计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(1):243−249 [doi: 10.15888/j.cnki.csa.007750]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 SSA-LightGBM 的交通流量调查数据趋势预测①

徐　磊, 孙朝云, 李　伟, 杨荣新

(长安大学 信息工程学院, 西安 710064)
通讯作者: 徐　磊, E-mail: 2191117492@qq.com

摘　要: 为了解决传统模型和机器学习模型对周期时间序列预测效果欠佳的问题, 本文以韩城高速公路交通流量调查数据为数据集, 提出了 SSA-LightGBM 交通流量调查数据预测模型. 对韩城高速数据进行当量计算, 然后对当量数据进行奇异谱分解得到周期项和随机项, 对周期项进行信号重建, 利用 LightGBM 预测随机项, 最后将预测随机项与周期延拓信号进行叠加得到最终的高速当量预测数据. 同时与 XGBoost 和 LightGBM 预测结果作对比, SSA-LightGBM 预测结果与真实值最为贴近, 且 *MAE*、*RMSE* 和 R^2 均优于 XGBoost 和 LightGBM 模型. 该结果对我国高速公路交通调查数据未来的变化趋势预测研究具有很好的指导意义, 可以为我国高速公路的整修和养护提供很好的参考价值.

关键词: 奇异谱分析; LightGBM 模型; 机器学习; 高速交通量调查数据; 预测

引用格式: 徐磊,孙朝云,李伟,杨荣新.基于 SSA-LightGBM 的交通流量调查数据趋势预测.计算机系统应用,2021,30(1):243–249. http://www.c-s-a.org.cn/1003-3254/7750.html

Traffic Volume Survey Data Trend Prediction Based on SSA-LightGBM

XU Lei, SUN Zhao-Yun, LI Wei, YANG Rong-Xin

(School of Information Engineering, Chang'an University, Xi'an 710064, China)

Abstract: In order to solve the problem that the traditional model and the machine learning model have poor performance in predicting the periodic time series, this study proposes the SSA-LightGBM prediction model which takes the Hancheng expressway traffic volume survey data as the data set. First, the Passenger Car Unit (PCU) of Hancheng expressway data is calculated. Second, the singular spectral analysis is applied on the PCU data to obtain periodic and random terms, the periodic term is reconstructed, and then the LightGBM is used to predict the random term. Finally, the predicted random term and the periodic extension signal are superimposed to obtain the prediction data of final expressway PCU. At the same time, compared with the prediction results of XGBoost and LightGBM model, the prediction results of SSA-LightGBM are closest to the true values, and *MAE*, *RMSE* and R^2 are better. This result has a good guiding significance for the research of the forecast on future change trend of our country's expressway volume survey data, and provides a good reference value for the renovation and maintenance of our country's expressways.

Key words: Singular Spectrum Analysis (SSA); LightGBM model; machine learning; traffic volume survey station data; prediction

① 基金项目: 陕西省交通运输厅交通科研项目 (18-22R); 国家重点研发计划 (2018YFB1600202); 高新技术研究培育项目 (300102240201)
Foundation item: Research Project of Transport Department, Shaanxi Province (18-22R); National Key Research and Development Program of China (2018YFB1600202); Cultivation Project of High Technology Research (300102240201)
收稿时间: 2020-06-02; 修改时间: 2020-06-30; 采用时间: 2020-07-10; csa 在线出版时间: 2020-12-31

随着国民经济的发展, 我国机动车数量快速增长, 道路拥堵时常发生, 高速公路作为我国公路的楷模, 其服务水平比一般的道路更好. 交通流量调 (交调) 数据是研究高速公路通行能力的重要参考指标, 为了提高高速公路的服务水平, 对高速公路进行更好的整修和养护, 亟需加强对交调数据的预测研究.

