

基于 CNN 和 LSTM 联合预测并修正的电量缺失数据预测^①



郭蕴颖^{1,2}, 丁云峰²

¹(中国科学院大学, 北京 100049)

²(中国科学院 沈阳计算技术研究所, 沈阳 110168)

通讯作者: 郭蕴颖, E-mail: gyy1.4@163.com

摘要: 数据是电网调度控制系统稳定运行的关键依据, 而因为硬件故障等原因导致数据采集过程中的数据缺失会影响到系统数据的完整性, 从而对电网调度的智能性和高效性产生相应的影响. 因此, 针对缺失数据的准确预测对于智能电网调度系统的建设有着重要的意义. 本文针对解决电网领域电能量采集系统的缺失数据预测问题对已有的基于 CNN 和 LSTM 联合预测方法进行改进和优化, 在联合预测模型基础上添加修正模型, 针对不同缺失数据段利用 CNN 卷积神经网络和电力数据里特有的对侧数据场景建模, 实验结果证明该方法将平均绝对误差值降到 0.142, 提高了现有预测模型的准确率, 对电网调度系统的智能性和高效性提供了数据完整性、准确性的保障.

关键词: 电量缺失数据预测; CNN; LSTM; 对侧数据

引用格式: 郭蕴颖, 丁云峰. 基于 CNN 和 LSTM 联合预测并修正的电量缺失数据预测. 计算机系统应用, 2020, 29(8): 192-198. <http://www.c-s-a.org.cn/1003-3254/7580.html>

Prediction and Correction of Power Loss Data Based on CNN and LSTM

GUO Yun-Ying^{1,2}, DING Yun-Feng²

¹(University of Chinese Academy of Sciences, Beijing 100049, China)

²(Shenyang Institute of Computing Technology, Chinese Academy of Sciences, Shenyang 110168, China)

Abstract: Data is the key basis for the stable operation of the power grid dispatching control system, in the process of data collection, the lack of data due to hardware failure and other reasons will affect the integrity of the system data, which will have a corresponding impact on the intelligence and efficiency of power grid dispatching. Therefore, the accurate prediction of missing data is of great significance for the construction of smart grid dispatching system. In order to solve the problem of missing data prediction of electric energy collection system in the field of power grid, this study improves and optimizes the existing joint prediction method based on CNN and LSTM, adds a modified model on the basis of the joint prediction model, and uses CNN convolution neural network and the unique opposite side data scene modeling in the electric power data for different missing data segments. The experimental results show that this method reduces the average absolute error value to 0.142, which improves the accuracy of the existing prediction model and accuracy guarantee for the intelligence and efficiency of power grid dispatching system.

Key words: prediction of power loss data; CNN; LSTM; contralateral data

1 引言

随着智能化系统以及各种先进智能技术在各领域

的流行, 电网领域的智能电网调度控制系统建设也应应运而生, 从而提高电量的使用率和调度工作效率. 在智

① 收稿时间: 2020-02-08; 修改时间: 2020-03-03; 采用时间: 2020-03-20; csa 在线出版时间: 2020-07-29

能电网调度控制系统中,电量数据是传输数据中的重要组成,而电量数据的获取是通过不同地区的观测点设备的读数记录的,为了提高系统跨区域电量数据传输效率,观测点设备每隔5分钟、10分钟或15分钟读取一次数据,每天记录的数据组成一个传输数据单元并打包传递给下一个记录读取观测点.在数据记录传输过程中,可能会因为设备故障等原因导致读取数据的缺失或错误,从而影响传输数据单元的数据完整性,进一步影响智能电网调度控制系统的智能性,因此,需要提出一种对于电量缺失数据精准预测的模型方法为智能电网调度控制系统提供支撑.

