据异常现象，这在一定程度上干扰了负荷规律的变化，导致负荷预测误差偏大，影响负荷预测模型的精确性，因此，需要对坏数据进行有必要的分类和处理。本文采用改进模糊 C 均值聚类方法[19]对坏数据进行有效分类，其步骤如下。

①采用减法聚类方法计算出聚类数目和初始聚类中心 $\boldsymbol{Z}_{c_l}$。

②计算隶属度

$$u_{ic}^{(N)}=\{\sum_{j=1}^{l}[\frac{\|\boldsymbol{Z}_c-\boldsymbol{P}_i^{(N)}\|^{\frac{1}{m-1}}}{\|\boldsymbol{Z}_c-\boldsymbol{P}_j^{(N)}\|}]\}^{-1} \tag{8}$$

式中：N 为迭代计数器；$\boldsymbol{Z}_c$ 表示第 c 个聚类中心；$\boldsymbol{P}_i^{(N)}$、$\boldsymbol{P}_j^{(N)}$ 分别表示第 N 次迭代下第 i 个和第 j 个聚类中心矩阵；m 为加权指数; $\boldsymbol{P}_i^{(N)}=\boldsymbol{Z}_c$。

③计算目标函数值

$$\Phi(u,p)^{(N)}=\sum_{c=1}^{l}\sum_{i=1}^{n}u_{ic}^{(N)}\|\boldsymbol{Z}_i-\boldsymbol{P}_c^{(N)}\|^2 \tag{9}$$

式中：$\boldsymbol{Z}_i$ 表示第 i 个聚类中心；$\boldsymbol{P}_c^{(N)}$ 表示第 N 次迭代下第 c 个聚类中心矩阵。

④判断 $\boldsymbol{\Phi}(u,p)^{(N)}$ 是否达到最小值，如果达到，则输出聚类类别数 $\boldsymbol{Z}$ 与聚类中心矩阵 $\boldsymbol{P}$；否则令 N=N+1 更新聚类中心矩阵并重新计算隶属度。聚类中心矩阵更新公式为

$$\boldsymbol{P}_i^{(N)}=\sum_{c=1}^{n}(u_{ic}^{(N)})^m\boldsymbol{Z}_c/\sum_{c=1}^{n}(u_{ic}^{(N)})^m \tag{10}$$

3）基于不一致率的特征选择。

海量负荷相关数据下特征选择的目的在于区别出于负荷相关性最强的特征子集，从而使得 L_2-Boosting 回归模型的输入向量具有较强的针对性，减少输入信息的冗余，从而提高负荷预测的精度。通过不一致率法进行快速特征选择，需要得知不一致率的计算方法。因此，假设所搜集的负荷数据拥有 g 项特征(如温度、湿度等)，分别由 $G_1,G_2,\ldots,G_g$ 的值表示；再假设标准类 M 拥有 c 个类别，具有 N 个数据实例，用 z_{ji} 表示特征 F_i 所对应的特征值，用 λ_i 表示 M 的类值，则数据实例可以表示为 $[\boldsymbol{z}_j,\lambda_i]$，其中 $\boldsymbol{z}_j=[z_{j1},z_{j2},\text{L}\ ,z_{jg}]$，则数据不一致率的计算公式[22]为

$$\tau=\frac{\sum_{k=1}^{p}(\sum_{l=1}^{c}f_{kl}-\max_l\{f_{kl}\})}{N} \tag{11}$$

式中：f_{kl} 为数据集中属于 x_k 模式的特征子集模式下数据实例的个数；x_k 为数据集共有 P 个特征划分区间模式(k=1,2,⋯,p；$p\le N$)。采用不一致率进行特征选择的步骤为：

①初始化最优特征子集为空集 $\boldsymbol{\Gamma}=\{\}$。

②计算 $\boldsymbol{\Gamma}$ 子集中与每个剩余特征组成的特征子集模式下数据集 $G_1,G_2,\ldots,G_g$ 的不一致率。

③选择最小不一致率所对应的特征 G_i 为最优特征，则更新最优特征集为 $\boldsymbol{\Gamma}=\{\boldsymbol{\Gamma},G_i\}$。

