长兴岛产业园区电力数据分析及用户需求方法研究

王月强，张文博，仇张权，朱 铮，童潇宁，王加祥

（国网上海市电力公司长兴供电公司，上海 201913）

摘要：在工业园区的迅速建设中，园区内的电力需求也在不断增加，电网对园区内的负荷分配也在不断改变。在建设过程中用户的负荷数据缺失、异常等问题，会对园区内电网规划和实施存在干扰。文章针对长兴岛产业园区已经通电的用户，对数据调研、数据分析和电力需求分析方法等进行了研究，为园区内电力需求和负荷预测研究提供了重要依据。

关键词：数据分析；电力需求；负荷预测；随机森林

Research on power data analysis and user demand method of Changxing island industrial park

WANG Yueqiang, ZHANG Wenbo, QIU Zhangquan, ZHU Zheng, TONG Xiaoning, WANG Jiaxiang

(Changxing Power Supply Company of Shanghai Electric Power Company, Shanghai 201913, China)

Abstract: In the rapid construction of industrial parks, the demand for power in the park is also increasing, and the load distribution of power grid in the park is also changing. In the process of construction, there may be some problems such as the lack of load data and abnormality of users, which may interfere with the planning and implementation of power grid in the park. Data investigation, data analysis and power demand analysis methods are carried out for the users who have been connected to Changxing island industrial park, which provides an important basis for the study of power demand and load forecasting in the park.

Key words: data analysis; power demand; load forecasting; random forest

0 引言

电网在运行环境中，需要对各个节点进行电量监控，监控数据的准确性对电网的安全运行起到重要作用，然而在实际测量中，时常有数据受污染产生噪声数据。可通过曲线外推方法处理异常数据，规定优先条件，为修复值赋权[1]；可通过历史数据，得出其上下限值，从而对异常数据进行清洗与预处理[2]；可通过小波分析修复异常数据，补全日电力负荷数据[3]。在处理好异常数据后，可基于历史数据分析用户的电力需求。综合用户负荷特性、自然环境影响因素预测电力需求[4]；考虑季节对用户电力需求的影响，分配各种因素的权值，来分析用户的短期电力需求[5]；可根据历史数据采集日负荷样本，利用随机森林方法对样本和相关影响因素进行训练，得出短期负荷预测的结果[6]。综合电网、分布式电源、区域资源和地理分布等分析园区内需求[7]，根据城市发展GDP增长，得出电力需求的增长规律[8]。

本文主要论述正在发展建设中的园区的负荷

数据和电力需求等问题。通过对园区内用户的总体情况调研，获取与用户相关联的电力数据，对历史数据进行比对，对异常数据的处理方法进行研究，提出用户电力需求分析的多种方法。

1 电力需求现状及调研

1.1 园区概况

长兴海洋装备产业基地开发有限公司为崇明县政府产业开发园区，位于长兴岛中东部，东至兴港路，西至兴冠路，北至潘圆公路，南至江南大道。规划面积为7.13 km^2，与长兴岛南岸的船舶、海洋工程装备及港口机械的总装基地相邻。目前，园区入驻的生产企业主要以船舶修造的相关配套企业为主，如沪东三造、油嘴油泵、线缆和船舶电器制造企业。长兴岛海洋装备产业园区内主要以工业用户为主，该园区内的电力用户以及主要类型为工业用电。

2015年12月，园区内已送电用户数共有9户，均为10 kV高压用户，其中基建施工临时用电用户2户，总容量为16 520 kV·A，2015年电量总电量为17 142 325 kW·h，园区内最高负荷为10.8 MW。2018年12月，园区内已经送电的用户达到18户，配套的设备和电力需求也相应增加。长兴岛海洋装备园区目正在建设中，园区内的用户和设备正在不断增加，用电的需求呈现上升的趋势。通过对园区内电力数据的分析和电力需求的研究，可以更进一步了解用户需求，对园区内电网的建设和运行具有重大意义。