关于电量缺失数据预测问题是一个时序预测问题,随着机器学习神经网络等算法的发展和日趋成熟,研究用于解决时序预测问题的时序预测模型也越来越多,比如:传统的平滑预测法、趋势预测法、自回归模型(AR)、移动平均模型(MA)、以及自回归移动平均模型(ARMA)和差分整合移动平均自回归模型(ARIMA)等.平滑法预测一般有移动平滑法和指数平滑法,移动平滑法是利用前 t 时刻的前 p 个时刻的平均值来预测 t 时刻的数值,而指数平滑法是在移动平滑法基础上发展起来的一种时间序列分析预测法,这种模型预测方法只需较少的实验数据,就可以预测出来所需要的结果,但是该模型算法在赋权重的时候,对远期数据赋较小的权重,近期数据赋较大的权重,因此,只能进行短期预测,预测范围较为局限;在现实工作预测中,为了提高工作效率选择趋势预测法,既将缺失数据两端直接连接的方法进行数据缺失段预测,该方法在实际项目工作中最大的优点就是方便,直接根据趋势线性预测就能得到预测值,但缺点显而易见,预测数据误差也是很高的,并不能完成精准预测;AR模型以时间序列的前一个值和当前残差来线性地表示时间序列的当前值,而MA模型则用时间序列的当前值和先前的残差序列来线性地表示时间序列的当前值,ARMA模型是前面两个模型的整合,这3种模型都适用于平稳时间序列的预测,而对于非平稳时间序列,可以通过差分过程将非平稳数据转换为平稳数据后用ARIMA模型可以解决.文献[1]通过实验比较了ARIMA模型与指数平滑法预测门诊量效果比较,从而得出ARIMA模型的预测效果更好的结论.但是,ARIMA模型对于趋势性较强的数据集预测效果更好,而且本质上只能捕捉线性关系,而对于电网数据来讲并不能完全的线性拟

合,因为用电量会有高峰期,亦或会在某些特殊地段的用电量会因为特殊事件发生突变,例如当附近的高铁经过时引起的电量变化情况.对于电力方面的预测,文献[2]提出了一种基于趋势变化分段的电力负荷组合预测方法,该算法是利用了电力负荷“三峰三谷”变化特性做出的研究,但本文要解决的问题是一天内用电量数据的缺失值预测,显然一天内的用电量数据值一定是递增的,不会出现“谷”,因此并不适合.对于时序序列的非线性预测方法,LSTM算法是较成熟切主流的时序预测方法.文献[3]提出基于LSTM算法的电力谐波监测数据预测分析的方法,通过对不同时间尺度谐波监测数据预测分析验证了方法的有效性和实用性,文献[4]提出基于LSTM模型的日销售额预测方法,但电量数据会因为某一时刻的突发事件发生电量数据的突增,如果仅仅通过LSTM模型根据历史数据进行预测误差一定会大,因此采用电网中的对侧数据这一特征数据进行校验,以此缩小预测值的预测误差.

基于对上述方法的总结并综合电网数据易突变的特性,本文对现有的LSTM预测方法联合CNN预测应用于电网领域,并在该联合预测模型的基础上添加对侧数据修正模块进行模型的进一步优化.通过CNN卷积神经网络将特征进行融合后重新提取新特征,通过电网中的对侧数据作为修正,从而降低仅通过LSTM预测的误差.根据缺失数据在传输数据段中的位置,本文将预测模型分成3组,分别是缺失数据段位于传输数据段的前段、中段和后段,将从某电网获取的实验数据分成3组,分别用于训练3种不同情况下的预测模型.将每组数据先通过数据预处理模块,将经过预处理的数据通过卷积神经网络进行模型新特征的提取,提高预测的准确率,并将新特征作为LSTM神经网络的输入进行LSTM模型的学习与训练,并反向计算每个神经元的误差项值,根据相应的误差项,计算每个权重的梯度,完成CNN和LSTM联合预测模型的训练和学习.将模型学习后输出的预测值输入到数据修正模块,该模块采用电网数据中特有的对侧数据属性进行数据修正,加入电网数据中特有的对侧数据这一特性均衡,减小预测值与实际值的误差大小,使预测模型的准确率更高.实验对比5种不同预测模型:趋势预测模型、ARIMA预测模型、LSTM预测模型、CNN和LSTM联合预测模型和CNN和LSTM联合预测并修正模型,通过对比5种预测模型的预测准确率,验证了