④采用顺序向前搜索策略，计算出特征子集的

不一致率统计表，并由小到大进行排列。

⑤选择特征个数尽量小的特征子集 $\boldsymbol{f}^{(k)}$，判断 $\varepsilon_{\boldsymbol{f}^{(k)}} \approx \varepsilon_{\Gamma}$ 是否成立，若成立，则 $\boldsymbol{f}^{(k)}$ 为被选中的特征子集。其中，$\varepsilon_{\boldsymbol{f}^{(k)}}$ 为被选中特征子集 $\boldsymbol{f}^{(k)}$ 的不一致率，ε_{Γ} 为最优特征子集 $\boldsymbol{\Gamma}$ 的不一致率。

在 Spark 框架的具体实现中，首先将搜集到的数据集封装于类 Datapoint[]中，该类由向量 $\boldsymbol{x}$ 与向量 $\boldsymbol{y}$ 组成，其中 $\boldsymbol{x}$ 代表模型输入向量，$\boldsymbol{y}$ 代表负荷。将全部训练数据集进行一次 RDD 处理，通过 Spark.Context 中的 parallelize()函数将其转化为 RDD 集，称为 RDD1；其次，将 RDD1 缓冲于内存中，调用 Spark 并行处理进行迭代计算，并将计算结果返回于 Feature[]集合中，再次将其进行 RDD 转化，称为 RDD2；再者，采用 map 算子将 RDD2 中的数据通过并行化处理构造输入向量权值，采用 reduce 算子将各模块中的向量值进行加和，计算出每次迭代总的权值向量 $\bar{\boldsymbol{w}}$，并储存于集合 weights[]中，记为 RDD3；最后，通过选择最优权值来构建 L_2-Boosting 负荷预测模型，并将测试集带入模型完成全部预测工作。

3 算例验证

算例数据选取北方某地区分布式能源系统 2015 年 4 月 20 日—2015 年 10 月 20 日的分布式负荷数据和气象数据，信息采集频率为 15 min，同时考虑最高温度、最低温度、风速、湿度、是否节假日、是否周末等作为特征选择候选集合。

本文数据表达方式设定为：令 y_i 为分布式负荷数据、x_{i1} 为温度、x_{i2} 为湿度、x_{i3} 为降雨量，以此类推直到全部候选特征设定完毕。由于采集的数据量有限，本文人为地将原数据集扩大到 2.48 G。其中，2015 年 4 月 20 日—2015 年 10 月 19 日被扩大的数据作为训练集，表示为 $[x_{i1}, x_{i2}, \mathrm{L}\ , x_{in}, y_i]$；2015 年 10 月 20 日的数据作为测试样本集，表示为 $[x_{i1}, \mathrm{L}\ , x_{im}, \mathrm{L}\ , x_{in}, y_i]$，如表 1 所示。

表 1 样本数据

Tab. 1 Sample data

样本	日期	时间	最高温度	最低温度	相对湿度/%	是否周末	是否节假日	分布式负荷/MW
训练样本	2015-04-20	00:00	17	6	82	—	—	5.68
	2015-04-20	00:15	17	6	86	—	—	6.04
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	2015-10-19	23:30	31	14	88	—	—	5.31
	2015-10-19	23:45	31	14	88	—	—	5.16
测试样本	2015-10-20	00:00	30	12	83	—	—	5.62
	2015-10-20	00:15	30	12	83	—	—	5.83
	2015-10-20	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	2015-10-20	23:30	30	12	87	—	—	4.97
	2015-10-20	23:45	30	12	87	—	—	4.70