1.3 园区电力数据的收集与整理

为完成长兴岛产业园区电力需求的研究，对园区内各个用户的负荷数据进行收集与整理，如表1所示。

园区内现已通电用户数量为12户，总运行容量为23 760 kW·A，其中大工业用户11户，非工业、学校用户1户。

在确定好园区内的用户数量和用户名称后，开始进行各用户电能量使用情况调研，如表2所示。

表1 长兴岛产业园区用户电量

Tab. 1 Customer power quantity of Changxing island industrial park

序号	单位	运行容量 / kV·A	类型
1	上海长兴金属处理有限公司	4 100	大工业用户
2	上海兴中船舶设备基地	1 000	大工业用户
3	上海工程技术管理学校	2 850	非工业（学校）
4	上海中华造船船用管件厂	3 830	大工业用户
5	沪东造船阀门有限公司	1 800	大工业用户
6	上海常鑫船舶设备有限公司	630	大工业用户
7	上海超成电子科技有限公司	1 260	大工业用户
8	上海中船长欣线缆配套有限公司	400	大工业用户
9	上海海洋高新技术投资发展有限公司	2 000	大工业用户
10	上海润稼农业科技有限公司	1 040	大工业用户
11	上海界宏包装有限公司	2 000	大工业用户
12	中国船舶重工集团公司第704 研究所	2 850	大工业用户

表2 长兴岛产业园区用户电能数据清单

Tab. 2 Energy data list of Changxing island industrial park user

序号	SNAME	时间粒度 /min	时间长度 / 月
1	cx_bhcb_K3_Ib	60	24
2	cx_lhsyz_K18_Ib	60	24
3	cx_lhsyz_K2_Ib	60	24
4	cx_tx_K12_Ib	5	24
5	cx_tx_K19_Ib	5	24
6	cx_tx_K13_Ib	5	24
7	cx_tx_K15_Ib	5	24
8	cx_tx_K16_Ib	5	24
9	cx_tx_K14_Ib	5	24
10	cx_tx_K17_Ib	5	24
11	##_hygx_K23_Ib	60	24
12	cx_hyzb_K31cx_Ia	5	24
13	cx_hyzb_K32hdfm_Ia	5	24
14	cx_hyzb_K33xc1_Ia	5	24
15	cx_hyzb_K34xc2_Ia	5	24
16	cx_hyzb_K35xn1_Ia	5	24
17	cx_tx_K328_Ib	5	24
18	cx_tx_K329_Ib	5	24
19	cx_tx_K7_Ib	5	24
20	cx_tx_K10_Ib	10	24
21	cx_tx_K5_Ib	5	24
22	cx_tx_K6_Ib	5	24

长兴岛产业园区内的电能源全部由同心站提供，同心站通过开关站、环网站或直接与用户站相连，将电能量传输至用户侧，所以同心站的输出线路能直接或间接反映园区内用户的电力需求现状。表2为长兴岛产业园区用户电能数据清单，SNAME为连接向各个用户的电网侧电站的线路名称，根据实际情况获取每条线路的电能输出数据，时间间隔为5 min或60 min，持续获取数据的时间为24个月。

2 电力数据分析和预处理

2.1 数据比对及分析

在用户电力数据分析中，以一个完整周期及1年的负荷数据为研究对象，研究用户的用电行为。一般变电站有两个变压器，以保证用户的供电安全，通过长兴岛海洋装备产业园区内全部负荷都需经过的变压器可反映出园区的总体负荷情况。产业园区变压器1和变压器2的负荷情况如图1～图6所示。

