大数据技术在电网数据上的应用

涂　炼[1]，张水平[1]，左　剑[1]，李　顺[1]，方海泉[2]，鲍　威[2]

（1. 广东电网有限责任公司电力调度控制中心，广东 广州 510030；
2. 浙江大学电气工程学院，浙江 杭州 310058）

摘要：为了有效预测电厂机组出力情况，不仅要采用先进的预测方法，还要对机组出力数据进行合理的预处理。提出了电厂机组出力数据分析的一整套流程，包括数据集成、异常值预处理、数据可视化和电厂机组出力预测。以某省 2018 年机组出力数据为研究对象，选取一个燃煤电厂为例。经过数据集成、数据预处理和可视化展示，并用长短期记忆（LSTM）神经网络对电厂出力进行预测，LSTM 预测得到的平均绝对百分比误差（MAPE）为 10.90%，预测结果优于误差反向传播（BP）神经网络，BP 神经网络预测得到的 MAPE 为 11.61%。说明经过预处理的机组出力数据再用 LSTM 模型预测能达到良好的预测准确度。

关键词：电厂；大数据；异常值；可视化；长短期记忆神经网络；预测

Application of big data technology in state grid data

TU Lian[1], ZHANG Shuiping[1], ZUO Jian[1], LI Shun[1], FANG Haiquan[2], BAO Wei[2]

(1. Electric Power Dispatching and Control Center of Guangdong Power Grid Company Limited, Guangzhou 510030, China; 2. College of Electrical Engineering, Zhejiang University, Hangzhou 310058, China)

Abstract: In order to effectively predict the output of power plant units, it is necessary not only to adopt advanced prediction methods, but also to reasonably preprocess the unit output data. A complete set of process of power plant unit output data analysis is proposed, including data integration, abnormal value preprocessing, data visualization and power plant output prediction, which can provide reference for power decision-making. We take the unit output data of Certain Province in 2018 as the research object, and take the coal-fired power plant as an example. After data integration, data preprocessing and visualization, the power plant output is predicted by using long short term memory (LSTM) neural network, and the mean absolute percentage error (MAPE) is 10.90%. The prediction result is better than the error back-propagation (BP) neural network, which MAPE is 11.61%. It is proved that the pre-processing unit output data can achieve good prediction accuracy by using LSTM model.

Key words: power plant; big data; abnormal value; visualization; long short term memory neural network; prediction

0　引言

机组出力是指发电机组的发电功率。随着信息技术的发展，现在已经能够实时获取并存储每一台发电机组的出力数据。大量的机组出力数据汇聚在一起构成一个机组出力大数据。

电厂机组出力预测，为降低运营成本和调节

南方电网公司科技项目资助（GDKJXM20190037）

电厂的负载均衡提供了参考依据。若与其他行业数据打通，对社会治理和城市管理将发挥更大的作用。

为了对电厂机组出力进行预测，需要先分析机组出力数据。机组出力数据具有五个特点：机组数量庞大，从区域范围看，全国各省市有众多电厂，每个电厂下面又有多个机组；机组类型多样，从机组发电类型看，有燃煤、水电、核电、风电和光伏等；机组出力数据具有实时性，每时每刻都在产生；机组出力数据具有时序性，比如每隔5 min记录一次数据；机组有新增或者退出。从这些特点可以看出机组出力数据具备大数据的特点，需要用大数据分析技术结合电力行业特点才能有效挖掘其中蕴含的价值。

大数据分析流程一般包括数据集成、数据预处理、可视化和预测。下面主要介绍数据集成和电厂机组出力预测相关研究进展。

数据集成是指按照一定的规则把不同来源的数据合并。数据集成对后续数据处理至关重要，在社会、医疗、电力和军事等领域得到广泛应用。文献[1]回顾了近年来数据集成技术。早期的数据集成技术主要采用字符串进行匹配[2]。随着互联网发展，有学者通过将互联网数据与企业数据相结合，可以生成更加精确的集成结果。随着大数据时代的到来，数据集成算法从传统的注重准确率开始向对算法效率和处理更加复杂的数据类型转变[3]。