预测研究的关键在于预测模型的建立, 提高预测结果精准度的重要方法便是选择合适有效的预测模型. 序列预测通常分为时间序列预测和多元回归预测, 目前在国内外研究中, 大多是通过传统的神经网络方法对序列进行预测. 例如, Zhou 等人针对传统数学统计方法的局限性等问题建立 BP 神经网络预测模型, 通过将 BP 算法与指数平滑和线性回归方法进行比较, 最后证明了 BP 算法预测结果能为企业制定库存计划提供更好的建议[1]. 成云等人针对现阶段城市道路交通流预测精度不高的局限性, 提出了一种基于差分自回归滑动平均和小波神经网络组合模型的预测方法来进行交通流预测, 最后证明了组合模型可以提高交通流预测精度[2]. Zhang 等人为了更准确的预测交通流量, 实现交通控制和交通拥堵管理, 提出了一种基于 XGboost 和 LightGBM 算法的组合预测模型, 实验结果表明组合模型比单个模型有更高的预测精度[3]. Wu 等人针对 2008–2018 年云南省货运数据及其影响因素建立了灰色神经网络模型, 对云南的货运量进行了预测, 并将其与 BP 神经网络预测结果进行比较最后证明了灰色神经网络具有更好的预测效果[4,5]. Li 等人考虑到植物蒸腾量对温室自动灌溉具有重要的作用, 在 2020 年建立了随机森林回归模型, 通过整合植物和环境参数建立植物蒸腾量预测模型, 该研究为温室种植蔬菜高效生产和智能灌溉提供了科学参考, 为节约水资源也提供了一种有效途径[6].

上述方法中无论是采用单纯的数学模型还是采用机器学习模型对长期的周期性比较强的数据均不能实现精准预测, 只能粗略的显示数据的大致走向, 局部详细的信息表现较差, 因此本文考虑到高速公路交调数据具有很强的周期性等特点提出了基于 SSA-LightGBM 的预测模型.

1 研究数据选取及算法流程

本文针对陕西省高速公路展开研究, 对韩城高速公路进行实地考察调研, 以韩城高速 7~8 月 1344 条交调数据进行研究.

1.1 数据介绍

表 1 为韩城高速交调数据表 (前 5 行数据). 从表中可以看出交调数据一共由观察日期、小时、观测站名称、观测里程、中小客流量、大客车流量、小货车流量、中货车流量、大货车流量、特大货车流量、集装箱流量组成.

表 1 韩城高速交调数据表

观测日期	小时	观测站名称	观察里程	中小客车流量	大客车流量	小货车流量	中货车流量	大货车流量	特大货车流量	集装箱流量
2019.07.01	1	韩城	18	71	0	5	1	4	6	2
2019.07.01	2	韩城	18	37	1	1	1	4	5	1
2019.07.01	3	韩城	18	41	1	4	1	2	7	0
2019.07.01	4	韩城	18	39	0	1	0	0	4	0
2019.07.01	5	韩城	18	56	0	2	0	0	0	0

为了使不同交通工具组成的交通流能够在同样的尺度下进行分析, 使其具有可比性, 在分析计算通行能力和服务水平时, 需要将各类车辆交通量换算成标准机动车当量, 需要用到车辆换算系数, 如表 2 所示, 这里只考虑汽车, 将其分为 3 档: 小型车、中型车、大型车, 其中小型车又分为中小客车和小型货车, 折算系数为 1; 中型车分为大客车和中型货车, 折算系数为 1.5; 大型车分为大型货车、特大货车和集装箱车, 大型货车折算系数为 3, 特大货车和集装箱车折算系数为 4.

表 2 折算系数参考表

车型	一级分类	二级分类	参考折算系数
汽车	小型车	中小客车	1
		小型货车	1
	中型车	大客车	1.5
		中型货车	1.5
	大型车	大型货车	3
		特大货车	4
		集装箱车	4

1.2 数据展示

通过对韩城高速交调数据进行换算, 如图 1 对机动车当量进行可视化, 展示了 2019 年 7 月 1 日–2019 年 8 月 25 日韩城高速公路交调数据当量变化. 从图中可以清晰的看出除了少数天呈现极小或者极大的情况, 其他天数整体呈现周期状态. 当量=小型车×1+中型车×1.5+大型货车×3+特大货车×4+集装箱车×4.