负荷预测结果评价指标采用最大、最小相对误差(MaxRE、MinRE)、平均绝对百分比误差(MAPE)和几何平均相对绝对误差(GMARE), 各评价指标表达式[23]分别为

$$R_{\mathrm{E}} = \frac{\max(\left|\frac{y_i - y_i'}{y_i}\right| \cdot 100\%)}{\min(\left|\frac{y_i - y_i'}{y_i}\right| \cdot 100\%)} \tag{12}$$

$$M_{\mathrm{APE}} = \frac{1}{n}\sum_{i-1}^{n}|(y_i - y_i') / y_i| \cdot 100\% \tag{13}$$

$$G_{\mathrm{MARE}} = (\prod_{i=1}^{n}|\frac{y_i - y_i'}{y_i - \hat{y}_i'}|)^{\frac{1}{n}} \tag{14}$$

式中：y_i 为分布式负荷原始值；y_i' 为分布式负荷预测值；$\hat{y}_i'$ 为基本预测方法得到的预测值。式(12)—(14)中，MAPE、RE、GMARE 的值越小，负荷预测结果越准确。

采用本文所建立的基于 Spark 平台的分布式能源系统负荷预测流程对样本数据进行处理，缺失数据、坏数据处理以及特征选择结果如下所示：

1）缺失数据处理结果。

基于 Neville 法将样本数据中的缺失点补全。其中，算法的参数 n=5，η=0.01。其中部分缺失数据补全结果如表 2 所示。由表 2 可知，Neville 算法对样本数据进行补全整体效果较好，缺失点补全后的误差均在 2%以内。

2）坏数据处理结果。

将缺失数据补全后，样本数据集则具有连续性。再用改进模糊 C 均值聚类方法对坏数据进行分类和识别。其中，分布式负荷坏数据识别结果如表 3 所示。表 3 中：漏检数为将坏数据检测为正常数据的次数；误检数为将正常数据检测为坏数据的次数。

由表 3 可知，检测错误率仅为 0.740‰，检测

表 2　部分缺失数据补全结果

Tab. 2　Results of partial missing data complement

缺失点	计算值/MW	实际值/MW	误差值/%
19	5.239 8	5.245 0	−0.10
88	5.486 1	5.456 1	0.55
265	6.258 3	6.208 1	0.81
441	6.073 2	6.058 7	0.24
879	5.139 5	5.174 2	−0.67
993	5.896 1	5.943 1	−0.79
1018	6.255 0	6.188 2	1.08
1112	6.297 5	6.288 7	0.14
1347	5.437 0	5.499 1	−1.13
1725	5.673 9	5.647 4	0.47
2453	6.817 4	6.854 5	−0.54
2811	5.431 9	5.468 0	−0.66

表 3　分布式负荷坏数据辨识结果

Tab. 3　Results of bad distributed load data identification

数据总量	坏数据数量	漏检数	误检数	检测错误率/%
17 568	187	9	4	0.740

结果很好。识别坏数据后并对坏数据进行调整，调整后的分布式负荷数据误差更小，负荷曲线更加平滑，从而保证了分布式负荷数据的准确性。

3）特征选择结果。

按照不一致率进行特征选择的步骤，各个候选特征子集的不一致率计算结果如表 4 所示。

表 4　不一致率计算结果

Tab. 4　Results of inconsistent rate

序号	候选特征集	不一致率/%	序号	候选特征集	不一致率/%
1	x_{i1} x_{i2} x_{i3}	2.03	7	x_{i1} x_{i2} x_{i3} x_{i8} x_{i9}	7.96
2	x_{i2} x_{i3} x_{i4}	2.17	8	x_{i1} x_{i2} x_{i3} x_{i5} x_{i6} x_{i7} x_{i8}	8.62
3	x_{i6} x_{i3} x_{i4}	3.04	9	x_{i1} x_{i2} x_{i5} x_{i6} x_{i8} x_{i9} x_{i10}	10.14
4	x_{i4} x_{i5} x_{i6}	4.46	10	x_{i1} x_{i4} x_{i5} x_{i6} x_{i8} x_{i9} x_{i10}	13.5
5	x_{i2} x_{i4} x_{i5} x_{i6}	6.23	11	x_{i1} x_{i2} x_{i3} x_{i4} x_{i5} x_{i6} x_{i11}	16.33
6	x_{i3} x_{i5} x_{i6} x_{i7}	7.68	12	x_{i1} ~ x_{i12}	20.69