电厂发电功率预测的相关研究较多。文献[4-6]应用长短期记忆神经网络对风电、光伏机组发电功率进行预测。文献[7]应用回声状态网络对光伏发电功率进行预测。文献[8]应用动态3次指数平滑法对火电厂发电量进行预测。

本文的创新之处包括两个方面：一方面，从一个局部来看，本文的创新之处体现在数据集成算法，通过数据集成算法可以把一个个孤立的数据表准确地对接起来，同时该算法的计算效率较高，对于数据量大的可以节省时间成本，经过数据集成和预处理可得到更加可靠的数据，为电厂机组出力的预测做好了铺垫；另一方面，从全局来看，把机组出力数据作为一个系统进行研究，给出机组出力数据分析的一整套流程，包括数据集成、异常值预处理、数据可视化和电厂机组出力预测，属于集成创新。

1 长短期记忆网络

人工神经网络（Artificial Neural Network, ANN）是一种数学模型，由大量的节点互相连接而形成的复杂网络结构，具有很好的函数逼近能力。用来处理序列数据的神经网络称为循环神经网络（Recurrent Neural Network, RNN）。长短期记忆（Long Short Term Memory，LSTM）网络是一种常见的循环神经网络，适用于时间序列预测。LSTM在自然语言处理、语言翻译、语音识别、视频分析和时间序列数据预测等领域都有广泛的应用，可以说LSTM非常适用于大数据的时间序列预测。

典型的LSTM单元结构如图1所示。

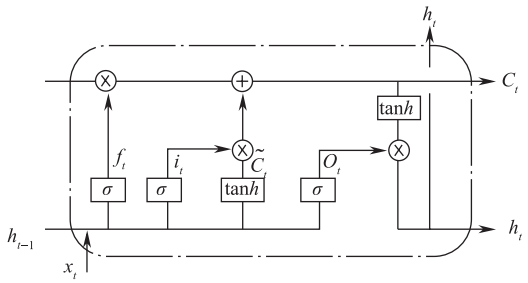

图 1 LSTM 单元结构图

Fig. 1 Unit structure diagram of the LSTM

该单元主要包含了三个门与一个记忆单元（Cell），三个门分别是遗忘门（Forget Gate）、输入门（Input Gate）和输出门（Output Gate）。方框内上方的那条水平线是单元状态（Cell State）。

LSTM第一步是用来决定什么信息可以通过单元状态，这个决定由遗忘门通过Sigmoid来控制，遗忘门的输出f_t与上一时刻的输出h_{t-1}和当前时刻的输入x_t有关，表达式为

$$f_t=\sigma(W_f[h_{t-1},x_t]+b_f) \tag{1}$$

式中，W_f和b_f分别为遗忘门中σ（Sigmoid）激活函数对应的权重参数和偏置项。

LSTM第二步是产生需要更新的新信息。当前时刻的单元状态C_t的计算方法如下

$$C_t = f_t C_{t-1} + i_t \tilde{C}_t \tag{2}$$

式中，$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i)$；$\tilde{C}_t = \tanh(W_c[h_{t-1}, x_t] + b_c)$；$i_t$表示输入门，决定哪些值用来更新；$W_i$和$b_i$分别为输入门中$\sigma$（Sigmoid）激活函数对应的权重参数和偏置项；W_c和b_i分别为tan$h$双曲正切激活函数对应的权重参数和偏置项；$\tilde{C}_t$表示生成新的候选值。

LSTM第三步是模型的输出，即

$$h_t = o_t \tanh(C_t) \tag{3}$$

式中，$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o)$，$W_o$和$b_o$分别为遗忘门中$\sigma$（Sigmoid）激活函数对应的权重参数和偏置项。

