

三种数学模型预测医院门急诊就诊人次数的价值比较

窦一峰, 崔金广, 蒙文涛, 吴秀春

(天津市宝坻区人民医院网络信息中心, 天津 301800)

摘要:目的 探索数学模型对医院门急诊就诊人次的预测效果,以期得到最佳的预测模型,更好的提升门急诊资源配置效能。**方法** 收集天津市某三级甲等综合性医院 2009~2019 年医院门急诊就诊数据,构建 NARNN、LSTM 和 SARIMA 模型,采用 2009 年 1 月~2018 年 12 月的数据用于模型的训练和测试,以 2019 年 1~12 月数据为预测集,比较三种模型的预测效果。**结果** NARNN、LSTM 和 SARIMA 模型在 MAPE 上的结果分别为 8.22%、4.32%和 3.55%,在 SMAPE 上的结果分别为 8.37%、4.33%和 3.58%,LSTM 和 SARIMA 的预测效果优于 NARNN,其中 SARIMA 模型在 4 个指标上的结果均优于 LSTM 和 NARNN,对门急诊人次拟合和预测效果较好。**结论** 三种数学模型均能对门急诊人次数据进行有效预测,其中 SARIMA 模型预测该院门急诊就诊人次的效果最优,可为决策提供预测数据支持。

关键词: 医院;门急诊;就诊人次;预测;NARNN;LSTM;SARIMA

中图分类号:R19

文献标识码:A

DOI: 10.3969/j.issn.1006-1959.2020.24.006

文章编号:1006-1959(2020)24-0018-05

Comparison of the Value of Three Mathematical Models on Prediction in the Number of Outpatient and Emergency Patients

DOU Yi-feng, CUI Jin-guang, MENG Wen-tao, WU Xiu-chun

(Network Information Center, People's Hospital of Baodi District, Tianjin 301800, China)

Abstract: Objective To explore the predictive effect of mathematical models on hospital outpatient and emergency visits, in order to obtain the best predictive model, and better improve the efficiency of outpatient and emergency resource allocation. Methods Collecting the outpatient and emergency department visit data of a tertiary A general hospital in Tianjin from 2009 to 2019, construct NARNN, LSTM and SARIMA models, and use the data from January 2009 to December 2018 for model training and testing. The data from January to December 2019 is a prediction set, and the prediction effects of the three models are compared. Results The results of the NARNN, LSTM and SARIMA models on MAPE were 8.22%, 4.32% and 3.55%, respectively, and the results on SMAPE were 8.37%, 4.33% and 3.58%, respectively. The prediction effects of LSTM and SARIMA were better than NARNN, of which SARIMA results of the model on the four indicators were better than LSTM and NARNN, and the effect of fitting and predicting the number of outpatient and emergency department was better. Conclusion The three mathematical models can effectively predict the number of outpatient and emergency department visits. The SARIMA model predicts the number of outpatient and emergency department visits in this hospital with the best effect, and can provide predictive data support for decision-making.

Key words: Hospital; Outpatient and emergency department; Visits; Prediction; NARNN; LSTM; SARIMA

门急诊是医院向患者提供服务的第一个窗口,门急诊患者流量的变化直接影响医院医疗服务工作的组织与管理。一方面,根据门急诊人次数据合理配置门诊医生等医疗资源,节省患者排队等候时间,从而提高患者满意度,增强患者获得感。另一方面也能够及时有效地为医院领导决策和制定长短期规划提供科学依据,因此如何正确反映医院门急诊人次数的动态变化趋势和规律,对医院长远发展来讲具有十分重要的意义。

1 资料与方法

1.1 资料来源 数据采集于天津市某三级甲等综合性医院 2009~2019 年医院门急诊就诊人次的月度数据,来源于《门急诊工作量月报表》,数据真实可靠。本文采用 2009 年 1 月~2018 年 12 月用于模型的训练和测试,2019 年 1~12 月的数据作为预测集检验模型的预测效果。

1.2 方法

1.2.1 NARNN 模型的建立

神经网络是一种模拟作者简介:窦一峰(1992.8-),男,天津人,硕士,助理工程师,主要从事医学统计与数据挖掘工作

通讯作者:蒙文涛(1985.5-),男,天津人,硕士,工程师,主要从事医疗信息化建设工作

大脑神经系统处理信息的方式而人为建立的能够实现某种功能的网络,由大量模拟生物系统中神经元之间突触连接的神经元形成,因此其相比其他数学模型具有非线性,鲁棒性,并行性和自适应性等特点^[1-3]。本文采用非线性自回归神经网络模型(nonlinear autoregression neural network, NARNN)^[4,5]将自身作为回归变量进行建模,利用前期多干时刻的随机变量的线性组合来描述后面时刻数值,其形式如下:

$$y_t = b_0 + b_1 y_{t-1} + b_2 y_{t-2} + b_3 y_{t-3} + \cdots + b_n y_{t-n} + \varepsilon_t$$