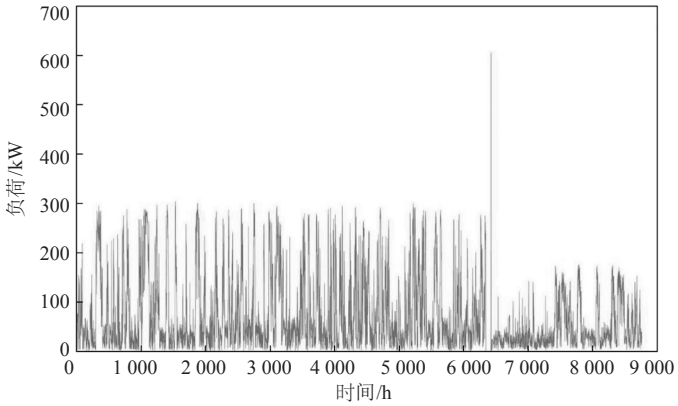

图1 变压器1年负荷曲线

Fig. 1 Transformer 1 year load curve

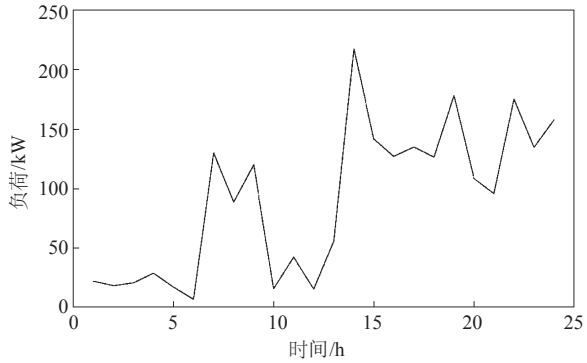

图2 变压器1日负荷曲线1

Fig. 2 Transformer 1 day load curve 1

图1为变压器1年负荷曲线，时间段为2017年9月1日至2018年8月31日，共8 760 h。2018 年5月23～26 日负荷数据异常（几乎为零）；2018年5月26日以后负荷水平明显下降，且日负荷曲线变化明显。

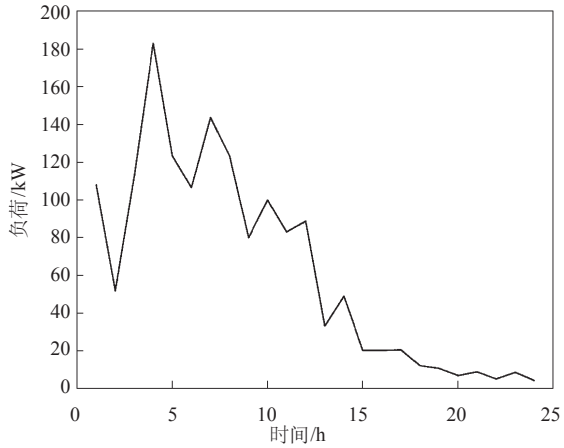

图3 变压器1日负荷曲线2

Fig. 3 Transformer 1 day load curve 2

图2、图3为变压器1的日负荷曲线。由于园区内的风力发电经由变压器1的原因，所以变压器1中日负荷特性变化剧烈，几乎没有日、周和年的规律性，与变压器2负荷特性差别较大。