CNN 和 LSTM 联合预测的可行性的同时也可以看出加入修正模块后模型预测误差的明显降低。

2 相关模型理论

(1) CNN 卷积神经网络

卷积神经网络 (Convolutional Neural Networks, CNN) 是一种包含卷积操作的前馈神经网络, 基本结构由输入层、卷积层、池化层、全连接层和输出层构成。卷积层和池化层一般会取若干个交替设置组合使用, 卷积层中输出特征图的每个神经元与其输入进行局部连接, 并通过对应的连接权值与局部输入进行加权求和再加上偏置完成特征提取。

考虑到要建立模型的目的是提高预测的准确率、减小预测值和实际值的误差值, 直接将经过预处理的数据作为输入数据训练 LSTM 网络, 并不能将输入数据的特征更好的融合, 只能单纯的预测出输入数据原本特性下通过 LSTM 神经网络学习出的预测值, 准确率和精准度不够高。而 CNN 卷积神经网络的原理在于通过对输入数据的卷积和池化操作, 将数据的隐藏结构特征进行提取, 并随着网络模型复杂度的提高, 特征提取的维度越来越高, 提取到的特征越来越抽象, 最后将抽象特征融合在一起, 得到提取的新特征。因此模型先将处理的数据通过 CNN 卷积神经网络进行新特征的提取, 提取的新特征更全面的融合了输入模块中输入数据的特征, 并作为输入对 LSTM 模型进行学习和训练, 能使得预测模型学习、训练的更精准, 误差更小。

(2) LSTM 长短期记忆网络

长短期记忆网络 (Long Short-Term Memory, LSTM) 是一种时间循环神经网络, 是循环神经网络 (Recurrent Neural Network, RNN) 的优化, 循环神经网络主要应用之一是时间序列的分析和预测, 而在电量缺失数据预测中, 当前预测值是由与之相邻的电量值根据变化趋势走向预测得到, 因此选择用循环神经网络建模预测。解决了 RNN 循环神经网络对远距离、长周期数据遗忘问题的 LSTM 神经网络在保持其原有模型结构的基础上, 设计隐藏层结构提高了对长序列的分析能力。LSTM 神经网络的关键在于细胞状态 (cell state), 细胞状态类似于输送带, 细胞的状态在整个链上运行, 通过设置门结构完成线性操作, 从而实现信息的删除和添加, 也避免了 RNN 神经网络训练过程中容易出现的梯度消失和梯度膨胀问题, 实现更精准的预

测学习。

LSTM 的实现由 3 个门: 遗忘门, 输入门, 输出门组成, 每个门负责是事情各不相同, 遗忘门负责决定保留多少上一时刻的单元状态到当前时刻的单元状态; 输入门负责决定保留多少当前时刻的输入到当前时刻的单元状态; 输出门负责决定当前时刻的单元状态有多少输出^[5]。每个 LSTM 包含了 3 个输入, 即上时刻的单元状态、上时刻 LSTM 的输出和当前时刻输入。

LSTM 模型第 1 步是通过遗忘门从细胞状态中丢弃无用的信息, 将前一时刻隐藏层的输出 h_{t-1} 和当前状态的输入 x_t 连接后通过遗忘门的权重矩阵 W_f 赋权重, 通过 Sigmoid 激活函数 σ 的计算决定有多少信息可以通过。公式如下:

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

第 2 步是通过输入门添加当前时刻的输入信息到信息流中, 和遗忘门的操作类似, 将前一时刻隐藏层的输出 h_{t-1} 和当前状态的输入 x_t 连接后通过输入门的权重矩阵 W_i 赋权重, 通过 Sigmoid 激活函数 σ 计算出通过输入门后的信息值。公式如下:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (2)$$

第 3 步是对信息流中细胞状态的更新, 首先定义一个当前输入的细胞状态 C_t , 该状态通过前一时刻隐藏层的输出 h_{t-1} 和当前时刻的输入 x_t 计算得到。公式如下:

$$\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (3)$$

将当前时刻遗忘门的计算值 f_t 与前一时刻细胞状态做叉乘, 当前时刻输入门的结果与当前时刻的细胞状态做叉乘, 并将两部分叉乘值相加对信息流中的细胞状态进行更新。公式如下:

$$C_t = f_t \times C_{t-1} + i_t \times \tilde{C}_t \quad (4)$$

第 4 步是通过一个 Sigmoid 激活函数, 决定要输出的细胞状态的部分信息作为输出门的数据结果。公式如下:

$$O_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (5)$$

第 5 步是将当前时刻的细胞状态通过 \tanh 将数值规范化, 并将该数值与输出门的结果 O_t 做叉乘作为当前细胞状态隐藏层的输出值 h_t 传递到下一时刻, 为下一时刻的细胞状态更新做准备。公式如下:

$$h_t = O_t \times \tanh(C_t) \quad (6)$$

至此, LSTM 模型的单次流程就结束了。

(3) 数据修正

数据通过预测模型的预测得到预测值, 因为 LSTM 模型的预测是通过历史数据的趋势走向为预测参照, 而在电力领域, 电量的变化并非是一直平稳, 每天内都会有用电高峰期, 在特殊地段的特殊时刻也会发生电量突变的情况, 比如在居民区的晚间用电量较白天用电量来说就是用电高峰; 再比如在某一时刻, 高铁火车的通过、天气炎热时空调的使用、工厂新设备的投运等, 都会使得该时刻的电量突增。此时如果单纯使用 CNN 和 LSTM 模型根据历史数据进行预测, 也会产生较大的预测误差, 为解决这个电力领域存在的特殊情况, 本文提出通过对侧数据进行数据修正的方案。

在电力领域, 电力的传输都是双向的, 每一个观测站点既接收前一个观测站点传过来的电力数据信息, 也会向下一个观测站点传递当前站点接收到的电力数据信息。该实验建模预测的数据是该工作站点接收到的电量数据, 对于该数据而言, 其对侧数据是前一个观测站点的发送出来的电量数据值。虽然前一个观测站点的电量数据到该工作站点的传输过程中会因为线损等不可避免的原因导致数据的损失, 但是数据变化的

走向和趋势是一致的, 可以准确的体现出该时刻数据的变化程度, 在电量因为某些特殊情况发生电量突变的时候, 通过该观测点对侧数据的走向和趋势可以对 CNN 和 LSTM 预测模型预测出来的数据进行数据的修正, 提高其预测的准确性。

该实验研究中, 对于预测数据的修正采取均值修正法, 将通过 CNN 和 LSTM 模型预测出来的当前时刻的缺失数据值 P_t 和该时刻该观测点电量的对侧数据值 OP_t 加和取平均作为最终的预测数据。

3 模型结构和实验

(1) 模型结构

根据上述模型理论, 建立 CNN 和 LSTM 联合预测和修正模型。总体架构图如图 1 所示。根据缺失数据在传输数据段中的位置, 将预测模型分为 3 种, 如图 2~图 4 所示: 缺失数据段在传输数据段前段时选择前向模型预测, 既使用后一时刻的数据反向预测; 缺失数据段在传输数据段中段时选择双向模型预测, 既前后双向预测两个预测值并取均值作为最终的预测值; 缺失数据段在传输数据段后段时选择后向模型预测, 既使用前一时刻的数据正向预测。