典型的 NARNN 包括滞后阶数,输入层,隐含层和输出层四个部分,基于自身数据作回归,形如 $y_t = f(y_{t-1}, y_{t-2}, y_{t-3}, \cdots, y_{t-lag})$ 。本文将数据分为用于训练调整网络的权值和阈值的训练集(80%),用于将训练过程中的过拟合现象降到最低的验证集(10%)和对网络结构进行最终调试的测试集(10%),以期得到最好的泛化能力。考虑到门急诊人次具有一定的季节性的特点,将滞后阶数初始化设定为 12,采取 Levenberg-Marquardt 函数进行网络训练,通过经验初始化隐含层神经元个数值位 3~200,进而不断调整尝试,根据拟合优度检验值 R 方最优和误差最小化原则综合确定最终隐含层神经元个数。训练完成后

采用前进递推预测法对未来一年数据进行预测,即根据拟合数据集预测(t+1)月医院门急诊人次,再将(t+1)月门急诊人次数据代入预测模型计算(t+2)院门急诊人次,依次计算得到最终结果,该模型使用 MATLAB R2016b 编程求解。

1.2.2 LSTM 模型的建立 由 Hochreiter & Schmidhuber 提出的长短时记忆网络(long short term memory network, LSTM)是一种在实际应用中能够学习长期依赖关系的循环神经网络^[6,7]。它改良了标准 RNN 中的仅有一种如 tanh 的重复神经网络模块的链式结构,LSTM 核心是 cell 状态,LSTM 网络能通过一种被称为门的结构对 cell 状态进行有选择性的决定让哪些信息通过,LSTM 具有 3 个由 sigmoid 层和点乘操作的组合的门结构,分别称作遗忘门,输入门和输出门,其中三类门共同控制信息进入和离开记忆细胞,输入门调节进入记忆细胞的新信息;遗忘门控制记忆细胞中保存多少信息;输出门定义可以输出多少信息。基于此,本文选取 LSTM 模型对医院门急诊人次进行预测,通过学习历史数据中存在的时间依赖关系来进行有效预测,该模型使用 PYTHON 编程求解。

1.2.3 SARIMA 模型的建立 AR/MA/ARMA 模型是分析时间序列的重要方法^[8]。某些时间序列,如医院门急诊就诊人次等存在明显的周期性变化,这种周期是由于季节性变化(季度、月度等)引起的,把这种基于季节性的时间序列预测方法叫做季节时间序列模型(seasonal arima model, SARIMA),也叫乘积 ARIMA 模型^[9-11],其标准格式为: SARIMAA(p,d,q)×(P,D,Q)_s。其中 p、d、q 分别表示传统 ARIMA 模型的自回归阶数、差分阶数和移动平均阶数, P、D、Q 分别表示季节性 ARIMA 模型的自回归阶数、差分阶数和移动平均阶数,s 值为 12。对于月度数据,其季节性变化周期为 12,将首先对数据进行平稳性检验,根据序列散点图、自相关函数(auto-correlation function, ACF)图、偏自相关函数(partial auto-correlation function, PACF)图等初步确定模型参数,并通过残差检验判断拟合模型是否有效,最后考察赤池信息准则(akaike information criterion, AIC)和贝叶斯信息准则(bayesian information criterion, BIC),依据信息准则最小化原则选取多个模型中结果相对最优的作为最终的预测模型,该模型使用 MATLAB R2016b 编程求解。

1.3 模型评价 主要采用均方误差(mean squared error, MSE)、均方根误差(root mean square error, RMSE)和平均绝对误差(mean absolute error, MAE),平均绝对百分比误差(mean absolute percentage error, MAPE),对称平均绝对百分比误差(sym-

metric mean absolute percentage error, SMAPE)来评价三个模型预测效果的优劣。所有评价指标的值越小,说明预测模型描述实验数据的准确度越高,模型越优。令预测值为 $y_{pred}=\{y_{pred1}, y_{pred2}, \dots, y_{predn}\}$, 真实值为 $y_{raw}=\{y_{raw1}, y_{raw}, \dots, y_{rawn}\}$ 具体计算公式依次如下:

$$MSN = \frac{1}{n} \sum_{i=1}^n (y_{predi} - y_{rawi})^2 \quad (1)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_{predi} - y_{rawi})^2} \quad (2)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_{predi} - y_{rawi}| \quad (3)$$