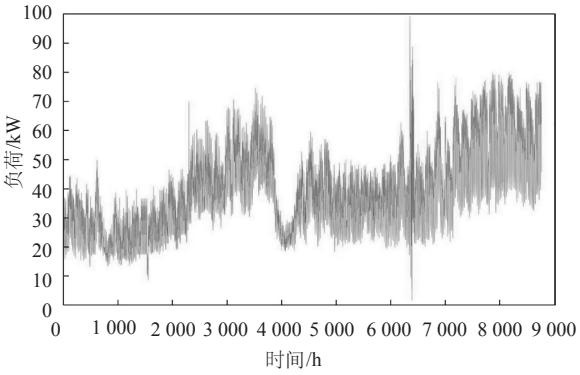

图4 变压器2年负荷曲线

Fig. 4 Transformer 2 year load curve

图4为变压器2年负荷曲线，时间段为2017年9月1日至2018年8月31日，共8 760 h。周期性明显。

图5、图6为变压器的日负荷曲线，比较符合电力用户的用电特性。

2.2 电力数据分析和预处理

根据对实际采集数据的整理发现原始数据中

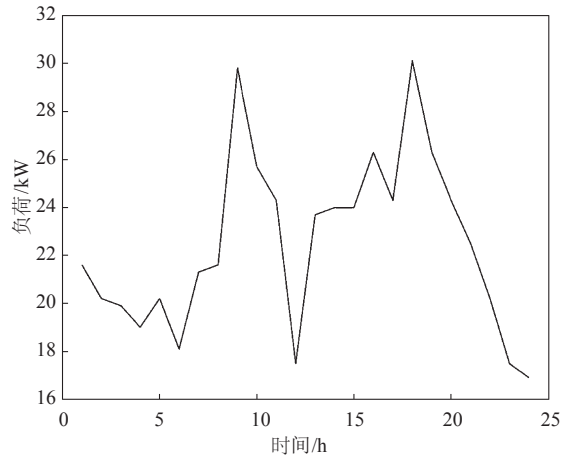

图5 变压器2日负荷曲线1

Fig. 5 Transformer 2 day load curve 1

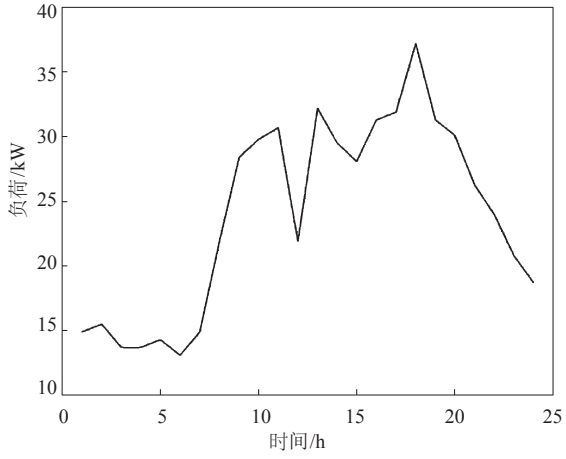

图6 变压器2日负荷曲线2

Fig. 6 Transformer 2 day load curve 2

存在有缺失数据，这些缺失数据的存在会影响负荷预测的准确度，不仅增加负荷预测的难度，而且还会增加数据序列的整体噪声。因此需要对这些缺失数据作处理，找到合理的数值去替换缺失数据，进一步提高数据的质量。在尽可能地保留原数据特性的同时，还要对缺失数据和异常数据进行修正，以得到可供算法输入的数据。处理缺失值的方法中较常用的是直接删除含有缺失值的记录，这种方法比较简单，但在负荷预测这个问题背景下，它不仅会减少数据的记录数，还会破坏数据的连续性，进而会改变用电负荷中潜在的特征。此外，还有一种比较常用的方法，该方法根据问题的实际背景用最合适的数值进行替代，可以使用全局变量、平均值、极值和众数等进行替换。由于负荷数据具有较强的周期性，影响因素也比复杂。因此，针对某一天缺失了很多数据的记录，利用相邻几天的正常数据进行加权处理，然后进行填充以减小误差，计算公式为

$$x(d,t)=\omega_1 x(d_1,t)+\omega_2 x(d_2,t) \quad (1)$$

式中，$x(d,t)$为第d天第t小时对应的负荷值；$\omega_1 x(d_1,t)$为第$d-1$天第t小时对应的负荷值；$\omega_2 x(d_2,t)$为第$d+1$天第t小时对应的负荷值，其中$\omega_1=\omega_2=0.5$分别表示前一天及后一天负荷的权重。