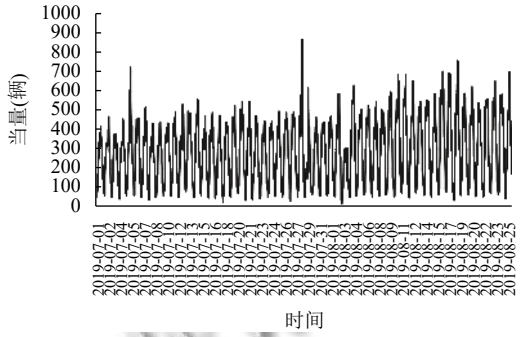

图 1 韩城高速交调当量变化图

1.3 算法流程图

本文的算法流程图如图 2 所示对原始数据进行整合, 计算出高速公路交调数据机动车当量, 然后对原始数据利用奇异普分解方法 (SSA) 进行数据分解, 得到周期信号和随机信号, 再利用机器学习模型 (LightGBM) 对随机项进行预测, 将预测结果和周期延伸信号结合得到最终的预测结果. 另一方面利用 XGBoost、LightGBM 等单独的机器学习方法对当量数据进行预测, 最后将预测结果同 SSA-LightGBM 方法得到的结果进行对比分析, 比较不同模型的预测效果.

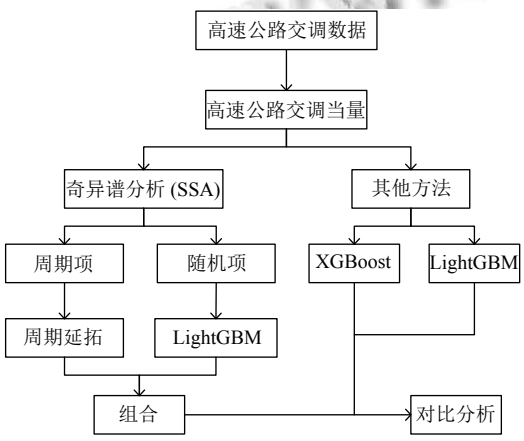

图 2 算法流程图

2 SSA-LightGBM 预测模型原理

2.1 奇异谱分解 (SSA)

奇异谱分解 (Singular Spectrum Analysis, SSA) 是通过分解原序列中的时间主分量, 得到不同层次上的分量序列, 然后将分解出来的低频视为序列变化的长期趋势, 最终得到数据序列的最佳谐波个数, 以此确定时间序列的周期信号.

本文将奇异谱分解分为 4 个步骤, 分别是:

(1) 嵌入. 选择适当的窗口长度 m, 将一维时间序列$Y(T)=((y_1),\cdots,(y_T))$转换为多维时间序列: $X1,\cdots Xn$, $(Xi=(y_i,\cdots,y_{i+m-1}),n=T-m+1)$, 得到轨迹矩阵, 即:

$$X=[X1,\cdots,Xn]=\left(x_{ij}\right)_{i,j=1}^{n,m}=\begin{pmatrix} y_1 & y_2 & \cdots & y_m \\ y_2 & y_3 & \cdots & y_{m+1} \\ \vdots & \vdots & \cdots & \vdots \\ y_n & y_{n+1} & \cdots & y_T \end{pmatrix}.$$

(2) 奇异值分解. 通过对矩阵 SVD 分解, 得到 XX^{T} 的 L 个特征值. 具体步骤是: 将 X 转置得到 X^{T}, 然后利用 XX^{T} 得到方阵, 再利用方阵的性质求得矩阵的特征值[7–9], 即利用$(X^{\mathrm{T}}X)v_i=\lambda_i v_i$, 求得$\sigma_i=\sqrt{\lambda_i}$, $\mu_i=\frac{1}{\sigma_i}Xv_i$. σ即是奇异值, μ是奇异向量.