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_{predi} - y_{rawi}}{y_{rawi}} \right| \times 100\% \quad (4)$$

$$SMAPE = \frac{1}{n} \sum_{i=1}^n \frac{|y_{predi} - y_{rawi}|}{(|y_{predi}| + |y_{rawi}|)/2} \quad (5)$$

2 结果

2.1 NARNN 模型 根据模型结果可知,模型误差在滞后阶数为 0 时最大,其余情况均在置信区间范围内,见图 1。当隐含层神经元个数 100 时,根据图 2 可知,模型在训练集上的 R 方为 1,在验证集上的 R 方是 0.93,在测试集上的 R 方为 0.95,模型总体的拟合优度值为 0.98,该模型十分理想,可用性强,其实际值与拟合值的误差对比见图 3。

2.2 LSTM 模型 本次构建的 LSTM 网络有 1 个输入层,1 个带有 15 个 LSTM 神经元的隐含层,选择 tanh 作为激活函数,以及一个进行多值预测的输出层。用平均绝对误差作损失函数,用随机梯度下降法 adam 作为优化器,设置早期停止函数,以监测损失函数 mse 的收敛状态,模型训练周期数为 200, batch_size 为 1,将所有训练数据和标签数据转换成 3 维张量,采取 80% 的数据作为训练集与 10% 的数据作为测试集进行实验,最后 10% 作为预测集,实验结果见图 4,结果显示在训练集上的 RMSE 为 6381.12,测试集上的 RMSE 为 7777.08,在预测集上的 RMSE 为 5186.30。

2.3 SARIMA 模型 首先对门急诊人次序列进行平稳性检验,经单位根检验(augmented dickey-fuller, ADF)^[12]后结果显示 Dickey-Fuller 为 -24.501, $P < 0.01$,该序列一阶单整,是平稳序列, DW 统计量为 1.9993,接近于 2,序列不存在一阶自相关性,可以建立 SARIMA 模型。根据图 5 中序列自相关和偏相关函数图的拖尾和截尾现象,尝试建立多种模型进行拟合,具体见表 1, with drift 代表有趋势,所以最终模型可以加上 d=1 去除趋势,结合拟合结果和残差 Q-Q 图确定本文 SARIMA 模型为, AIC = 2128.25, AICc = 2129.73, BIC = 2149.55, Box-Ljung 检

验卡方统计量为 0.005706, P 值为 0.9398, 利用模型预测 2019 年 1~12 月数据的效果见图 6, 其中红色部分为 95% 置信区间水平。

2.4 拟合与预测效果比较 分别采用 MSE、RMSE、MAE、MAPE 和 SMAPE 对数据进行拟合和预测效果评价, 预测数据集上的结果见表 2, NARNN、LSTM

和 SARIMA 模型在 MAPE 上的结果分别为 8.22%、4.32% 和 3.40%, 在 SMAPE 上的结果分别为 8.37%、4.33% 和 3.42%, LSTM 和 SARIMA 两者预测效果优于 NARNN, 其中 SARIMA 模型在 4 个指标上的结果均优于 LSTM 和 NARNN, 针对门急诊人次数据拟合和预测效果较好, 三种模型预测效果见图 7。