这种方式综合考虑了缺失值最接近两天的负荷值，弱化了其他因素的影响，可以更合理地填充有空缺的负荷值，进而保留了负荷数据的连续性。

由于采集的数据为母线电流值，因此在负荷数据的处理过程中发现在某些字段的一些字段值会与其他大部分记录有明显的不同，且没有其他具体的含义，那么就认为这些记录是异常的数据。如果不进行处理，将它们和其他正常数据混淆在一起，会增大噪声进而影响预测结果，因此需要对异常值进行处理。同样也有两种方案处理异常值：一种是直接删除包含异常值的记录；另一种是对异常值进行平滑处理。同缺失值填充方法一致，直接删除含有异常值的记录虽然简单但会导致数据不连续的问题。因此，本文通过箱线图分析找到异常值后，选择第二种方式对异常数据进行平滑处理。该方法只考虑异常值前一天和后一天对应的数据，并对其进行加权平均替代异常值。异常值的处理和缺失值的填充方法是相似的，都是以一个更合理的数值替换不合理的数值。式（1）的平滑操作考虑到了负荷数据的时间特性，即相邻几天同一时刻的负荷值是差不多的，因此能够找到一个较接近真实值的合理数据来替换。

由于原始数据中不同特征的值域存在较大的差异，例如：母线的电流数据小于5。如果不对这些数据作处理而直接分析，值域范围大的特征将掩盖掉值域范围小的特征，值域小的特征会被用电量这种值域大的特征掩盖掉，进而无法得到有效利用。为此需要对原始数据的各个特征进行规范化，常见的规范化方法如下。

（1）区间规范化

根据特征值的区间的边界值将数值归一化到[0，1]区间，计算公式为

$$x'_i = \frac{x_i - X_{\min}}{X_{\max} - X_{\min}} \quad (2)$$

式中，$X_{\min}$、$X_{\max}$分别为特征X的最小值和最大值；X_i为特征X的某个值；X'_i为规范化后的值。

该方法通过特征中的最大、最小值将特征值规范化到[0，1]。

（2）标准规范化

根据特征值得统计特性将数值归一化，但不一定在[0，1]区间上，计算公式为

$$x'_i = \frac{x_i - \bar{X}}{s} \quad (3)$$

式中，$\bar{X}$、s分别为特征X的平均值和标准差；x_i为特征X的某个值；x'_i为规范化后的值。

该方法要事先计算特征的均值及标准差这两个统计特性，需要多次遍历数据表。本项目使用的是区间规范化方法对原始数据进行处理，将所有特征值映射到[0，1]区间。该方法计算量较小，同时较好地保留了原特征的分布。此外，如果在进行规范化过程中，发现某个特征的最大值和最小值相等，即该特征所有的取值都相同，这将会导致式（2）中的分母为0。此时，不对该特征规范化，而是将所有的数值替换为1。

由于原始数据中包含许多无意义的值，如果不进行处理直接进行建模则会影响负荷预测的结果。为了减少预测算法中的数据处理量，提高算法的预测效率，还需要对数据作特征选择。特征选择的过程主要包括统计分析方法以及特征选择方法来进行。以下分别对负荷数据和天气数据的特征选择进行阐述。经过统计分析，发现在负荷数据表中存在大量空值、重复值，或一些对分析作用不大的特征，并对时间尺度进行统一。例如，对STIME进行了筛选，仅留下了整小时数据。因此，在保证数据完整性的同时还能够减少数据量的前提下，对这些特征进行删除或者采取去重操作。

3 电力需求及预测方法研究

3.1 回归分析法

电力负荷的趋势变化部分代表了用电水平在较长发展周期内的变化趋势，主要受到宏观经济的影响。在发达国家，电力负荷趋势变量变化趋于平稳，用电保持较高水平。在我国，受到经济高速发展的影响，总体用电水平不断上升，趋势变量变化较大。为了能够体现在现行经济政策下用电水平的总体趋势，采用多元线性回归模型对趋势变量部分进行预测。

含误差项的广义多元线性回归模型为

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1} + \varepsilon_i \quad (4)$$

式中，$\beta_0 \cdots \beta_{p-1}$为回归系数；$x_{i1} \cdots x_{i,\,p-1}$为影响因素变量，在广义多元线性回归模型中，影响因素变量可以是连续量、离散量或者指示变量；ε_i为服从$N(0, \sigma^2)$分布的随机变量。

