

基于 Prophet-LSTM-PSO 组合模型的 医院住院量预测研究

徐 佩¹, 樊重俊¹, 朱人杰², 黄 耐¹

(1. 上海理工大学 管理学院, 上海 200093; 2. 同济大学附属东方医院, 上海 200120)

摘要: 针对医院住院量预测问题, 首先利用先知模型 (Prophet) 与长短期记忆循环神经网络 (LSTM) 预测方法对 2015 年 1 月到 2019 年 12 月上海市东方医院呼吸内科住院量的时序数据进行建模分析, 然后利用粒子群算法 (PSO) 求出两种模型对应的组合系数, 从而得到最终的 Prophet-LSTM-PSO 组合模型, 并通过 RMSE 和 MAE 统计学指标将组合模型与单一模型进行对比, 同时利用公开数据集进行对比实验。结果表明, Prophet-LSTM-PSO 组合模型较 Prophet、LSTM、移动平均自回归模型等 (ARIMA) 等单一模型有效地降低了医院住院量预测的偏离性, 提高了预测精度。

关键词: Prophet 模型; LSTM 模型; 粒子群算法; 时间序列模型; 住院量预测

中图分类号: TP 183

文献标志码: A

Prediction of hospital inpatients based on combined Prophet-LSTM-PSO model

XU Pei¹, FAN Chongjun¹, ZHU Renjie², HUANG Nai¹

(1. Business School, University of Shanghai for Science and Technology, Shanghai 200093, China;

2. Shanghai East Hospital Affiliated to Tongji University, Shanghai 200120, China)

Abstract: Based on the prediction of hospital inpatients, first Prophet model and long- and short-term memory (LSTM) recurrent neural network prediction method were used to design a combined prediction model based on particle swarm optimization (PSO). Both Prophet and LSTM neural network models were used to simulate and analyze the time series data of the inpatients of respiratory medicine in Shanghai Oriental Hospital from January 2015 to December 2019. Then the particle swarm algorithm was used to find the corresponding combination coefficients of the two models to obtain the final Prophet-LSTM-PSO combined model. By the RMSE and MAE statistical indicators, the combined model was compared with the single model. At the same time, comparative experiments were carried out with open datasets. Results show that Prophet-LSTM-PSO combined model can effectively reduce prediction deviation and improve prediction accuracy compared with single models including Prophet, LSTM, and autoregressive integrated moving average model (ARIMA).

Keywords: Prophet model; LSTM model; particle swarm algorithm; time series model; hospitalization prediction

收稿日期: 2020-03-08

第一作者: 徐 佩 (1996-), 女, 硕士研究生. 研究方向: 智能预测. E-mail: 18309294965@163.com

通信作者: 樊重俊 (1963-), 男, 教授. 研究方向: 大数据分析. E-mail: Fan.chongjun@163.com

2019 年末我国武汉地区爆发新型冠状病毒肺炎,由于爆发在春运期间,严重性不言而喻。举全国之力共抗疫情,但在武汉地区,医院可供床位却远远不足。虽然这属于突发事件,但同时说明对医院住院量的研究非常必要,通过对医院住院量未来动态变化的预测研究,可以提高医院床位的使用率,合理配置各科室医疗床位,最大限度发挥现有医疗资源的作用。

在医疗床位的预测方法中,时间序列预测方法是较为常用的预测方法,并在多种实际应用场景中预测效果较好^[1]。2008 年徐建伟^[2]建立了贝叶斯正规化 BP(back propagation)神经网络,对医院床位数据进行拟和预测,并与指数平滑和自回归模型进行比较,实验结果表明,贝叶斯正规化 BP 神经网络模型较传统模型预测的精度更高一些。2013 年龙会典等^[3]建立了由自回归移动平均模型(ARIMA)、灰色动态模型(GM 模型)和 BP 神经网络集成的时间序列预测模型。首先建立 GM(1, 1)模型,对序列的趋势项进行预测,然后建立基于 ARIMA 和 BP 神经网络的组合模型,对序列的周期波动项进行预测,最后利用乘积模型集成两者预测值。实证结果表明,集成模型的预测效果显著高于单一模型。2014 年曹净等^[4]提出了一种基于小波变换、粒子群优化的最小二乘支持向量机(PSO-LSSVM)和 ARIMA 的基坑变形时间序列预测方法,并成功地应用于昆明某基坑工程的深层水平位移预测中。综合以上研究表明,组合预测模型在不同领域的预测效果均优于单一模型的预测效果。