表 4 中：x_{i1}—x_{i12} 依次代表最高温度、最低温度、平均温度、前一天平均气温、前一天最高温度、是否周末、风速、前一天分布式负荷、相对湿度、前一天最低温度、降水量、是否节假日。因此，通过计算得出候选特征集 x_{i1}、x_{i2}、x_{i3}、x_{i8}、x_{i9} 的不一致率比最低，即通过不一致率法得到的特征选择结果为：最高温度、最低温度、平均温度、前一天分布式负荷、相对湿度。

通过上述数据处理及特征选择结果建立 Spark-L_2-Boosting 预测模型，并与支持向量机(support vector machine，SVM)回归预测模型、BP(back propagation)神经网络预测模型、RBF(radial basis function)神经网络预测模型进行比较，SVM 回归预测模型、BPNN 预测模型和 RBF 预测模型在电力负荷短期预测中表现除了出色的学习能力和泛化能力，通过与这 3 种算法的比较，测试本文所提出的 Spark-L_2-Boosting 分布式能源系统短期负荷预测模型的性能。4 种模型的预测结果如图 3 所示，模型评价指标计算结果如表 5 所示，模型预测结果误差对比如图 4 所示。

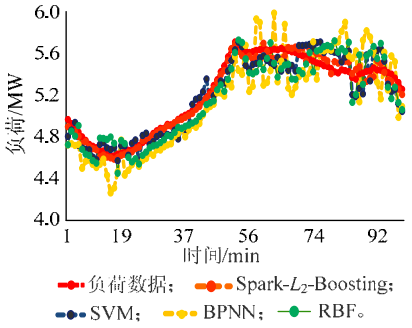

图 3　各模型预测结果对比

Fig. 3　Forecasting results of each model

表 5　模型评价指标计算结果

Tab. 5　Computation results of evaluation indexes of each model

指标	指标计算结果/%			
	Spark-L_2-Boosting	SVM	BP 神经网络	RBF 神经网络
R_E 最大值	1.96	4.04	8.53	5.48
R_E 最大值	0.01	1.01	1.06	1.13
M_{APE}	0.86	2.21	4.05	2.78
G_{MARE}	0.58	0.72	0.81	0.76

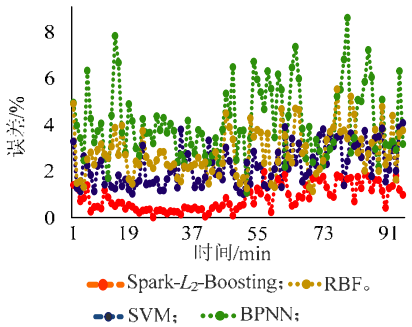

图 4　各模型预测结果误差对比

Fig. 4　Forecasting errors of each model

由表 5 可见，Spark-L_2-Boosting 预测模型的 M_{APE} 值为 0.86%，而 SVM、BP 与 RBF 预测模型的 M_{APE} 值分别为 2.21%、4.05%和 2.78%，这表明 Spark-L_2-Boosting 模型的预测精度更高，这是因为通过数据处理以及特征提取使得样本的可学习性能增强，从而降低了数据维度的复杂性和抽象性；此外，L_2-Boosting 模型通过不断迭代建立，这使得模型对训练样本数据的学习能力更强、泛化性能更优。Spark-L_2-Boosting 预测模型的 R_E 最大值为

1.96%，而 SVM、BP 与 RBF 预测模型的 R_E 最大值分别为 4.04%、8.53%和 5.48%；Spark-L_2-Boosting 预测模型的 R_E 最小值为 0.01%，而 SVM、BP 与 RBF 预测模型的 R_E 最小值分别为 1.01%、1.06%和 1.13%，该指标结果表明 Spark-L_2-Boosting 预测模型精度更高，预测更准确，预测效果更好。从评价指标 G_{MARE} 也可以看出，Spark-L_2-Boosting 预测模型的 G_{MARE} 值低于 SVM、BP 与 RBF 预测模型，负荷预测结果最为准确。