两边同取期望可得

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{p-1} x_{p-1} \quad (5)$$

式中，$x_1 \cdots x_{p-1}$为$p-1$个影响因素变量。

由式（5）可知，观测值y_i服从期望值为$E(y)$，方差为σ^2的正态分布。

因此，可以利用式（5）进行负荷预测，其中y为预测负荷值，$x_1 \cdots x_{p-1}$为影响负荷的相关因素，回归系数$\beta_1 \cdots \beta_{p-1}$可由线性回归的方法求出。在多数情况下，影响因素变量是定量的，例如随着用户数量的增加，负荷随之上升，若将负荷预测模型建立为关于用户数的线性模型，则用户数就是一个定量的变量。然而式（4）的广义多元线性回归模型并没有限制影响因素变量为定量变量，定性变量如日期类型，也可以虚拟变量的形式加入多元线性回归模型之中。

3.2 时间序列分析模型

时间序列分析模型自提出以来，被广泛地应用于各类预测之中，也是短期负荷预测的常用方法。时间序列预测的基本思想是：首先假设所分析的时间序列是由某个随机过程产生的，然后利用时间序列的原始数据建立一个描述该过程的模

型，并进行参数估计，之后运用所建立的模型，在已知时间序列在过去和现在的观测值的情况下，求得时间序列未来的预测值。

对于平稳性时间序列，时间序列分析模型分为三种：自回归模型（Auto Regression，AR）、移动平均模型（Moving Average，MA）和自回归－移动平均模型（Auto Regressive Moving Average，ARMA）。而对于非平稳序列，则可以采用累积式自回归－移动平均模型（Auto Regressive Integrated Moving Average，ARIMA）。一般电力负荷具有较强的周期性，且经过小波分析可将周期变化部分与趋势变化部分分离开来，因此可以应用X-12-ARIMA模型进行时间序列预测。

X-12-ARIMA程序处理过程主要由三个部分组成：regARIMA部分提供前向预测、后向预测和各种效应的先验调整，此部分用于预测以实现时间序列的延伸；X-11部分用于进行季节调整可以生成季节指数；诊断部分则主要检验经季节调整后的时间序列的稳定性。X-12-ARIMA程序处理的基本流程图如图7所示。

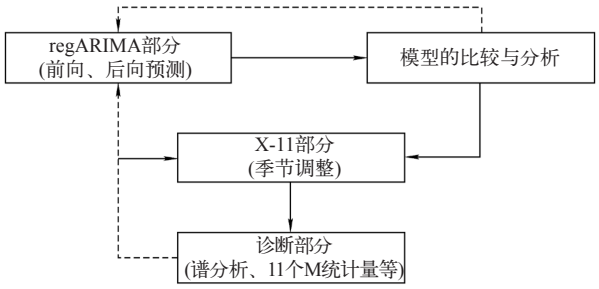

图7 X-12-ARIMA程序基本流程图

Fig. 7 Basic flow chart of X-12-ARIMA program

其中，ARIMA模型处理非平稳序列时，对其进行了一次或多次差分，从而将非平稳序列转化为平稳序列。将∇记为差分算子，则有

$$\nabla^2 y_r = \nabla(y_t - y_{t-1}) = y_t - 2y_{t-1} + y_{t-2} \quad (6)$$

对于延迟算子B有

$$y_{t-p} = B^p y_r, \forall p \geqslant 1 \quad (7)$$

因此可以得出$\nabla^k=(1-B)^k$。

设有d阶齐次非平稳时序列y_t，那么有$\nabla^d y_t$是平稳时间序列，则可以利用自ARMA(p，q)模型，即

$$\lambda(B)(\nabla^d y_t) = \theta(B)\varepsilon_t \quad (8)$$

或者

$$\lambda(B)(1-B)^d y_t = \theta(B)\varepsilon_t \quad (9)$$

式中，$\lambda(B)=1-\lambda_1 B-\lambda_2 B^2-\cdots-\lambda_p B^p$，$\theta(B)=1-\theta_1 B-\theta_2 B^2-\cdots-\theta_q B^q$分别为自回归系数多项式和滑动平均系数多项式；$\varepsilon_t$为零均值白噪声序列。