本文针对上海市东方医院呼吸内科住院量的时间序列数据进行研究,提出了一种基于 Prophet-LSTM-PSO 组合模型的预测方法,分别运用先知模型(Prophet)和长短期记忆循环神经网络(LSTM)对 2015 年 1 月到 2019 年 12 月上海市东方医院呼吸内科住院量的时序数据进行建模,然后利用粒子群算法(PSO)算法求出两种模型对应的组合系数,从而得到最终的 Prophet-LSTM-PSO 组合模型,并与单个预测模型 Prophet, LSTM, ARIMA 模型的预测结果进行比较。结果表明,在医院住院量时序数据预测中,本文提出的组合模型 Prophet-LSTM-PSO 的预测效果优于单个模型的预测效果,预测性能更佳。

1 研究理论与方法

1.1 基于医院住院量的 LSTM 模型

LSTM 是循环神经网络(RNN)的一种特殊形

式,通过遗忘门、输入门和输出门 3 个“门”控制信息量的传入,并通过引入“记忆细胞状态”对其进行长期保存。LSTM 单元结构如图 1 所示。

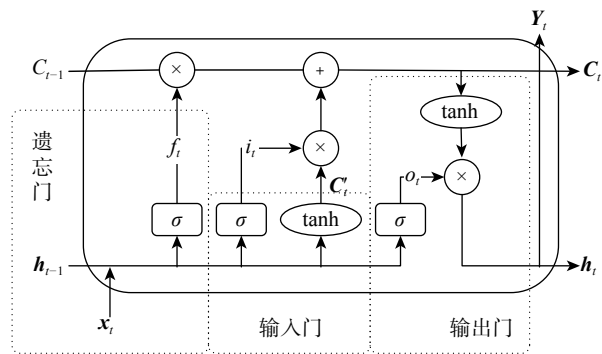


图 1 LSTM 单元

Fig.1 LSTM unit

在医院呼吸内科住院量的时序数据预测中, f_t 为遗忘门的控制函数,通过输出 0~1 之间的数用于对上一时刻的呼吸内科住院量向量数据进行选择。 h_t 为当前时刻的呼吸内科住院量转换向量, x_t 为当前时刻的呼吸内科住院量向量, $[h_{t-1}, x_t]$ 是 2 个向量连接的长向量。 Y_t 为 t 时刻呼吸内科住院量的输出向量。 t 为当前时刻, $t-1$ 为上一时刻。 W 为各层对应的权重矩阵, b 为偏差向量,下标 f, i, o 分别代表遗忘门、输入门和控制门。

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

i_t 为输入门的控制函数,与 f_t 类似,通过改变权重矩阵的值对新输入的呼吸内科住院量数据进行选择,以避免相关性较低的呼吸内科住院量值加入记忆细胞。 C'_t 是当前时刻真实的呼吸内科住院量的 tanh 函数转换值,作为当前时刻呼吸内科住院量的信息向量。

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

$$C'_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (3)$$

当前时刻记忆细胞的存储向量 C_t 用于对当前时刻呼吸内科住院量向量进行选取、转换、存储, C_t 是将上一时刻的存储向量和当前时刻新输入的呼吸内科住院量的信息向量分别与遗忘门控制函数向量、输入门控制函数向量进行逐元素相乘,进而实现对较长时间的呼吸内科住院量相关信息量的记忆存储。