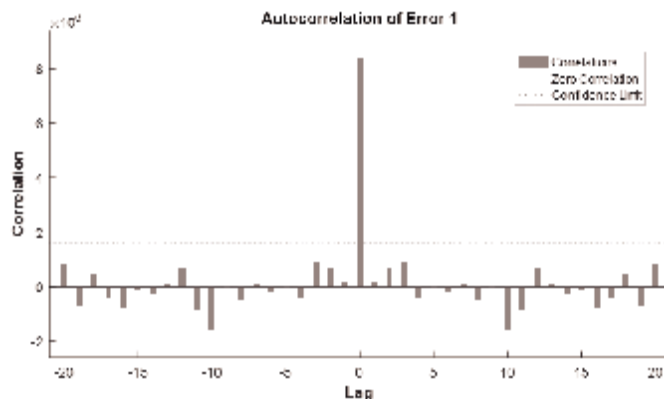


图 1 滞后阶数与模型误差变化

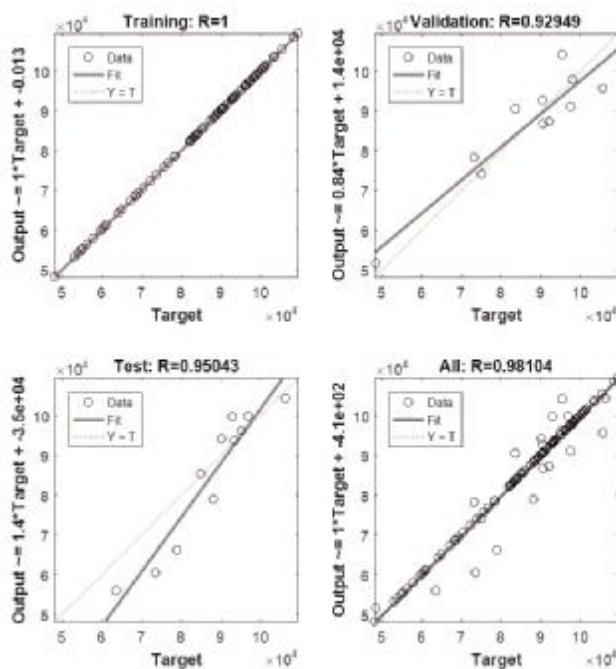


图 2 NARNN 运行结果

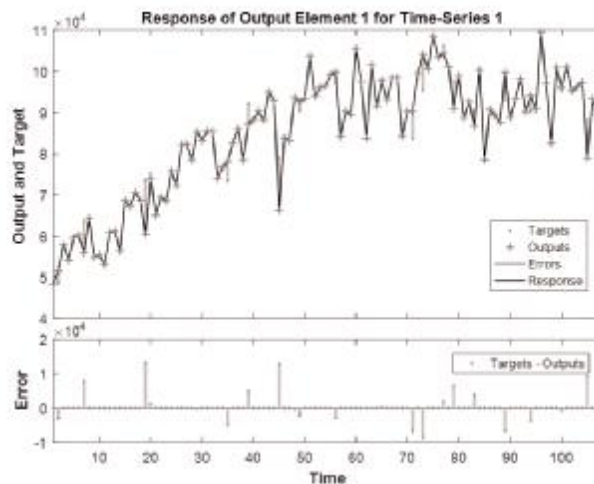


图 3 实际值与拟合值误差对比

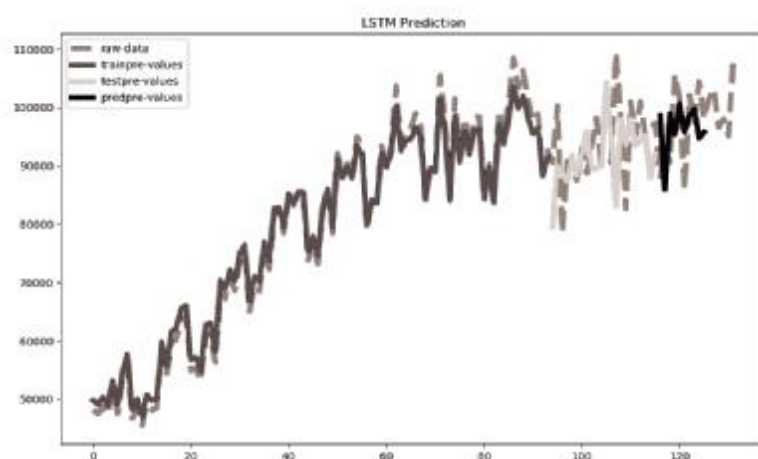


图 4 STM 预测结果图

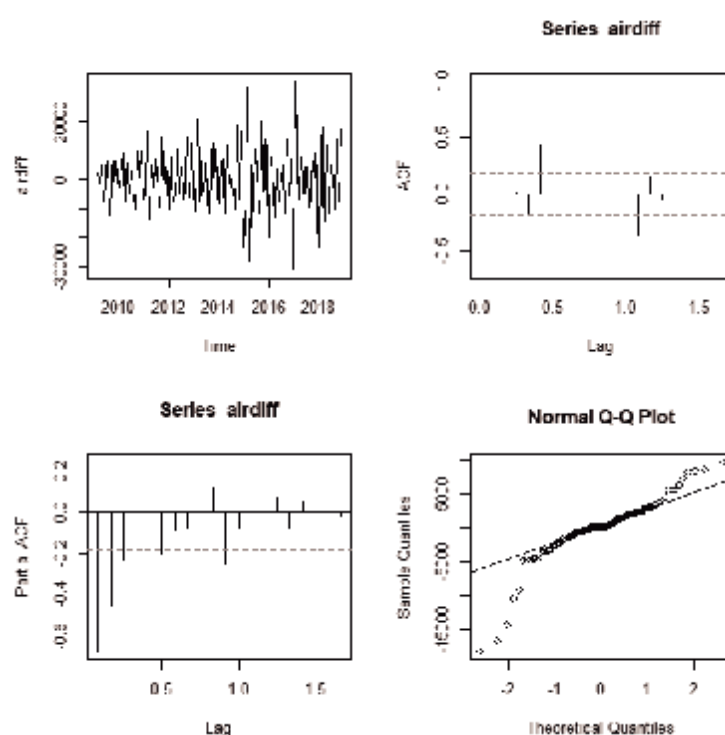


图 5 序列差分、ACF/PACF、Q-Q 图

表 1 过程估计表

模型形式	是否有趋势	AICc 值	模型形式	是否有趋势	AICc 值
ARIMA(0,0,0)(0,1,0) ^[12]	with drift	2276.686	ARIMA(3,0,0)(1,1,1) ^[12]	with drift	2137.897
ARIMA(1,0,0)(1,1,0) ^[12]	with drift	2185.398	ARIMA(4,0,0)(2,1,0) ^[12]	with drift	2132.7
ARIMA(0,0,0)(0,1,0) ^[12]		2274.61	ARIMA(4,0,0)(1,1,0) ^[12]	with drift	2142.014
ARIMA(1,0,0)(0,1,0) ^[12]	with drift	2224.525	ARIMA(4,0,0)(1,1,1) ^[12]	with drift	2134.842
ARIMA(1,0,0)(2,1,0) ^[12]	with drift	2172.923	ARIMA(5,0,0)(2,1,0) ^[12]	with drift	2132.111