图 3 为测试样本真实分布式负荷与各个模型预测值的对比图。从图 3 可以看出，Spark-L_2-Boosting 模型的预测值曲线较其他模型更为平滑，且与真实分布式负荷曲线的贴进度更高，而其他 3 种算法的预测曲线出现较大波动，与真实分布式负荷数据的贴进度相对较低。负荷预测曲线进一步证明了本文所提 Spark-L_2-Boosting 算法预测性能更佳。如图 4 所示，Spark-L_2-Boosting 模型预测结果误差曲线最靠近横轴，表明其预测结果误差最小，分布式负荷预测效果整体最优。其次，SVM 模型预测效果优于 RBF 模型，而 BPNN 模型预测结果误差波动最大，预测效果较差。

此外，本文亦利用 Hadoop 平台与 L_2-Boosting 模型结合进行短期负荷预测，并与本文所提 Spark-L_2-Boosting 模型的并行计算效率进行对比。图 5 为 2 种算法并行计算效率对比结果。由图 5 可以看出：当样本数据小于某个值时，2 种并行算法的计算效率差别不大；但随着数据量的进一步增大，Spark-L_2-Boosting 算法的计算效率越来越高，同等计算量所需时间远远小于 Hadoop-L_2-Boosting 算法，这是因为 Spark 平台将数据转化所得的 RDD 集直接存储于内存中，实现了并行结构之间的数据共享，使得数据提取更加方便、有效率，避免 Hadoop 中需要返回 HDFS 再提取数据的麻烦。

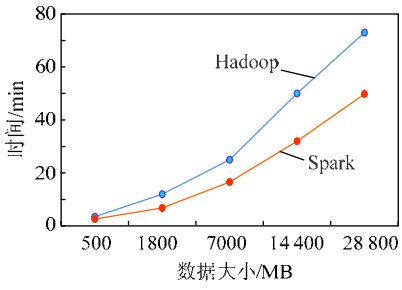

图 5 模型并行计算效率对比

Fig. 5 Parallel computation efficiency of each model

4 结论

本文针对分布式能源系统大数据特点，在 Neville 法的缺失数据处理、改进减法聚类的坏数据分类以及不一致率的特征选择的基础上，建立了基于 Spark 平台与多变量 L_2-Boosting 回归模型的分布式能源系统短期负荷预测方法，通过算例验证，主要得出以下结论：

1）通过建立 NGLF 损失函数和 CWLLS 弱学习过程的 Boosting 迭代算法，可帮助提高多元线性回归模型的学习能力和泛化能力，使模型在分布式能源短期负荷预测中的适用性更强、预测精度更高。

2）在分布式能源系统短期负荷预测中，本文所建立的 Spark-L_2-Boosting 模型的预测精度高于常规单机负荷预测模型(如 SVM、BPNN 等)，这是因为 Spark-L_2-Boosting 模型能够在海量高维数据下承担大规模计算消耗，使其更加适用于当前分布式能源系统短期负荷预测。

3）在分布式能源系统海量在线数据的背景下，Spark 平台的数据处理效率要远远高于 Hadoop 平台，原因在于 Spark 平台可通过系统内存实现数据自由共享，这不仅缩短了预测所消耗的时间，还提高了资源利用效率，显著提高了短期电力负荷预测的精度。

本文所提出的 Spark-L_2-Boosting 模型为分布式能源系统负荷预测提供了新的方法和思路，具备一定的可操作性和可扩展性；在下一步的工作中，如何将其他回归模型(如支持向量机回归模型、神经网络回归模型)与 Spark 平台相结合，如何将其他数据处理方法应用于 Spark 平台中，以及如何更加精确地在众多分布式能源系统负荷影响因素中选取关联性大的特征等将是今后重点研究方向。

参考文献

[1] 曾鸣. 新能源电力系统与能源互联网[N]. 国家电网报, 2015-06-01(6).

[2] 杨勇平. 分布式能源系统[M]. 北京：化学工业出版社, 2011：274.