(3) 分组. 假设有 N 个奇异值$\sigma_1,\sigma_2,\cdots,\sigma_N$, 定义第 i 个奇异值的方差贡献率为$\sigma_i\Big/\sum_{k=1}^{N}\sigma_k=p(i)$, 由大到小选择前 M 个奇异值, 使其方差贡献率之和大于一定阈值.

(4) 确定最佳谐波个数. 利用前 M 个奇异值的方差贡献率之和大于一定阈值 (0.85) 来确定最佳谐波个数 M. 当 $P(i)\geq 0.85$ 时, 我们认为此时的 i 即为最佳的谐波个数, 即 $M=i$. 然后利用三角函数的性质将数据构造成周期函数.

2.2 周期项和随机项的确定

设$\{y_t,t=1,2,\cdots,N\}$为时间序列, 若将其看作由周期项和随机项组成, 可以用组合模式$y_t=p_t+x_t$描述, 其中p_t为周期项, x_t为随机项. p_t的表达式如下:

$$P_t=a_0/2+\sum_{k=1}^{M}(a_k\cos(2\pi kt/N)+b_k\sin(2\pi kt/N))$$

其中, M 为谐波个数, $a_k=(2/N)\sum_{t=1}^{N}y_t\cos(2\pi kt/N),k=0,1,\cdots,M$, $b_k=(2/N)\sum_{t=1}^{N}y_t\sin(2\pi kt/N),k=0,1,\cdots,M$

2.3 LightGBM 模型

LightGBM 是一种梯度 Boosting 框架, 使用基于决

策树的学习算法, 具有快速、高效、支持并行化学习、可以处理大规模数据等优点.

(1) 梯度提升.

梯度提升是在不断的迭代过程中, 对模型不停的增加子模型, 同时保证最终的损失函数值在不断的下降. GBDT 是一种梯度提升决策树, 是由多个决策树组成, 利用最速下降的近似方法, 即利用损失函数的负梯度在当前模型的值作为我们回归提升树算法的残差的近似值, 来拟合一个回归树[10–12].

假设我们每一个单独的子模型为 $f_i(x)$, 我们的复合模型为:

$$F_m(x)=\partial_0 f_0(x)+\partial_1 f_1(x)+\cdots+\partial_m f_m(x)$$

损失函数为 $L[F_m(x),Y]$, 每一次对模型中添加新的子模型后, 使得我们的损失函数不断朝着 0 发展.

(2) LightGBM 原理

LightGBM 是在传统的梯度提升树 (GBDT) 上使用直方图优化 (Histogram) 算法, 先把连续的特征值离散化成 k 个整数, 同时构造一个宽度为 k 的直方图. 在遍历数据的时候, 根据离散化后的值作为索引在直方图中累积统计量, 当遍历一次数据后, 直方图累积了需要的统计量, 然后根据直方图的离散值, 遍历寻找最优的分割点. 同时使用带深度限制的 Leaf-wise 的叶子生长策略, 经过一次数据可以同时分裂同一层的叶子, 容易进行多线程优化, 也好控制模型复杂度, 不容易过拟合[13–15]. 表 3 给出 LightGBM 模型主要参数含义.

表 3 LightGBM 模型参数含义

模型参数	含义
learning_rate	学习率
num_leaves	叶子数量, 默认为31
max_bin	feature将存入的bin的最大数量
n_estimators	迭代次数
bagging_fraction	每次迭代时用的数据比例
objective	模型所选回归任务
feature_fraction	每次迭代中随机选择一定比例的参数来建树

3 SSA-LightGBM 模型预测分析

3.1 对数据进行奇异谱分解

此处将 2019 年的 7 月 1 日到 2019 年 8 月 25 日的 1344 条数据分为两部分, 利用前 70% 的 941 条数据预测后 30% 的 403 条数据并将其与对应的真实数据进行对比. 对前部分数据进行奇异谱分析时选取的窗口长度为 200, 可以得到方差贡献率图 (见图 3). 从图中可以看出, 当 i≥56 时, $P(i)$≥85%, 满足本文的数据处理要求, 即当显著谐波个数 M 为 56 时, 对应个三角函数信号能表征序列的最主要趋势.