2 机组出力数据分析

机组出力数据分析流程如图2所示，包括数据集成、数据预处理、数据可视化以及电厂机组出力预测四个阶段。

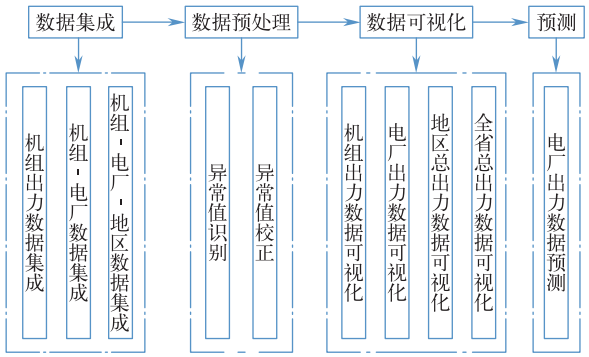

图 2 机组出力数据的分析流程

Fig. 2 Analysis process of unit output data

2.1 数据集成

数据集成阶段包括三个步骤：先对机组出力数据集成；再与电厂信息表集成得到机组-电厂数据；还需要与地区信息表集成得到机组-电厂-地区数据表。

2.1.1 机组出力数据集成

（1）机组出力数据介绍

机组出力数据是动态数据，每隔5 min记录并存储为一张Excel文件，文件名以时间命名，如“20180103_023500.xls”可以看出时间为2018年1月3日2点35分。以某省的机组出力数据为例，打开数据表“20180103_023500.xls”，可以看到该表记录了这个时刻全省所有机组的出力，该表有817行6列，这里只用到前两列，第一列是机组名称，第二列是机组对应的出力数据。每一行是一个机组的名称和该机组的出力数据，机组出力数据的单位为MW。

由于5 min一张数据表，一个小时有12张数据表，一天有288张数据表，一个月按30天计算有8 640张数据表，一年12个月就有103 680张左右的数据表，实际情况一年的数据表会少于这个数量，因为有少部分时间点的数据会有缺失。

（2）机组出力数据集成的算法

机组出力数据集成的目的是要把一张张分开的Excel表合并为一张表，合并好的这张表的每一行是一个机组的名称和该机组所有的出力数据，该机组的出力数据从左到右按时间序列排序。

由于每张机组出力数据表的第一列机组名称数量和排列顺序不一定一致，不能直接拼接，导致数据合并有一定难度。

把机组出力这样一个数据集成问题抽象出一个数学问题：已知有n个矩阵按时间序列排序命名为$\boldsymbol{H}_1$，$\boldsymbol{H}_2$，…，$\boldsymbol{H}_n$，每个矩阵的第一列是变量名，第二列是变量的取值，每个矩阵的第一列变量名数量不一致且排序也不一致，如何把这n个矩阵里出现的所有变量以及取值按时间序列合并为一个矩阵$\boldsymbol{A}_1$。

这里提出以下算法：

1）取第一个矩阵$\boldsymbol{H}_1$的第一列为一个列表记为List1，记这一列表元素的个数为m_1，定义一个（m_1+m_2）行n列的0矩阵$\boldsymbol{A}$（注：这里m_2是为将来新出现的变量预留位置，m_2可以人为设置，比如可以设置为100）。

2）依次对矩阵$\boldsymbol{H}_j$（j=1，2，…，n)进行遍历，对$\boldsymbol{H}_j$第1列的每一个变量$b_{j,k}$（k=1，2，…，w，这里w为$\boldsymbol{H}_j$第1列变量的个数）进行遍历，相当于两个“for”循环，在列表List1里查找该变

量，查找方法可用字符串匹配，查找结果有两种可能：①如果变量$b_{j,k}$在列表List1里，可以得到变量$b_{j,k}$在List1里的位置$i0$，把H_j中变量$b_{j,k}$对应的取值放到矩阵A的第$i0$行、第j列；②如果变量$b_{j,k}$不在列表List1里，在列表List1的底部新增该变量名，新增后的列表依然记为List1，这时List1的列表元素个数记为$i1$，把H_j中变量$b_{j,k}$对应的取值放到矩阵A的第$i1$行第j列。