[3] 黎祚, 周步祥, 林楠. 基于模糊聚类与改进 BP 算法的日负荷特性曲线分类与短期负荷预测[J]. 电力系统保护与控制, 2012, 40(3)：56-60.
Li Zuo, Zhou Buxiang, Lin Nan. Classification of daily load characteristics curve and forecasting of short-term load based on fuzzy clustering and improved BP algorithm[J]. Power System Protection and Control, 2012, 40(3)：56-60(in Chinese).

[4] Kang J, Zhao H. Application of improved grey model in long-term load forecasting of power engineering[J]. Systems Engineering Procedia, 2012(3)：85-91.

[5] 刘文颖, 门德月, 梁纪峰, 等. 基于灰色关联度与 LSSVM 组合的月度负荷预测[J]. 电网技术, 2012, 36(8)：228-232.
Liu Wenying, Men Deyue, Liang Jifeng, et al. Monthly load forecasting based on grey relational degree and least squares support

vector machine[J]. Power System Technology, 2012, 36(8): 228-232(in Chinese).

[6] Bashir Z A, El-Hawary M E. Applying wavelets to short-term load forecasting using PSO-based neural networks[J]. IEEE Transactions on Power Systems, 2009, 24(1): 20-27.

[7] 沈沉, 秦建, 盛万兴, 等. 基于小波聚类的配变短期负荷预测方法研究[J]. 电网技术, 2016, 40(2): 521-526.
Shen Chen, Qin Jian, Sheng Wanxing, et al. Study on short-term forecasting of distribution transformer load using wavelet and clustering method[J]. Power System Technology, 2016, 40(2): 521-526(in Chinese).

[8] 熊浩, 李卫国, 黄彦浩, 等. 基于模糊粗糙集理论的综合数据挖掘方法在空间负荷预测中的应用[J]. 电网技术, 2007, 31(14): 36-40.
Xiong Hao, Li Weiguo, Huang Yanhao, et al. Application of comprehensive data mining method based on fuzzy rough set in spatial load forecasting[J]. Power System Technology, 2007, 31(14): 36-40(in Chinese).

[9] 万昆, 柳瑞禹. 区间时间序列向量自回归模型在短期电力负荷预测中的应用[J]. 电网技术, 2013, 36(11): 77-81.
Wan Kun, Liu Ruiyu. Application of interval time-series vector autoregressive model in short-term load forecasting[J]. Power System Technology, 2012, 36(11): 77-81(in Chinese).

[10] 刘文博, 傅旭华, 王蕾, 等. 电力负荷无迹卡尔曼阈值多频级 WNN 区间预估[J]. 电网技术, 2016, 40(2): 527-533.
Liu Wenbo, Fu Xuhua, Wang Lei, et al. WNN interval estimation algorithm for electric load forecasting based on threshold multi-frequency unscented Kalman filter[J]. Power System Technology, 2016, 40(2): 527-533(in Chinese).

[11] 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 Dewen, Sun Zhiwei. Big data analysis and parallel load forecasting of electric power user side[J]. Proceedings of the CSEE, 2015, 35(3): 527-537(in Chinese).

[12] Xiao T, Zhu J, Liu T. Bagging and Boosting statistical machine translation systems[J]. Artificial Intelligence, 2013, 195(1): 496-527.

[13] 周国雄, 沈学杰, 李琳, 等. 基于 AdaBoost 的网络入侵智能检测[J]. 系统仿真学报, 2014, 26(7): 1517-1521.
Zhou Guoxiong, Shen Xuejie, Li Lin, et al. Network intrusion intelligent detection based on AdaBoost[J]. Journal of System Simulation, 2014, 26(7): 1517-1521(in Chinese).

[14] Harirchi F, Radparvar P, Moghaddam H A, et al. Two-level algorithm for MCs detection in mammograms using diverse-adaboost-SVM[C]//2010 20th International Conference on Pattern Recognition (ICPR). Istanbul, Turkey: IEEE, 2010: 269-272.

