文章编号：2095-2163(2020)02-0204-04 中图分类号：TP183 文献标志码：A

集成气象环境数据的门诊量预测研究

张家艳，郑建立
（上海理工大学，上海 200093）

摘 要：为了合理配置卫生资源，建立门诊量精确预测模型，本文使用门诊历史数据结合气象数据和环境监测数据，采用差分处理的 xgboost 方法进行预测。结果表明，此模型在测试集上决定系数 R^2 为 0.805，平均绝对百分比误差 *mape* 为 4.7%，优于未加入气象数据及环境监测数据的门诊量预测（R^2 为 0.757，*mape* 为 5.3%）。该模型能够对门诊量进行较为准确的预测，为日值卫生资源的合理分配提供依据。

关键词：气象数据；环境监测因素；差分；xgboost；门诊量

Research on outpatient forecast with integrated meteorological and environmental data

ZHANG Jiayan, ZHENG Jianli
(University of Shanghai for Science and Technology, Shanghai 200093, China)

【Abstract】 In order to allocate the hospital´s health resources rationally and establish an accurate model for outpatients volume prediction, history outpatient data combined with meteorological data and environmental monitoring data is used with differential processing and xgboost methods for prediction. The results show that the determination coefficient R^2 is 0.805 and the mean absolute percentage error *mape* is 4.7% in test dataset. The R^2 and *mape* of outpatients volume prediction which did not include meteorological data and environmental monitoring data is 0.757 and 5.3%. This model can predict the outpatient volume of the hospital accurately and provide a basis for allocating the daily health resources rationally.

【Key words】 meteorological data; environmental monitoring data; difference; xgboost; outpatient visits

0 引 言

门诊是医院对外服务的窗口。人口老龄化导致患者人数的增加以及人们越来越关注自身健康状况，每年的门诊压力越来越大。此外，门诊量与体检和住院服务的工作量直接相关[1]。对门诊病人数量进行准确和可靠的预测，有助于科学合理地分配医院的人力物力资源如医生坐诊人数、医疗设备等，从而能更好应对门诊压力。

在研究门诊量随时间变化过程中，影响其变化的因素太多，难以考虑全面。由于时间序列模型仅考虑日期因素，故常把门诊预测当作时间序列数据分析。在时间序列模型中，最常见的是差分整合移动自回归模型（ARIMA）[2-3]。ARIMA 起初是出于经济学目的设计，现已广泛用于医学领域。如范晓欣等人[4]用 ARIMA 预测门急诊人次，*mape* 为 7.01%。近年来，人们采用深度学习等新技术预测，如 Wang 等人[5]将时间序列分解，再用广义回归神经网络模型预测；Huang 等人[6]使用经验模式分解结合粒子群算法优化的反向传播人工神经网络预测；相比传统技术，均得到了更准确的预测结果。但深度学习技术在大数据量上效果较好，针对少数据量的情况，常采用机器学习方法。Islam 等人[7]用支持向量回归预测社区医院的门诊人次。Yang 等人[8]用多层感知器预测门诊就诊上呼吸道感染人数。2016 年以来，xgboost[9]在 Kaggle 等各大数据科学比赛中都得到广泛应用，成为比赛中的高分模型。

空气污染是一个重大的全球性问题，空气里的污染物能够影响人体健康。同时，天气的变化也能给人带来不适，这些可能都影响门诊量的变化。Seo 等人[10]采用环境监测数据与气象数据建立韩国结膜炎门诊量预测模型，发现门诊量 O_3 浓度相关系数为 0.49。经学者研究发现，台湾干眼病与环境监测因素如一氧化碳、二氧化氮等的含量正相关（$P < 0.05$）[11]。因此，在对门诊量进行预测时，选择气象因素及环境监测因素作为门诊预测因素是必要的。

1 xgboost 算法

极端梯度提升（Extreme Gradient Boosting，xgboost）是在集成学习 GBDT 的基础上对目标函数进行了二阶泰勒展开，在陈刚等人[12]提出之后，就得到了广泛的应用，在许多问题上得到了优胜的解

作者简介：张家艳（1995-），女，硕士研究生，主要研究方向：医学信息处理；郑建立（1965-），男，博士，副教授，主要研究方向：医学信息集成。

收稿日期：2019-11-20

决方案。

1.1 xgboost

xgboost 是在 GBDT 上进行改进的算法，故也是由 k 个 cart 树集成学习而来。但不同的是 GBDT 的基函数为决策树，而 xgboost 的基函数为其他的机器学习器。在 xgboost 中，损失函数的计算公式在 GBDT 的损失函数的基础上加上了正则化项 $\Omega(h_t)$，即损失函数为：

$$L_t = \sum_{i=1}^{m} L(y_i, f_{t-1}(x_i) + h_t(x)) + \Omega(h_t), \quad (1)$$

其中，$\Omega(h_t)$ 计算公式为：

$$\Omega(h_t) = \gamma J + \frac{\lambda}{2}\sum_{j=1}^{J} w_{t,j}^2, \quad (2)$$