$$C_t = f_t \odot C_{t-1} + i_t \odot C'_t \quad (4)$$

o_t 为输出门的控制函数,用于对当前时刻呼吸内科住院量预测向量的输出控制。

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

$$h_t = o_t \odot \tanh C_t \quad (6)$$

$\bar{L}_t$ 为经转换之后的呼吸内科住院量预测值, $\min J(\theta)$ 代表 LSTM 的损失函数。

$$\bar{L}_t = W_y \cdot h_t \quad (7)$$

$$\min J(\theta) = \sum_{i=1}^T (\bar{L}_t - L_t) \quad (8)$$

式中: W_y 为输出门上一时刻呼吸内科住院量的存储值权重矩阵; L_t 为呼吸内科住院量实际值。

LSTM 模型通过采用损失函数对输出值的倒数作为误差项随时间进行反向传播, 该算法称为基于时间的反向传播算法(BPTT), 其基本思想主要分为4个步骤:

步骤1 运用式(1)~(7)前向计算出各个神经元的输出值 f_t , i_t , C_t , o_t , h_t , $\bar{L}_t$;

步骤2 从时间和网络层级2个方向反向计算每个神经元的损失函数对其输出值的倒数, 即每个神经元的误差项;

步骤3 根据每个神经元的误差项值对其相应的权重计算梯度;

步骤4 运用梯度下降法更新权重矩阵和偏差向量。

1.2 基于医院住院量的 Prophet 模型

在使用 LSTM 对原始数据进行预测的同时, 采用 Prophet 模型进行预测。

2017年 Facebook 发布了 Prophet 时序先知模型, 主要研究时序数据的特征和时序变化规律, 同时预测未来的走势。Prophet 的核心是分析包含周期性、趋势性以及节假日效应等的时间序列特征, 它是一种针对时序的有效集成解决方案, Prophet 模型的构成为

$$P(t) = g(t) + s(t) + h(t) + \varepsilon \quad (9)$$

式中: $g(t)$ 为趋势函数, 用来分析时间序列中的非周期性的变化; $s(t)$ 代表周期性的变化, 例如, 一年或一周的周期性; $h(t)$ 代表节假日等偶然一天或几天造成的影响; ε 为随机波动, 代表模型没有考虑到的误差的影响^[5]。

a. 趋势项。Facebook 采用改进的 Logistic 增长模型, 其中, 饱和值是随时间动态变化的, 而且增长率也会随着新产品等因素发生变化。

$$g(t) = \frac{C(t)}{1 + \exp(-(k + a(t)^T \delta)) (t - (m + a(t)^T \gamma))} \quad (10)$$

$$a(t) = \begin{cases} 1, & t > S_j \\ 0, & \text{其他} \end{cases} \quad (11)$$

式中: $C(t)$ 为随时间变化的饱和值(承载能力);

$k + a(t)^T \delta$ 为随时间变化的增长率, $(m + a(t)^T \gamma)$ 为相应的偏置参数; S_j 为若干个转折点; δ 为转折点处增长率的变化量。

b. 周期项。Prophet 模型依靠傅里叶级数来构造灵活的周期性模型, 基本形式为

$$s(t) = \sum_{n=1}^N \left(a_n \cos\left(\frac{2\pi n t}{p}\right) + b_n \sin\left(\frac{2\pi n t}{p}\right) \right) \quad (12)$$

式中: p 为时间序列的周期长度; N 为周期数; a_n , b_n 为需要估计的参数。

1.3 基于医院住院量的 PSO 模型

在多元回归中, 求解回归参数的常用方法有最小二乘法、梯度下降法及牛顿法等, 但是, 这些方法涉及大量的矩阵计算, 且计算速度较慢。而 PSO 算法最初是基于自然界中存在的鸟类捕食活动现象所产生的智能模拟算法, 该算法相较于遗传算法(GA)、蚁群算法(ACO)等具有简单高效、参数少及收敛速度快等特点。因此, 本文借助粒子群算法确定组合模型的权重系数。