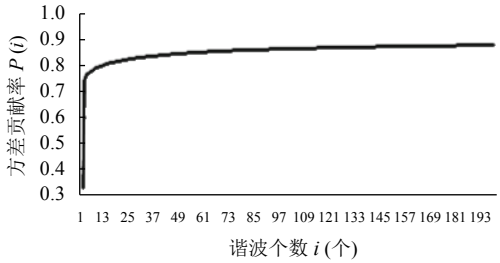

图 3 方差贡献率图

然后, 可构造周期函数, 得到图 4. 周期函数对应的公式为:

$$P_t=a_0/2+\sum_{k=1}^{M}(a_k\cos(2\pi kt/N)+b_k\sin(2\pi kt/N))$$

式中, M 为谐波个数 56, $a_k=(2/N)\sum_{t=1}^{N}y_t\cos(2\pi kt/N),k=0,1,\cdots,M$, $b_k=(2/N)\sum_{t=1}^{N}y_t\sin(2\pi kt/N),k=0,1,\cdots,M$

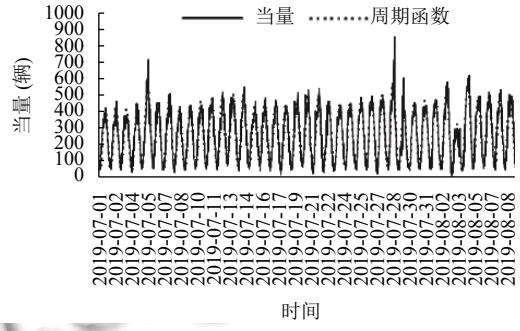

图 4 周期函数对比图

表 4 详细介绍了周期函数 Pt 中每一个参数的具体含义和取值. 根据周期函数的性质, 利用周期不变性对周期函数进行延伸得到 2019 年 8 月 9 日到 8 月 25 日周期函数延拓图 (图 5). 然后由交调当量原始数据减去周期项 ($x_t=y_t-p_t$), 可得到图 6 所示的随机项.

表 4 周期函数各参数含义

参数	意义
P_t	t时刻周期函数对应的函数值, 即t时刻对应的周期交调当量值
a_0	周期函数初始值, 此处为588.7374
a_k/b_k	不同谐波个数对应的正余弦函数的系数
y_t	t时刻的交调当量数值
M	周期函数的谐波个数, 其值等于方差贡献率大于0.85时刻的i值此处为56
N	当量序列的长度, 此处为1344

本文采用 LightGBM 机器学习方法对交调当量随机项数据进行预测, 表 5 为 LightGBM 模型常用参数的取值, 其他参数均采用默认. 图 7 中曲实线为交调当量随机项前 941 条数据, 后面虚线为使用 LightGBM 预测的 403 条随机当量数据.

4 预测效果评价

我们将 SSA-LightGBM 模型得到的 2019 年 8 月 9 日到 8 月 25 日周期函数延拓结果和随机项预测值相叠加得到最终交调当量预测值 (见图 8, 红色曲线). 为了验证 SSA-LightGBM 模型的预测效果, 本文还分别用 XGBoost 和 LightGBM 机器学习方法对数据进行了预测 (利用 2019 年 7 月 1 日到 8 月 8 日的数据预测 2019 年 8 月 9 日到 8 月 25 日的数据). 由于这 2 种模型为单纯机器学习预测方法, 且不是本文的研究对象, 故此处不详细描述各种方法的预测过程. 将不同预测模型得到的预测结果与韩城交调当量的原数据进行对比, 得到图 8. 图中黑色曲线为韩城高速交调数据的原数据, 红色曲线为 SSA-LightGBM 模型的预测结果, XGBoost 和 LightGBM 模型的预测结果分别对应蓝色和绿色虚线. 从图中可以发现, 单独使用 XGBoost 和 LightGBM 模型得到的曲线整趋势能表示原始数据, 但是整体均低于原始数据不能完整的表示当量的局部特征. 只有橙色虚线 SSA-LightGBM 模型预测的结果能够实现高精度稳定预测 2019 年 8 月 9 日到 8 月 25 日交调当量变化趋势.