3.3 随机森林模型

随机森林（Random Forest，RF）是一种统计学习理论，其利用Bootstrap重抽样方法从原始样本中抽取多个样本，对每个Bootstrap样本进行决策树建模，然后组合多棵决策树的预测，通过投票得出最终预测结果。简而言之，随机森林是一种组合分类器，通过随机的方法建立由多棵决策树组成的森林，而且每棵决策树之间没有关联。大量理论和实验都证明了随机森林具有很高的预测准确度，对异常值和噪声具有很好的容忍度，而且不易出现过拟合。因此，适合于负荷数据中随机变化部分的预测。

随机森林的基本步骤如下，说明图如图8所示。

1）假定原始数据的记录数为N，随机森林的树木个数为k，每棵树的训练集样本个数为n，采用Bootstrap重采样技术从原始数据集N中随机抽取k个规模为n的训练样本集合。

2）假设有T个输入变量，每个节点都可以随机选择$t(t<T)$个指定的变量，接着运用这t个变量来决定最优的分裂点。在决策树的整个生成过程中，t值是保持恒定不变的。

3）每一棵决策树尽可能地生长，而不需要进行剪枝，进而构成k个决策树组成随机森林。

4）每输入一个新的样本，森林里面的每一棵决策树都会进行判断，判断该样本是属于哪一类。对于分类算法，最终会根据少数服从多数的方法进行决策，而对于回归算法则采用平均加权得到最终的预测结果。

在随机森林的步骤1）中，每棵树的建立依赖于从同一个数据集中独立抽取的样本集，从而保证了森林中的每棵树都具有相同的分布。步骤2）中，采用随机的方法去分裂每一个节点，然

后比较不同情况下产生的误差进而进行特征选择。它能够检测到内在估计误差、分类能力和相关性并依此决定选择特征的数目。步骤3）并不需要对每棵决策树进行剪枝操作，简化了决策树的生成过程，加快了决策树生成。虽然单棵树的分类能力有限，但是通过随机产生众多的决策树进行独立决策，再经过统计分析后，测试样本就可以选择最可能的分类。随机森林的随机性主要体现在两个方面：一方面是每棵树的样本都是通过Bootstrap放回抽样随机产生；另一方面，每个节点变量都是随机产生的，即在生成每棵树的时候，每个节点变量都只是从随机选出的少数变量中产生。

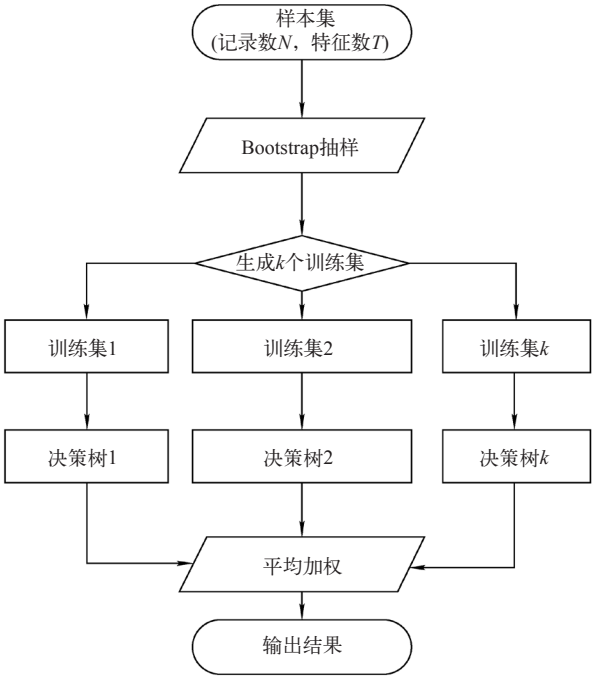

图8　随机森林预测模型说明图

Fig. 8 Illustration chart of stochastic forest prediction mode

4　结束语

本文主要针对长兴岛海洋装备产业园区电力需求的问题进行了研究。在课题研究之前，对园区内现有的用户及用电情况进行了调研；获取了该园区内用户的负荷数据并进行了比对；针对历史负荷数据缺失或异常的情况，给出了数据预处理的方法，保证数据的连续性、有效性等。在对园区内的历史负荷数据进行了异常剔除、缺失补偿后，提出了三种电力需求预测方法模型，能合理地对园区内的电力需求进行分析和预测。