PSO 算法将回归系数组合 (α, β) 看成一个粒子, 随机初始化多个粒子, 每个粒子通过位置和速度更新方程, 向最优解“靠拢”。

$$v_i(k+1) = wv_i(k) + c_1 r_1 (p_i - x_i(k)) + c_2 r_2 (p_g - x_i(k)) \quad (13)$$

$$x_i(k+1) = x_i(k) + v_i(k+1) \quad (14)$$

式中: $v_i(k)$, $x_i(k)$ 分别为粒子的速度和位置; c_1 , c_2 分别为粒子移动到该粒子历史最优位置 p_i 和当前所有粒子最优位置 p_g 的加速项因子; r_1 , r_2 为0~1之间的随机因子; w 为惯性因子, 是粒子之前运动方向在当前运动方向上的惯性。

最后, 通过定义适应度函数

$$Q(\theta) = \sum_{i=1}^k (\alpha \bar{L}_t + \beta \bar{P}_t - L_t)^2$$

来评价组合模型的拟合优劣性。 $\bar{L}_t$ 为 LSTM 模型的预测值, $\bar{P}_t$ 为 Prophet 模型的预测值。

2 Prophet-LSTM-PSO 模型构建与预测

2.1 数据概况及预处理

为了验证模型的可行性与有效性, 选取了上海市东方医院(本院及南院)呼吸内科 2015 年 1 月到 2019 年 12 月这 60 个月住院量数据为研究对象。将数据中存在的缺失值, 运用历史同期和近 3 期数据的平均值进行填充, 获得完整的原始输入

数据。对原始输入数据, 首先将其转换为与模型输入维度相同的类型, 然后将转换后的数据集分为两类, 其中, 80% 的数据即从 2015 年 1 月到 2018 年 12 月的 96 个数据(本院 48 个, 南院 48 个)用于训练模型, 剩余的 2019 年 1 月至 2019 年 12 月的 24 个(本院 12 个, 南院 12 个)数据作为测试集, 用于对模型拟合效果进行验证。

2.2 Prophet-LSTM-PSO 组合模型构建

时间序列数据中蕴含着大量不确定信息, 应用单一模型的预测效果往往不是很理想。因此, 为了提高预测的准确率, 将 LSTM 和 Prophet 两种模型进行组合, 应用 PSO 算法求解两种模型的组合系数以提高求解速度, 充分发挥组合模型的优势。

- 基于 LSTM 和 Prophet 组合模型的建模步骤:
- 步骤 1 对数据中缺失的小部分数据, 结合近 3 个月数据趋势和每一年同期数据, 采用平均法进行填充;
 - 步骤 2 对经过步骤 1 处理的原始数据进行归一化处理, 使得训练数据的极差尽可能小。然后运用式(1)~(8)对时间序列 $\{X_t\}$ 进行训练和预测, 得到 LSTM 模型的预测值 $\bar{L}_t$;
 - 步骤 3 对经过步骤 1 处理的原始数据进行 log 处理, 使其尽量符合正态分布。然后运用式(9)~(12)对时间序列 $\{X_t\}$ 进行训练和预测, 得到 Prophet 模型的预测值 $\bar{P}_t$;
 - 步骤 4 通过构建适应度函数 $Q(\theta)$, 使得组合模型预测值尽可能逼近真实值, 并采用 PSO 算法对组合模型的系数 α 和 β 进行求解;
 - 步骤 5 组合两种模型的预测结果, 获得组合模型预测值 $\bar{X}_t = \alpha \bar{L}_t + \beta \bar{P}_t$ 。

2.3 Prophet-LSTM-PSO 组合模型训练

2.3.1 LSTM 模型训练参数

在 Python 3.6 软件的基础上, 将 LSTM 模型分别应用于训练集和测试集, 获得单一模型下的各种具体参数, 其中, 输入向量为(2, 1)的维度的时序数据, 输入层神经元数为 8, 隐藏层为 1 层, 神经元数为 128, 隐藏层激活函数为 RELU 函数, 即为 $f(x) = \max(0, x)$, 输出层神经元数为 1。本实验中训练数据“轮”次数为 100, 采用随机梯度下降算法。