其中，γ，λ 为正则化系数；J 为叶子节点的个数；$w_{t,j}$ 为对应叶子节点 $R_{t,j}$ 的输出值。

在 GBDT 中，损失函数仅仅对误差部分做负梯度、即一阶泰勒展开，但在 xgboost 中对误差部分做二阶泰勒展开，从而使拟合结果更准确，即：

$$L_t = \sum_{i=1}^{m}\left[L(y_i, f_{t-1}(x_i)) + \frac{\partial L(y_i, f_{t-1}(x_i))}{\partial f_{t-1}(x)} h_t(x_i) + \frac{\partial^2 L(y_i, f_{t-1}(x_i))}{\partial f_{t-1}^2(x)} h_t^2(x_i)\right] + \gamma J + \frac{\lambda}{2}\sum_{j=1}^{J} w_{t,j}^2, \quad (3)$$

在最小化损失函数的过程中，由于 $L(y_i, f_{t-1}(x_i))$ 为常数，故不影响最小化的过程，可省略。同时，由上知 $w_{t,j}$ 的定义，而 $h_t(x_i)$ 为 x_i 经过第 t 个决策树处理后在子节点区域的输出值，故式(3) 可改为：

$$L_t \cong \sum_{j=1}^{J}\left[\sum_{x_i \in R_{t,j}} g_{t,i} w_{t,j} + \frac{1}{2}\sum_{x_i \in R_{t,j}} (h_{t,i} + \lambda) w_{t,j}^2\right] + \gamma J. \quad (4)$$

其中，

$$g_{t,i} = \frac{\partial L(y_i, f_{t-1}(x_i))}{\partial f_{t-1}(x)}, h_{t,i} = \frac{\partial^2 L(y_i, f_{t-1}(x_i))}{\partial f_{t-1}^2(x)}。$$

1.2 xgboost 的差分化处理

机器学习技术广泛应用在机器翻译，语音识别等领域，而这些领域的数据有些是非数值数据，很少有数值不平稳的情况。但作为时间序列数据，门诊量经常会存在不平稳的情况。非平稳序列包含了季节、趋势等因素，这些不确定因素使预测结果的准确性降低[13]。故在数据预处理时，先对数据进行平稳性检测，如不平稳，常用的方法是进行差分化处理，即将数据的后一个数减去前一个数，依次相减得到数据集，重复检测直到数据转化成平稳序列，最后将预测结果进行反差分化得到最终预测结果。

本文数据采用 ADF 平稳性检测结果得到 p - $value$ 值为 0.874 3，即数据是不平稳的。将数据集进行一次差分后便发现 p - $value$ 值变为 0，数据已经变成平稳序列了。

2 差分化 xgboost 门诊量预测

2.1 数据获取与预处理

本文采用 kettle 工具抽取了上海市某三甲医院 2017/01～2019/05 年的门诊日值数据。对假期和周末的门诊异常值，汇合气象环境数据后一起进行处理。

在中国气象数据网上，下载了对应 2017/01～2019/05 的气象数据。这些数据总共包括 22 个气象特征，对其中缺失值和异常值采用均值法进行处理。选取的主要气象数据特征及值见表 1。

表 1　部分气象数据特征及值

Tab. 1　Some characteristics and values of meteorological data

气象特征	值(某一天的平均值)
地表气温/℃	9.7
风速/(m·s^{-1})	1.4
降水量/mm	0
气压/hpa	1 025.4
日照时数/(0.1h)	2.2
气温/℃	9.1
相对湿度/%	88
蒸发量/mm	0.6

同期环境监测数据来自于 pm2.5 历史数据网站[13]，其中的数据全部来自于国家环境保护部。对数据中的缺失数值采用均值处理。环境监测数据特征及值见表 2。

表 2　环境监测数据特征及值

Tab. 2　Characteristics and values of environmental monitoring data

环境监测数据特征	值(某一天)
空气质量指数(AQI)	127
细颗粒物(pm2.5)	96
可吸入颗粒物(pm10)	95
二氧化硫浓度(SO_2)	18
一氧化碳浓度(CO)	1.2
二氧化氮浓度(NO_2)	74
8 小时臭氧浓度(O_3_8h)	73