[15] 胡德华, 郑东健, 付浩雁. AdaBoost-BP 模型在大坝变形预测中的应用[J]. 三峡大学学报: 自然科学版, 2015, 37(5): 5-8.
Hu Dehua, Zheng Dongjian, Fu Haoyan. Application of AdaBoost-BP model to dam deformation prediction[J]. Journal of China Three Gorges University: Natural Sciences, 2015, 37(5): 5-8(in Chinese).

[16] 蒋翠侠, 许启发. 再论线性回归模型的最小二乘估计与线性方程组的解[J]. 统计与决策, 2014(6): 8-11.
Jiang Cuixia, Xu Qifa. Further discussion about least squares estimation of linear regression model and the solution of linear equations[J]. Statistics and Decision, 2014(6): 8-11(in Chinese).

[17] Roman W L, Peter B. Boosting for high-multivariate responses in high-dimensional linear Regression[J]. Statistica Sinica, 2006(16): 471-494.

[18] 王想, 邓光明, 蒋远营. 零均值高斯 AR(p)模型参数的最小化残差平方和(RSS)下条件最大似然估计及其最优性[J]. 统计与决策, 2009(6): 35-37.
Wang Xiang, Deng Guangming, Jiang Yuanying. Conditional maximum likelihood estimation and its optimality under the condition of the minimum residual sum of squares of the parameters of the zero mean Gauss AR (p) model[J]. Statistics and Decision, 2009(6): 35-37(in Chinese).

[19] 王诏远, 王宏杰, 邢焕来, 等. 基于 Spark 的蚁群优化算法[J]. 计算机应用, 2015, 35(10): 2777-2780.
Wang Zhaoyuan, Wang Hongjie, Xing Huanlai, et al. Ant colony optimization algorithm based on Spark[J]. Journal of Computer Applications, 2015, 35(10): 2777-2780(in Chinese).

[20] 李霄, 贺成龙, 张广庆, 等. 基于 Spark 平台的海量电子对抗数据分析[J]. 指挥信息系统与技术, 2015, 6(2): 53-56.
Li Xiao, He Chenglong, Zhang Guangqing, et al. Analysis of massive electronic countermeasure data based on Spark platform[J]. Command Information System and Technology, 2015, 6(2): 53-56(in Chinese).

[21] 蒋雯倩, 李欣然, 钱军. 改进 FCM 算法及其在电力负荷坏数据处理的应用[J]. 电力系统及其自动化学报, 2011, 23(5): 1-5.
Jiang Wenqian, Li Xinran, Qian Jun. Application of improved FCM algorithm in outlier processing of power load[J]. Proceedings of the Chinese Society of Universities for Electric Power System & Its Automation, 2011, 23(5): 1-5(in Chinese).

[22] 陈铁明, 马继霞, Samuel H.Huang, 等. 一种新的快速特征选择和数据分类方法[J]. 计算机研究与发展, 2012, 49(4): 735-745.
Chen Tieming, Ma Jixia, Samuel H.Huang, et al. Novel and efficient method on feature selection and data classification[J]. Journal of Computer Research & Development, 2012, 49(4): 735-745(in Chinese).

[23] Meng M, Niu D, Sun W. Forecasting monthly electric energy consumption using feature extraction[J]. Energies, 2011, 4(10): 1495-1507.

收稿日期：2015-12-30。

作者简介：


马天男

马天男(1992), 男, 博士研究生, 研究方向为输电线路覆冰预测、电力负荷预测、技术经济评价及预测, E-mail：matiannan_1234@126.com；

牛东晓(1962), 男, 博士, 教授, 博士生导师, 本文涉及课题负责人, 研究方向为项目预测与决策理论及其应用、项目综合评价方法及其应用等, E-mail：niudx@126.com；

黄雅莉(1991), 通信作者, 女, 硕士研究生, 研究方向为输电线路覆冰预测、输配电网评估方法及应用, E-mail：huangyali5210@163.com；

杜振东(1971), 男, 高级工程师, 本科, 研究方向为电网规划设计、输变电工程设计, E-mail：hzdzd@163.com。

（责任编辑　徐梅）