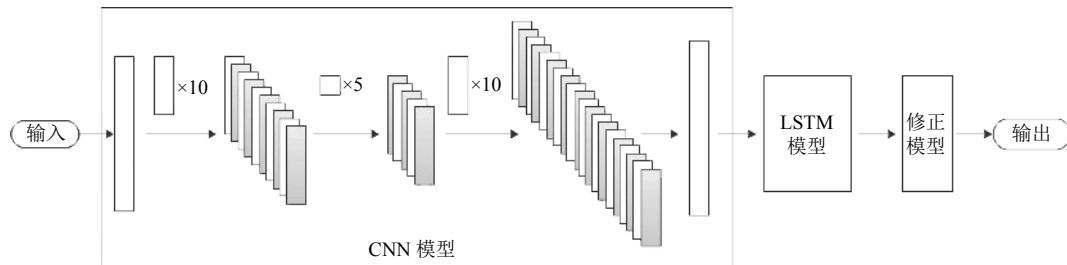


图 1 模型总体架构图

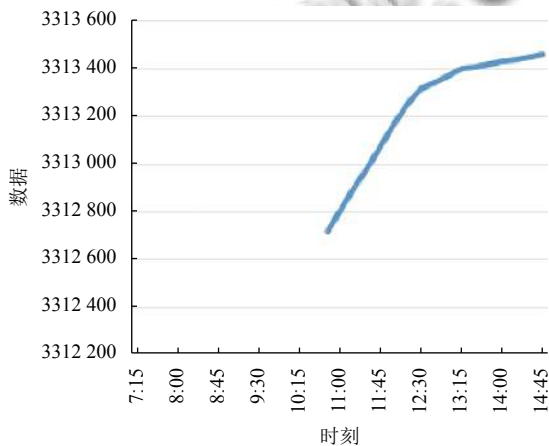


图 2 前段数据缺失图

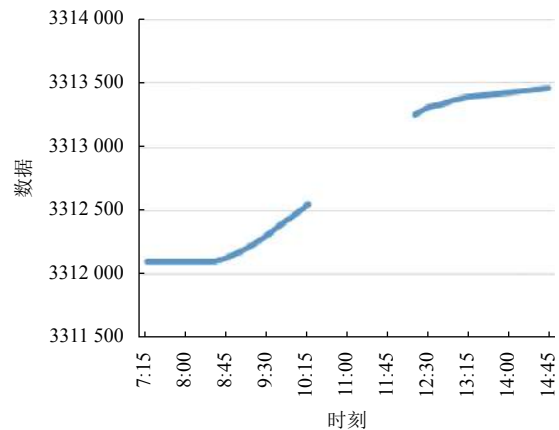


图 3 中段数据缺失图

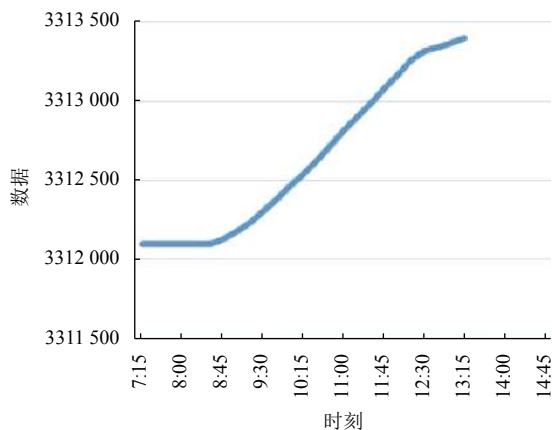


图4 后段缺失数据图

① 数据预处理

该实验数据存储使用的数据库为国产的达梦数据库,数据为某电网采集到的某段时间内的电量数据,数据预处理阶段主要完成电量数据的提取,将电表信息表和电量表通过电表ID进行表连接,通过电表信息表中的正向有功和反向有功对应的电表ID获取到电表中的电量数据值并分别提取,将正向电量数据作为实验数据进行模型的学习和训练,反向电量数据作为修正模块数据。

② CNN 卷积神经网络

该实验的卷积神经网络设计有1个输入层、3个卷积层和1个全连接层构成,每一次训练选取16个电量数据为一组,组成一个 16×1 的向量矩阵作为单次模型训练的输入,并以一个步长为单位选取30次。每一次训练分别用10个 3×1 的卷积核进行第1次的卷积操作,并使用ReLU激活函数激活避免训练过程中的梯度爆炸和梯度消失的问题,得到10个 14×1 的向量矩阵。用5个 1×1 的卷积核进行第2次的卷积操作,主要达到降维的目的,减小模型的复杂度,得到5个 14×1 的向量矩阵。用20个 3×1 的卷积核进行第3次的卷积操作,得到20个 12×1 的向量矩阵。将20个 12×1 的向量矩阵输入到全连接层进行全连接操作,得到最终 30×1 的特征向量值。

③ LSTM 神经网络

该实验的LSTM神经网络的训练首先是通过数据获取模块,将CNN卷积神经网络训练学习得到的 30×1 的特征向量获取作为一次模型训练学习的数据样本,通过得到的30个新的特征数据值训练学习,模型中的隐藏层设计为10个神经元,用来记忆和存储历史

状态数据,输出层为1个神经元,得到预测值,根据预测值和实际值完成一次误差反向传播和参数更新。

④ 数据修正

该实验的修正模块采用的是均值修正法,将该缺失数据预测值与该观测点对侧数据值加和取均值作为最终预测值,通过对侧数据值将突变的电量数据预测中和。

(2) 实验数据

实验数据来自某电网某段时间电量数据,电量采集是半自动化采集,每隔15分钟采集一次观测点仪器设备值作为该观测点的电量数据,每一天为一个传输单元进行数据传输,因为实验数据量过大,文章篇幅有限,因此选取某一个观测点一天内部分电表数据来展示实验数据。