整合上述三份数据，删除其中的周末以及假期数据，共得到 565 份数据。从这 565 份数据中选取 508 份作为训练集，将剩下的 57 份数据作为预测集。

2.2 xgboost 模型训练

2.2.1 超参数取值

本算法在调参时首先采用随机搜索调参法，确定大致的参数范围，然后采用网格搜索调参法获取最优的参数组合。

在本次随机搜索调参时，采用三折交叉验证，*n_iters* 选择为 10 即搜索次数为 10。在随机搜索调参的结果上，取每个超参数左邻和右邻几个数一起作为网格搜索参数的初始值，最终得到的网格搜索参数结果见表 3。

表 3 超参数取值

Tab. 3 Hyper-parameter values

超参数	取值
Learning_rate	0.1
max_depth	4
min_child_weight	1
colsample_bytree	0.84
subsample	0.69
Reg_lambda	0.999
reg_alpha	0.001 1
gamma	0.000 099 9
n_estimators	400

2.2.2 评价标准

对于门诊量预测的结果，采用平均百分比误差（mean absoulute percent error，*mape*）来衡量预测值与真实值之间的差距，采用模型拟合度 R^2 来衡量模型的拟合程度，其计算公式具体如下：

$$mape = \sum_{t=1}^{n} \left| \frac{y_{real} - y_{pred}}{y_{real}} \right| \times \frac{100}{n}, \tag{5}$$

$$R^2 = 1 - \frac{\sum (y_{real} - y_{pred})^2}{\sum (y_{real} - y_{avg})^2}. \tag{6}$$

其中，y_{real} 表示实际的门诊量值；y_{pred} 表示门诊量预测值；y_{avg} 表示门诊量平均值。

mape 越小说明预测的准确程度越高，R^2 越大说明模型选择越合理。

2.3 结果

2.3.1 门诊量预测

在模型确定之后，便可以训练模型进行门诊量预测。未来 50 天日门诊量预测值与真实值的对比曲线如图 1 所示。其中，虚线即为未来 50 天的预测值，实线为未来 50 天的真实值，由图 1 可以看出除了在最高值或最低值处有部分偏差之外，预测走向基本一致。

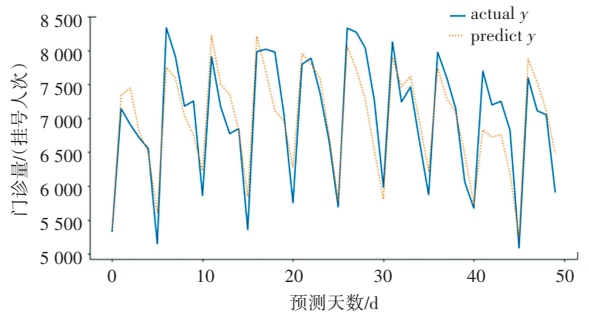

图 1 未来 50 天日门诊量预测值与真实值

Fig. 1 Forecast and actual values of daily outpatient visits in future 50 d

2.3.2 差分化处理对预测结果的影响

在数据预处理后，将数据进行差分化处理，转化为平稳序列，再进行预测，最终得到的预测结果是差分数据。在计算 R^2 和 *mape* 时，需要将差分数据反差分化得到最终数据集。差分处理与未进行差分处理的 R^2 和 *mape* 值见表 4。可以看出，对于非稳定的时间序列数据，差分处理对预测结果的准确性影响明显，因此在预测之前进行差分处理是必要的。

表 4 差分对模型结果的影响

Tab. 4 The influence of difference on model results

模型	R^2	*mape*/ %
差分化 xgboost 模型	0.805	4.7
原 xgboost 模型	0.766	5.1

2.3.3 气象及环境监测因素对预测结果的影响

对于门诊量的预测，传统的方法就是采用日期和门诊量数据当作时间序列数据进行预测。这种方法解决了门诊量影响因素太多无法选取全部因素的问题。在此方法中，只用时间变量来替代所有变化的因素，从而达到大致较好的预测结果。但时间因素是个笼统的特征，内在的变量太多，时间变量并不能完全替代这些变量。本文将对疾病影响较大的气象因素及环境监测数据再加上时间变量一起预测门诊量，比仅采用时间变量预测效果好。结果见表 5。

表 5 预测结果对比

Tab. 5 Comparison of prediction results

特征	评估 R^2	评估 mape/%
环境监测+气象+时间	0.805	4.7
时间	0.757	5.3

3 结束语

门诊量数据为时间序列数据，由于数据随着时间变化存在波动现象，即数据是不稳定的，常见的机器学习和深度学习领域，很少需要时间序列处理，故该领域的常规化处理思路便没有差分化这一方法。

为了改善预测效果，翻阅了大量统计学文献后，进行了差分化处理，结果显示采用差分化处理后的 xgboost 模型预测方法，得到的预测结果之平均绝对百分比误差低于原生数据的 xgboost 模型，展现了强大的预测能力。而且，相对于仅用时间来预测门诊量的方法，将气象、环境监测因素引入门诊量预测的方法，其平均绝对百分比误差及模型拟合度均获得更好的效果。与其他门诊量预测的研究相比[7-8]，本文提出的模型的预测结果高于平均水平。由于数据集较小，深度学习方法效果不好，在后续的研究中，可以考虑抽取更多的临床数据以扩大数据量以及引入经济因素，并采用优化的深度学习模型，进一步增大预测的准确性。

参考文献

[1] LUO Li, LUO Le, ZHANG Xinli, et al. Hospital daily outpatient visits forecasting using a combinatorial model based on ARIMA and SES models[J]. BMC health services research, 2017, 17(1): 469.