参考文献

[1] 孙舒，刘莉，张瑛．基于典型负荷模式的异常数据处理[C]．北京：中国高等学校电力系统及其自动化专业学术年会，2008：1623-1626．

[2] 乔博鑫，章坚民．基于历史负荷的配电网量测数据关联清洗方法研究[J]．信息与电脑，2018(8)：70-71．
QIAO Boxin，ZHANG Jianmin．Research on relevant cleaning method of distribution network measurement data based on historical load[J]．China Computer and Communication，2018(8)：70-71．

[3] 朱加豪，赵振华，李建涛．基于小波分析的电力负荷数据研究[J]．中国新技术新产品，2017(10)：4-5．
ZHU Jiahao，ZHAO Zhenhua，LI Jiantao．Research on power load data of information and electricity based on wavelet analysis[J]．New Technology and New Products of China，2017(10)：4-5．

[4] 李黎，杨升峰，邱金鹏，等．电力系统供电短期负荷预测方法仿真研究[J]．计算机仿真，2017，34(1)：104-108．
LI Li，YANG Shengfeng，QIU Jinpeng，et al．Simulation study on short-term load forecasting method for power supply in power system[J]．Computer Simulation，2017，34(1)：104-108．

[5] 李啸骢，李春涛，从兰美，等．基于动态权值相似日选取算法的短期负荷预测[J]．电力系统保护与控制，2017，45(6)：1-8．
LI Xiaocong，LI Chuntao，CONG Lanmei，et al．Short-term load forecasting based on dynamic weight similar days selection[J]．Power System Protection and Control，2017，45(6)：1-8．

[6] 张冰，周步祥，石敏，等．基于灰色关联分析与随机森林回归模型的短期负荷预测[J]．水电能源

科学，2017，35(4)：203-207.
ZHANG Bing，ZHOU Buxiang，SHI Min，et al. Short-term load forecasting based on grey correlation analysis and random forest regression model[J]. Water Resources and Power，2017，35(4)：203-207.
[7] 李斌，丁艺．区域电力需求分析[J]．统计与决策，2007(3)：90-92.
LI Bin，DING Yi. Regional power demand analysis[J]. Statistics and Decision，2007(3)：90-92.
[8] 柴建，卢全莹，张钟毓，等．工业化、城镇化进程中电力需求分析及预测[J]．运筹与管理，2015(1)：164-172.
CHAI Jian，LU Quanying，ZHANG Zhongyu，et al. Analysis and forecast of electric power demand in the process of industrialization and urbanization[J]. Operations Research and Management Science，2015(1)：164-172.

作者简介

王月强（1982－），男，高级工程师，主要研究方向为电力自动化。

(收稿日期：2019－02－12)EA

王月强

(上接第 95 页)

指标检测与分析研究[J]．自动化仪表，2015，36(9)：84-87.
CUI Jian，WANG Jinmei，CHEN Jie，et al. Detection and analysis of performance index of photovoltaic grid-connected inverter[J]. Process Automation Instrumentation，2015，36(9)：84-87.
[8] 光伏发电专用逆变器技术规范：NB/T 32004—2013[S].
[9] Test procedure of islanding prevention measures for utility-interconnected photovoltaic inverters：IEC 62116—2008 [S].
[10] 光伏系统并网技术要求：GB/T 19939—2005[S].
[11] 光伏电站接入电网技术规定：Q/GDW 617—2011[S].
[12] 光伏电站接入电网测试规程：Q/GDW 618—2011[S].

作者简介

王康坚（1983－），男，助理工程师，主要研究方向为新能源调度。

(收稿日期：2018－08－15)EA

王康坚