表1中为部分电量实验数据。

表1 部分电量实验数据

设备编号	观测时间	电表数值
30001766	00:00	3312 100
30001766	01:00	3312 100
30001766	08:00	3312 100
30001766	08:30	3312 110
30001766	08:45	3312 130
30001766	09:00	3312 180
30001766	09:15	3312 240
30001766	09:30	3312 310
30001766	09:45	3312 390
30001766	10:00	3312 470
30001766	10:15	3312 550
30001766	10:30	3312 630
30001766	10:45	3312 720
30001766	11:00	3312 810

实验数据分成3部分,用来学习和训练缺失数据段位置在传输数据段位置不同情况下的3种模型,并在每个模型训练过程中选取该部分所有电量数据的80%的数据作为实验数据,剩余20%作为验证数据。

(3) 实验模型对比

实验采取了现阶段较为流行的时序预测方法进行缺失数据预测实验,为了更直观的体现不同预测模型的性能好坏以及数据预测的准确性,本文将模型预测的平均绝对误差值作为评价指标,通过对比不同模型预测的平均绝对误差值的大小得出模型性能对比结果。

平均绝对误差值是所有单个观测值与算术平均值的偏差的绝对值的平均。平均绝对误差可以避免误差相互抵消的问题,因而可以通过模型的平均绝对误差

来准确反映模型实际预测误差的大小。

$$\text{平均绝对误差值} = \sum |\text{预测值} - \text{真实值}| \div \text{样本总数}$$

实验结果如图5、图6及表2所示。

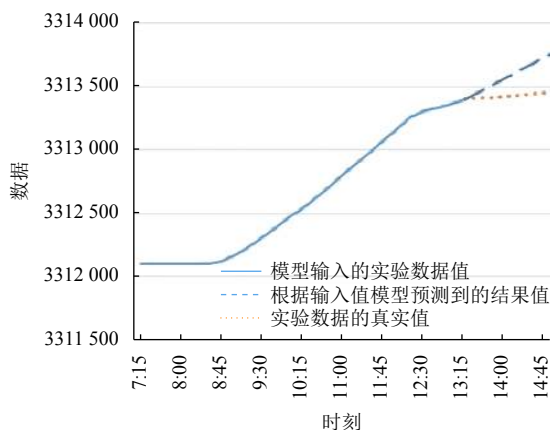


图5 趋势预测后段缺失数据图

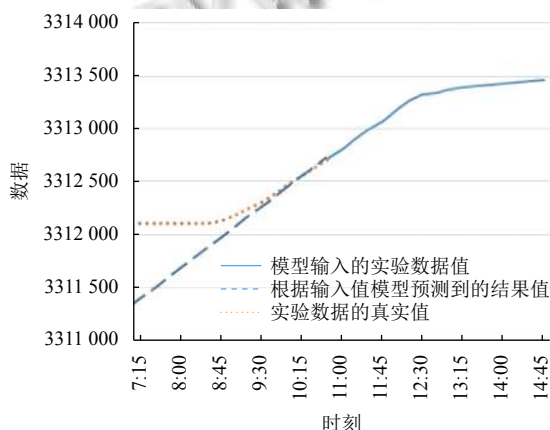


图6 趋势预测前段缺失数据图

表2 模型平均绝对误差值结果对比

模型	缺失前段数据	缺失中段数据	缺失后段数据	误差均值
LSTM	38.7690	59.6572	28.9834	42.4699
ARIMA	107.9023	24.2704	87.0923	73.0883
CNN+LSTM	40.7601	43.7867	11.9055	32.1508
CNN+LSTM+修正	10.9760	21.7603	9.8634	14.1999
数据趋势预测	290.7141	543.4963	368.5224	400.9109

方法一采用趋势预测法, 缺失数据段在传输数据段的前段和后段时, 按已知数据的增长趋势进行直接补齐, 完成预测; 缺失数据段在传输数据段的中段时, 直接将缺失数据的两端连接完成数据预测. 该预测方式优点是预测方便, 但准确率是很低的, 有很大的偶然性, 因此并不能达到精准预测的预期效果。