[2] ZHOU Ruyi, WU Dasheng, LI Yang, et al. Relationship between air pollutants and outpatient visits for respiratory diseases in Hangzhou[C]// 2018 9th International Conference on Information Technology in Medicine and Education (ITME). Hangzhou, China: IEEE Computer Society, 2018: 275.

[3] HE Z, TAO H. Epidemiology and ARIMA model of positive-rate of influenza viruses among children in Wuhan, China: A nine-year retrospective study[J]. International Journal of Infectious Diseases, 2018, 74: 61.

[4] 范晓欣, 隋虹. ARIMA 乘积季节模型在医院门急诊人次预测中的应用[J]. 中国医院管理, 2015, 35(4): 41.

[5] WANG Yongming, GU Junzhong, ZHOU Zili, et al. Diarrhoea outpatient visits prediction based on time series decomposition and multi-local predictor fusion[J]. Knowledge-Based Systems, 2015, 88: 12.

[6] HUANG D, WU Z. Forecasting outpatient visits using empirical mode decomposition coupled with back-propagation artificial neural networks optimized by particle swarm optimization[J]. Plos One, 2017, 12(2): e0172539.

[7] ISLAM K S, SHAHARIA A, ISLAM N, et al. Improving healthcare services of community clinics using machine learning techniques[C]//2018 International Conference on Innovations in Science, Engineering and Technology (ICISET). Chittagong, Bangladesh: IEEE, 2018: 437.

[8] YANG P H, HSIEH M T, LIN G M, et al. Prediction of outpatient visits for upper respiratory tract infections by machine learning of PM2. 5 and PM10 levels in Taiwan[C]//2018 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW). Taichung, Taiwan: IEEE, 2018: 1.

[9] CHEN T, GUESTRIN C. XGBoost: A scalable tree boosting system[C]// Proceedings of 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, California, USA: ACM, 2016: 785.

[10] SEO J W, YOUN J S, PARK S J, et al. Development of a conjunctivitis outpatient rate prediction model incorporating ambient Ozone and meteorological factors in South Korea[J]. Frontiers in pharmacology, 2018, 9: 1135.

[11] ZHONG Jiayu, LEE Y C, HSIEH C J, et al. Association between dry eye disease, air pollution and weather changes in Taiwan[J]. International Journal of Environmental Research and Public Health, 2018, 15(10): 2269.

[12] 陈刚，丁慧玲. 基于主成分分析的模糊时间序列模型的平稳化算法[J]. 控制与决策，2018，33(9)：1643.

[13] 王杰. 历史数据[EB/OL]. [2019-11-13]. https://www.aqistudy.cn/historydata/weather.php，2018-04-01/.

(上接第 203 页)

3　结束语

(1)改进了三自由度汽车模型的数学表达式，降低了因参数过多过繁而导致的模型复杂程度，使其更容易理解和应用。

(2)对比了在不同车速下低自由度汽车模型的角阶跃输入响应，结果表明，本文建立的三自由度汽车模型是准确可靠的，基本可以反映出汽车的操纵特性。

(3)三自由度汽车模型可以反映汽车车身的侧倾状态，有利于研究汽车的侧倾稳定性，而且汽车转弯行驶时随着车速的提高，汽车车身的侧倾状态越来越严重。

参考文献

[1] 余志生. 汽车理论[M]. 5 版. 北京：机械工业出版社，2009.

[2] 金贤建. 分布式驱动电动汽车状态参数估计与侧向稳定性鲁棒控制研究[D]. 南京：东南大学，2017.

[3] 金智林. 运动型多功能汽车侧翻稳定性及防侧翻控制[D]. 南京：南京航空航天大学，2008.

[4] 李双，刚宪约，于海兴. 汽车操纵动力学三自由度模型与转向特性仿真[J]. 机械设计与制造，2015(3)：260.

[5] 喻凡，林逸. 汽车系统动力学[M]. 2 版. 北京：机械工业出版社，2017.

[6] 郭孔辉. 汽车操纵动力学[M]. 长春：吉林科学技术出版社，1991.