ARIMA(1,0,0)(1,1,1) ^[12]	with drift	2174.282	ARIMA(5,0,0)(1,1,0) ^[12]	with drift	2141.073
ARIMA(0,0,0)(2,1,0) ^[12]	with drift	2230.569	ARIMA(5,0,0)(1,1,1) ^[12]	with drift	2134.268
ARIMA(2,0,0)(2,1,0) ^[12]	with drift	2148.954	ARIMA(5,0,0)(2,1,0) ^[12]		2129.73
ARIMA(2,0,0)(1,1,0) ^[12]	with drift	2160.598	ARIMA(5,0,0)(1,1,0) ^[12]		2138.758
ARIMA(2,0,0)(1,1,1) ^[12]	with drift	2150.869	ARIMA(5,0,0)(1,1,1) ^[12]		2131.888
ARIMA(3,0,0)(2,1,0) ^[12]	with drift	2135.64	ARIMA(4,0,0)(2,1,0) ^[12]		2130.368
ARIMA(3,0,0)(1,1,0) ^[12]	with drift	2145.585			

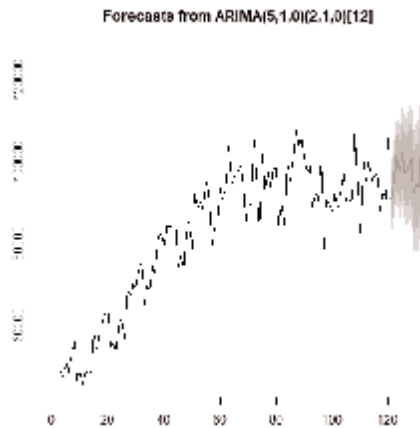


图 6 SARIMA 预测效果

表 2 三种模型在预测数据集上的评价指标结果

评价 指标	MSE	RMSE	MAE	MAPE (%)	SMAPE (%)
NARNN	103338025.80	10165.53	8124.71	8.22	8.37
LSTM	26897672.57	5186.30	4243.70	4.32	4.33
SARIMA	22198571.29	4711.54	3277.05	3.40	3.42