方法二单纯采用 LSTM 进行预测, 只考虑历史数据的变化趋势, 该方法的实验结果可以证明 LSTM 时序预测的可行性; 方法三采用了 ARIMA 模型预测, 因为电量数据值并不满足 ARIMA 模型对数据稳定性的要求, 因此首先将数据通过差分达到数据稳定的要求再进行缺失数据预测, 该方法的模型简单, 预测的误差相比方法一有明显的降低; 方法四和方法五都主要采用 CNN 和 LSTM 联合预测的模型, 区别在于方法五根据电网数据的特殊性添加了用电量数据的对侧数据对预测数据的修正模块, 用均值修正法将对侧数据和预测数据取平均作为最终预测结果从而降低预测值与真实值直接的误差大小. 通过对比方法二和方法三的实验结果可以得出在解决电网缺失数据预测问题中, LSTM 模型的预测误差要小于 ARIMA 模型的预测误差, 从而得出 LSTM 更适合解决电量缺失数据预测问题. 方法二、方法四和方法五这 3 种预测模型的实验结果可以得出本文提出的 CNN 卷积神经网络进行数据特征提取、对侧数据进行预测数据修正的必要性。

(4) 实验结果分析

通过实验得到的平均绝对误差值的结果值, 可以看出 3 种不同情况下的误差均值最小的是 CNN 和 LSTM 联合预测和修正模型, 可以看出本文提出的电量缺失数据预测方法的可行性、合理性和准确性。

数据趋势预测模型的高误差可以看出该方法的预测准确率较低, 相比于该方法, LSTM 模型预测和 ARIMA 模型预测的误差均值都有了较明显的降低, 实验数据对比可以看出, 虽然 ARIMA 模型的中段数据预测误差小于 LSTM 模型预测误差, 但是从误差均值整体看来 LSTM 模型优于 ARIMA 模型; 在 LSTM 预测模型基础上添加利用 CNN 卷积神经网络进行特征提取的模块后, 虽然前段数据的预测并无优化, 但是中段、后段和误差均值都有了一定程度的优化; 而本文根据电力数据特点加入对策数据修正模块的模型在各个阶段都有了优化, 不同缺失数据段的误差较其他预测模型都有了一定程度的降低, 误差均值也达到了实验模型中的最优值. 以此可以得出, 本文提出的 CNN 和 LSTM 联合预测并修正的模型在预测误差达到了最小值, 对于预测问题准确性的要求完成度最高。

4 结语

本文以解决电网传输过程中电量数据缺失问题为

背景,在 CNN 和 LSTM 联合预测的模型基础上增加了根据电网数据特有的对侧数据作为基础参考数据的数据修正模块,对模型预测的数据做一定的误差校正,通过实验结果也验证了该方法的可行性以及修正模块的必要性.该方法的研究可以提高现有预测模型的模型准确率,其对于电力调度系统的应用有着举足轻重的作用,确保了传输数据段数据的完整性,也对电网企业里进行电力调度业务运营的智能性、准确性提供了数据支持.

参考文献

- 1 王晨,郭倩,周罗晶. ARIMA 模型与指数平滑法预测门诊量效果比较. 预防医学, 2018, 30(11): 1152-1155.
- 2 谭风雷,张军,马宏忠. 基于趋势变化分段的电力负荷组合预测方法. 华北电力大学学报(自然科学版), 2020, 47(2): 17-24.
- 3 刘启斌,尹温硕,胡卫华,等. 基于 LSTM 算法的电力谐波监测数据预测. 电力电容器与无功补偿, 2019, 40(5): 139-145.
- 4 吴娟娟,任帅,张卫钢,等. 一种基于 LSTM 模型的日销售额预测方法. 计算机技术与发展, 2020, 30(2): 133-137. [doi: 10.3969/j.issn.1673-629X.2020.02.026]
- 5 Lin T, Guo T, Aberer K. Hybrid neural networks for learning the trend in time series. Proceedings of the 26 International Joint Conference on Artificial Intelligence. Melbourne, Australia. 2017. 2273-2279.