3）当所有的矩阵H_1，H_2，…，H_n计算完之后，计算列表List1元素个数为$i2$，则说明最多有$i2$个变量，选取矩阵A的前$i2$行n列记为矩阵A_0，再把列表List1与矩阵A_0按行拼接得到矩阵A_1。矩阵A_1就是合并后得到的矩阵。

2.1.2 机组 - 电厂数据集成

机组出力数据集成得到矩阵A_1，需要进一步与电厂信息表集成。电厂信息表包括机组名称、机组发电类型和机组所属电厂名称3列。电厂信息表是一个静态信息表。

机组-电厂数据集成的算法思路如下：

1）把矩阵A_1分成两部分，第一列为列表List1，其他列为矩阵A_0。在列表List1的右边新增两列，用于存放机组发电类型和机组所属电厂，记为列表List2。

2）以矩阵A_1的第一列机组列List1为基准，对List1的每一个机组名称进行遍历，在电厂信息表的机组名称列通过字符串匹配查找，查找结果有两种可能：①如果查找到，则把电厂信息表里该机组对应的机组发电类型和机组所属电厂放到列表List2；②如果没有查找到，说明机组发电类型和机组所属电厂为空，则在机组名称列的后面以“NULL”+“NULL”填补，也放到列表List2中。

3）遍历完之后得到的列表List2与矩阵A_0拼接得到矩阵A_2，可以知道A_2的前3列为列表List2，矩阵A_2即为机组-电厂数据集成表。

2.1.3 机组 - 电厂 - 地区数据集成

地区信息表包括电厂名称和所属地市两列。地区信息表也是一个静态信息表。

机组-电厂-地区数据集成的算法思路如下：

1）把矩阵A_2分成两部分，前3列为列表List2，后面其他列为矩阵A_0。在列表List2的右边新增1列，用于存放电厂所属地市，记为List3。以列表List2的第3列电厂名称列为基准，对List2的每一个电厂名称进行遍历，在地区信息表查找该电厂所属地市，查找到的结果放到列表List2第4列。若没有查找到，以“NULL”填补。

2）遍历完之后得到的列表List3与矩阵A_0拼接得到矩阵A_3，A_3为机组-电厂-地区数据集成表。

为了便于分析，矩阵A_3需要加上表头，表头为一行，这一行的前4个元素分别为机组名称、机组发电类型、机组所属电厂名称和机组所属地市名称。表头这一行的其他元素为时间，由于每个机组出力数据表的文件名是以时间命名的，所以可以得到矩阵A_0每一列的时间。矩阵A_3加上表头后记为矩阵A_4。

2.2 数据预处理

数据预处理主要包括异常值识别与校正。对于燃煤电厂，一般认为机组出力数据大于中位数的3倍或者小于中位数的1/3为异常值，把异常值替换为0。异常值校正方法是用中位数替换0值。

2.3 数据可视化

“机组-电厂-地区数据集成表”经过异常值处理之后可以对数据进行可视化展示。根据需要可做出机组出力、电厂出力、地区总出力和全省总出力的时间序列图。电厂出力、地区总出力和全省总出力的计算方法是把目标范围的机组选出并把数据对应相加得到一行新的时间序列数据，从而可做出时间序列图。

2.4 电厂机组出力预测

经过预处理得到电厂出力数据可预测电厂机组出力情况。电厂机组出力预测包括四个步骤：①建立电厂机组出力预测模型，这里采用LSTM神经网络模型，与传统的BP神经网络模型进行对比；②准备数据，包括训练集数据和测试集数据，以一个燃煤电厂为例，从经过预处理的电厂

出力数据中选取该电厂每天一个时间点的数据，比如上午10点的数据，得到时间序列数据；③把训练集数据输入模型并进行训练；④模型训练好之后，用测试集数据评估模型的预测效果。