2.3.2 Prophet 模型训练参数

在 Python 3.6 软件的基础上, 将 Prophet 模型分别应用于 log 处理后的数据集, 获得单一模型下的

各种具体参数, 输入向量为(2, 60)的时序数据, 趋势项为改进的 Logistic 增长模型, 周期项取傅里叶级数, 拟合函数为 Fit 函数。

2.3.3 组合模型训练与结果分析

将两种单一模型的预测结果进行组合, 并应用 PSO 算法求解组合模型的系数 α 和 β , 从而获得组合模型的预测值。

本院的组合预测模型为

$$\bar{X}_t = 0.301\ 510\ 37\bar{L}_t + 0.711\ 636\ 83\bar{P}_t$$

南院的组合预测模型为

$$\bar{X}_t = 0.152\ 072\ 93\bar{L}_t + 0.858\ 231\ 58\bar{P}_t$$

3 实验结果与分析

将组合模型与常用的单一模型(Prophet, LSTM, ARIMA 模型)在相同的测试集进行定量分析, 结果表明, LSTM-Prophet-PSO 组合模型的预测结果较单一模型要好, 预测值与实际值的误差值较小, 为了更加直观地观察预测结果, 将上海市东方医院(本院)住院量的 4 种模型实际值与预测值进行可视化, 如图 2 所示。从图 2 中可以看出, 组合模型拟合效果最好, 其次是 Prophet 模型, 再次是 ARIMA 模型, 最后是 LSTM 模型。与其他几种单一模型相比, 组合模型在训练集和测试集上都有较好的拟合效果, 符合时序数据的大致趋势。

为了对模型的效果进行科学的评估, 分别计算了各个模型的整体均方根误差(RMSE)和平均绝对误差(MAE), 并将 LSTM-Prophet-PSO 组合模型与 Prophet, LSTM, ARIMA 模型进行对比, 以验证该方法的可行性与有效性。不同模型在相同指标下的对比结果如表 1 所示。

为了增加本文的可信度, 采用公开数据集, 即我国新型冠状病毒肺炎疫情的累计确诊量的时序数据进行对比实验, 实验数据为 2020 年 1 月 24 日到 2020 年 4 月 19 日我国累计确诊人数, 应用本文所建立的模型, 得到组合模型为 $\bar{X}_t = -0.064\ 883\ 51\bar{L}_t + 1.064\ 045\ 72\bar{P}_t$, 并分别计算 LSTM-Prophet-PSO 组合模型与 Prophet, LSTM, ARIMA 模型的 RMSE 和 MAE, 如表 2 所示。

从表 1 和表 2 可以看出, 组合模型与其他 3 种单一模型相比, 预测精度最高, 其相对于 Prophet, LSTM, ARIMA 这 3 种单一模型在 RMSE 和 MAE 上都有明显的提高。