3 讨论

医院门急诊人次数月度数据是典型的时间序列数据,其主要特点表现为受季节影响显著,也会受社会因素、公众假期以及患者口碑等影响明显。NARNN 模型在小样本非线性低维数据预测中具有自适应能力强的特点,适合于解决非线性的时间序列问题。LSTM 模型基于机器学习的思想,不仅集成了传统 RNN 算法的特点,还能够捕获时序数据中的依赖关系,根据长短时记忆特点对未知数据进行充分预测,研究显示该模型在时间序列预测中能产生良好的预测效果。SARIMA 模型能够解决时间序列中受季节性影响较为显著的问题。基于此,本文选取了三种不同的模型对医院门急诊就诊人次数进行拟合预测,也是机器学习方法与传统时序预测方法在解决医疗门急诊就诊人次数预测问题中的一次探索,以期选取较优的预测模型。

本文构建的三种模型通过在训练集上训练预测模型,并对该院门急诊就诊人次数未来一年的数据进行预测,结果显示 SARIMA 模型的预测效果最优,其次是 LSTM, NARNN 模型的预测效果相对较差,在预测精度要求不是非常高的情况下,三种模型均可应用于预测医院门急诊就诊人次数数据的未来变化趋势。SARIMA 模型在对医院门急诊就诊人次数的趋势具有很好的预测能力,泛化能力强,鲁棒性高,能为医院管理决策部门提供更准确的预测数据,从而提高医院在门急诊人财物资源配置方面的效能,在医院服务质量评价体系中具有一定的意义。由于医院门急诊就诊人次数的月度时间序列数据同时具有线性特征和非线性特征,单一模型的预测在

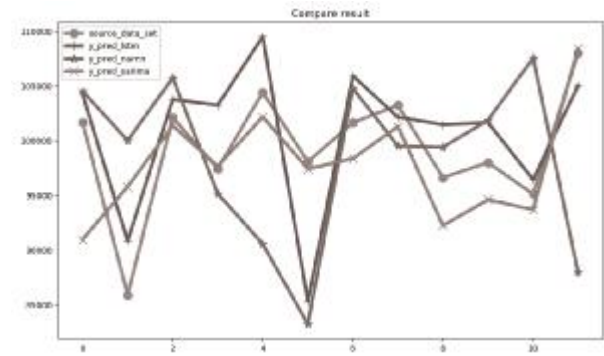


图 7 模型预测效果图

时序预测中存在一定的局限性,未来考虑将模型进行信息融合,以期充分发挥组合模型的优势,得到更精准的预测结果。

参考文献:

- [1]Dudziak T,Gajewski P,Sniezynski B,et al.Neural network modelling studies of steam oxidized kinetic behaviour of advanced steels and Ni-based alloys at 800 °C for 3000 h[J].Corrosion Science,2018,133(APR):94-111.
- [2]傅裕,鲍迎秋,谢沂伯,等.人工智能技术在皮肤病辅助诊断的应用研究[J].中国数字医学,2018,13(10):29-31,38.
- [3]朱大奇,史慧.人工神经网络原理及应用[M].北京:科学出版社,2006.
- [4]Ren G,Gao J.Comparison of NARNN and ARIMA Models for Short-Term Metro Passenger Flow Forecasting [C]//19th COTA International Conference of Transportation Professionals. 2019.
- [5]袁鲁山.基于 NAR 神经网络的车速预测及应用[D].大连理工大学,2016.
- [6]Gers FA,Schmidhuber J,Cummins F,et al.Learning to Forget: Continual Prediction with LSTM [J].Neural Computation, 2000,12(10):2451-2471.
- [7]李琳,王哲,张学良,等.基于 LSTM 深度神经网络的月门诊量预测精度研究[J].中国数字医学,2019,14(1):14-17.
- [8]于畅,张羽.基于 ARMA 模型拟合分析方法的地磁监测数据研究--以长春地磁台观测数据为例 [J]. 防灾减灾学报, 2019,35(3):58-63.
- [9]孙娜,许小珊,冯佳宁,等.ARIMA 与 GM(1,1)模型对我国肺结核年发病人数预测情况的比较 [J]. 中国卫生统计,2019,36 (1):73-76.
- [10]言晨绮,王瑞白,刘海灿,等.ARIMA 模型预测 2018-2019 年我国肺结核发病趋势的应用[J].中华流行病学杂志,2019,40 (6):633-637.
- [11]李芷云,常薇,陈平,等.基于 ARIMA 模型结合回归分析在产科工作量预测中的应用[J].中国医学创新,2019,16(10):161-166.
- [12]左秀霞.带高次趋势项的 ADF 单位根检验[J].数量经济技术经济研究,2019,36(1):152-169.

收稿日期:2020-08-25;修回日期:2020-09-04

编辑/成森