3 实例研究

3.1 研究对象

选取某省发电机组出力数据为研究对象，时间范围为2018年1月1日～12月31日一整年。发电机组发电类型包括燃煤、水电、核电、风电和光伏等，本文以燃煤电厂为例，对机组出力数据进行整理、分析，包括数据集成、数据预处理、数据可视化和机组出力预测。

3.2 计算环境

由于数据量较大，数据集成在服务器上运行完成。再把集成好的数据从服务器传到个人计算机。数据预处理、数据可视化和电厂机组出力预测在Win10系统上运行，在Anaconda下使用Spyder开发程序，程序语言为Python。机组出力预测是在TensorFlow框架下用Keras搭建LSTM神经网络。

3.3 数据集成

把机组出力数据、电厂信息表和地区信息表集成，表1为数据集成的部分示例。

表 1 机组 - 电厂 - 地区数据集成表

Tab. 1 Unit-plant- regional data integration

机组名称	类型	电厂	地市	20180101_000000	20180101_000500	20180101_001000
ZA1 厂 /20 kV.1 号机	燃煤	BLC	NB	371.10	371.10	375.53
ZA1 厂 /20 kV.2 号机	燃煤	BLC	NB	0.00	0.00	0.00
ZA1 厂 /20 kV.3 号机	燃煤	BLC	NB	387.57	387.57	407.79
ZA1 厂 /20 kV.4 号机	燃煤	BLC	NB	387.09	387.09	395.40
ZA1 厂 /20 kV.5 号机	燃煤	BLC	NB	391.34	391.34	407.86
ZA1 厂 /27 kV.6 号机	燃煤	BLC	NB	626.48	626.48	626.18
ZA1 厂 /27 kV.7 号机	燃煤	BLC	NB	616.45	616.45	637.89
ZA1 厂 /22 kV.1 号机	燃煤	LXC	JH	0.00	0.00	0.00
ZA1 厂 /22 kV.2 号机	燃煤	LXC	JH	383.87	383.87	398.57
ZA1 厂 /22 kV.3 号机	燃煤	LXC	JH	380.81	380.81	394.59
ZA1 厂 /22 kV.4 号机	燃煤	LXC	JH	390.66	390.66	401.34

3.4 数据可视化

对数据集成表预处理之后，可进行数据可视化。以电厂出力数据可视化为例，一个电厂可能有多个机组，对经过预处理的机组-电厂-地区数据表，把一个电厂下所有的发电机组出力数据对应相加得到该电厂的出力。图3是电厂BLC 2018年的出力时间序列图。通过图像可以很直观地看出该电厂出力一年的变化情况，具有一定的季节变化规律，春节期间电厂出力降低，夏季电厂出力升高，国庆节期间电厂出力降低，冬季电厂出力缓慢升高。

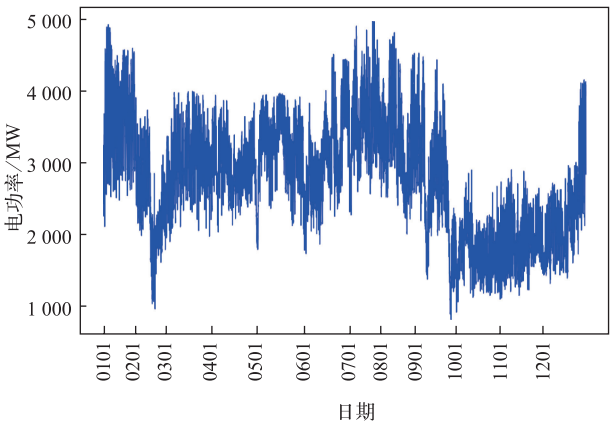

图 3 电厂 BLC 出力时间序列图

Fig. 3 Time series diagram of BLC power plant output

3.5 电厂出力预测

以电厂BLC为例，从经过预处理后得到的该电厂数据中选取2018年每天上午10的数据，得到365个时间序列数据，选取前80%的数据作为训练数据，后20%的数据作为测试数据。模型经过训练之后，用测试数据进行检测，检测结果如图4、图5所示，BP神经网络的平均绝对百分比误差（MAPE）为11.61%，LSTM神经网络的MAPE为10.90%。可以看出，用LSTM神经网络对电厂出力预测优于BP神经网络。