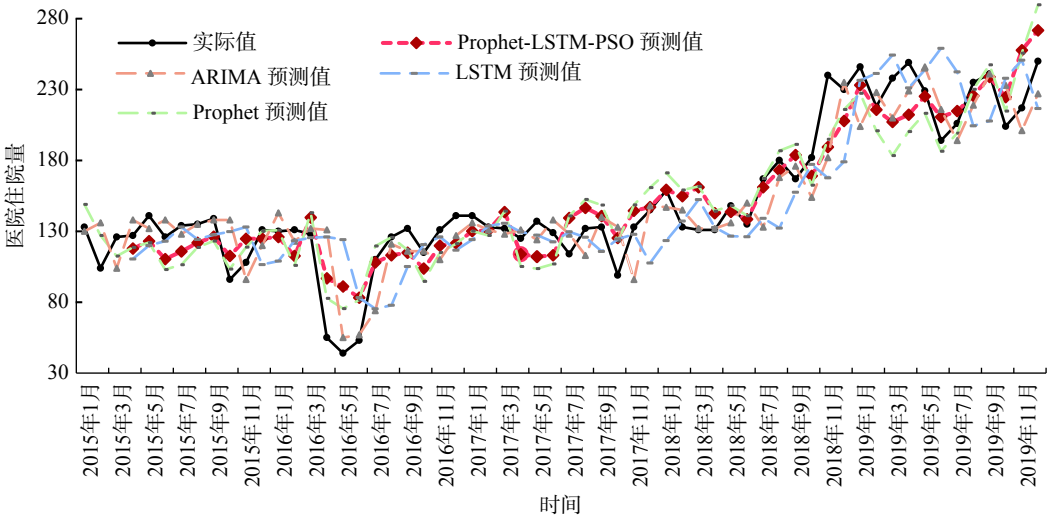


图2 模型的实际值与预测值对比

Fig.2 Comparison of actual and predicted values of the model

表 1 实验组：不同模型的整体均方根误差与平均绝对误差对比

Tab.1 Trial group: comparison of RMSE and MAE of different models

模型类别	整体均方根误差	平均绝对误差
ARIMA(0,1,2)	21.488 5	16.789 4
LSTM	29.594 2	22.733 8
Prophet	21.685 7	18.067 1
LSTM-Prophet-PSO	19.641 3	15.984 3

表 2 对比实验组：不同模型的整体均方根误差与平均绝对误差对比

Tab.2 Control group: comparison of RMSE and MAE of different models

模型类别	整体均方根误差	平均绝对误差
ARIMA(0,2,1)	17.704 5	7.236 5
LSTM	26.795 6	12.823 7
Prophet	11.716 1	6.228 9
LSTM-Prophet-PSO	10.648 4	6.013 6

4 结束语

通过分析医院住院量的时序数据的特征与趋势，进一步分析医院住院量预测的准确度，对于医院及时调整医院资源配置具有重要的参考意义。根据上海市东方医院呼吸内科住院量时序数据的规律，提出了一种基于 Prophet 模型和长短期记忆网络 (LSTM) 的组合模型预测方法，通过粒子群算法确定组合模型中的权重系数，并与 Prophet, LSTM, ARIMA 模型这 3 种单一模型的预测效果进行对比。利用公开数据集，即我国新型冠状病毒肺炎疫情的累计确诊量的时序数据进行对比实

验。实验结果表明，通过 PSO 计算得出的组合模型的预测性能优于单个 Prophet, LSTM, ARIMA 模型的预测性能。本文对医院住院量的研究只考虑了历史数据的影响，在进一步预测时可考虑多个影响因素，例如，气温、湿度等因素对呼吸内科住院量的影响^[6-9]。

参考文献：

[1] 葛娜, 孙连英, 石晓达, 等. Prophet-LSTM 组合模型的销售量预测研究 [J]. 计算机科学, 2019, 46(S1): 446–451.

[2] 徐建伟. 贝叶斯正规化 BP 神经网络在我国医院床位预测中的应用 [J]. 山西医科大学学报, 2008, 39(9): 833–835.

[3] 龙会典, 严广乐. 基于 SARIMA、GM(1,1) 和 BP 神经网络集成模型的 GDP 时间序列预测研究 [J]. 数理统计与管理, 2013, 32(5): 814–822.

[4] 曹净, 丁文云, 赵党书, 等. 基于 LSSVM-ARMA 模型的基坑变形时间序列预测 [J]. 岩土力学, 2014, 35(S2): 579–586.

[5] 李丽萍, 段桂华, 王建新. 基于 Prophet 框架的银行网点备付金预测方法 [J]. 中南大学学报 (自然科学版), 2019, 50(1): 75–82.

[6] 王晓飞, 王波, 陆玉玉, 等. 基于 Prophet-LSTM 模型的 PM_{2.5} 浓度预测研究 [J]. 软件导刊, 2020, 19(3): 133–136.

[7] 王敏珍, 郑山, 王式功, 等. 气温与湿度的交互作用对呼吸系统疾病的影响 [J]. 中国环境科学, 2016, 36(2): 581–588.

[8] 莫运政, 郑亚安, 陶辉, 等. 日均气温与呼吸系统疾病急诊人次相关性的时间序列分析 [J]. 北京大学学报 (医学版), 2012, 44(3): 416–420.

[9] 樊重俊, 金阳, 杨云鹏. 基于节点活跃度与疾病交互双重影响因素的传染病模型 [J]. 系统管理学报, 2017, 26(5): 941–946.