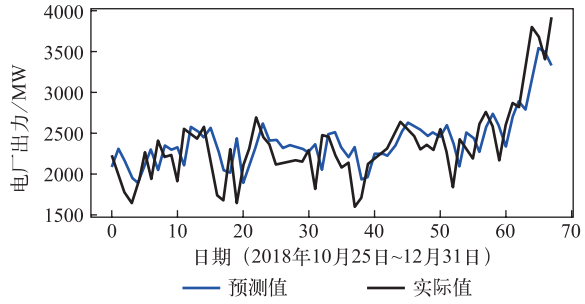

图 4 BP 神经网络预测结果

Fig. 4 Prediction results of BP neural network

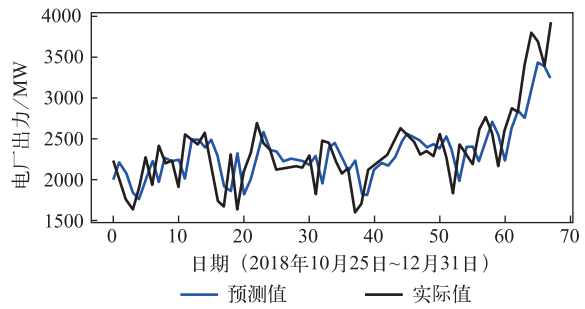

图 5 LSTM 神经网络预测结果

Fig. 5 Prediction results of LSTM neural network

4 结束语

电厂机组出力数据是电力行业重要的数据资源，需要有效的数据分析技术才能挖掘其中的价值。本文提出了电厂机组出力数据分析的一整套流程和具体的算法，包括数据集成、异常值预处理、数据可视化和电厂出力预测。以某省发电机组出力数据进行实证分析，通过数据集成和异常值处理，对电厂机组出力一年的数据可视化可看出变化趋势，通过LSTM神经网络模型对电厂出力进行预测，LSTM预测得到的MAPE为10.90%，预测结果优于BP神经网络，BP神经网络预测得到的MAPE为11.61 %。本文提出的数据集成和机组出力预测方法对国家电网机组出力大数据分析有一定的参考应用价值。EA

参考文献

[1] 王淞，彭煜玮，兰海，等．数据集成方法发展与展望[J]．软件学报，2020，31(3)：893-908．

[2] RAHM E，BERNSTEIN P A．A survey of approaches to automatic schema matching[J]．The VLDB Journal，2001，10(4)：334-350．

[3] YING Q，LI Y，SONG J，et a1．Discovering complex semantic matches between database schemas[J]．Computer Engineering & Science，2008，29(10)：61-63．

[4] 韩朋，张晓琳，张飞，等．基于AM-LSTM 模型的超短期风电功率预测[J]．科学技术与工程，2020，20(21)：8594-8600．

[5] 王炜，刘宏伟，陈永杰，等．基于LSTM 循环神经网络的风力发电预测[J]．可再生能源，2020，38(9)：1187-1191．

[6] 管军霖，智鑫．基于LSTM 的超短期光伏发电功率预测[J]．通信电源技术，2020，37(8)：123-125．

[7] 安鹏跃，孙堃．基于相似日和回声状态网络的光伏发电功率预测[J]．智慧电力，2020，48(8)：38-43．

[8] 黄伟建，张一帆，黄远．动态三次指数平滑法的火电厂发电量预测研究[J]．现代电子技术，2020，43(17)：147-154．

作者简介

涂 炼

涂炼（1989—），男，硕士，工程师。主要从事电网运行策划工作。

(收稿日期：2021-03-03)