另外, 本文还采用常用的平均绝对值误差 (*MAE*)、均方根误差 (*RMSE*) 和相关系数 (R^2) 分析评价了不同模型预测结果的精度 (表 6), 各指标的公式、含义及评价标准见表 7.从表中可以发现本文的模型 *MAE* 和 RMSE 均低于 XGBoost 和 LightGBM 模型说明模型的稳定性和平均误差优于单独的机器学习模型, R^2 大于另外两个模型说明预测效果好相关性强.

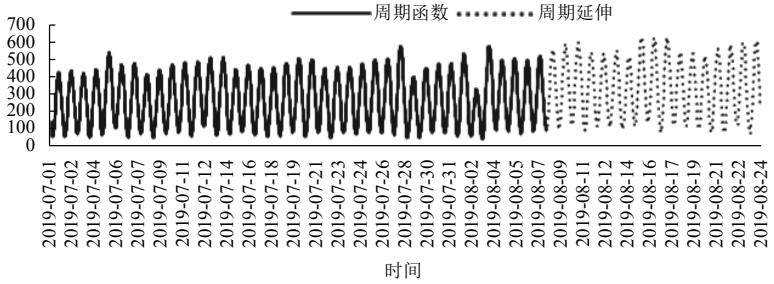

图 5 周期函数延拓图

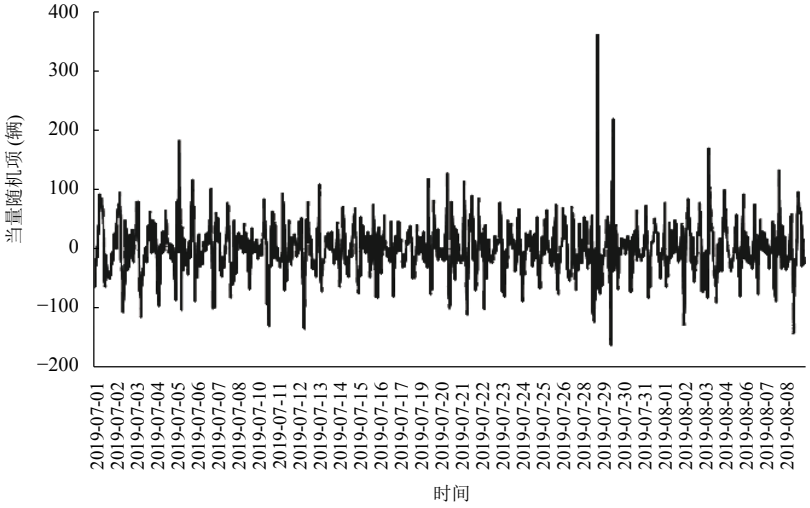

图 6 交调当量随机项

表 5 LightGBM 模型参数取值

模型参数	learning_rate	num_leaves	max_>bin	n_estimators	bagging_fraction	objective	feature_fraction
取值	0.2	3	50	700	0.9	regression	0.25

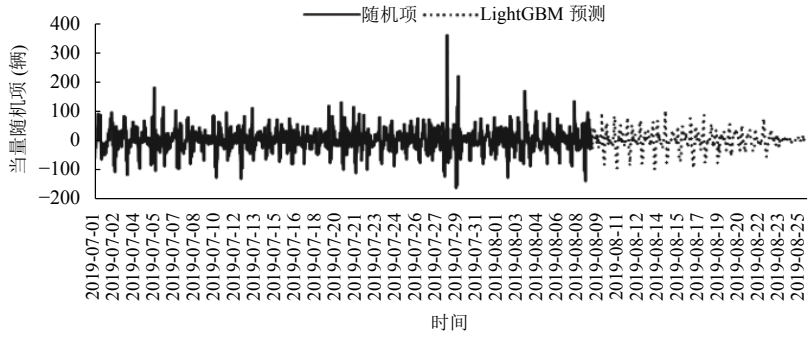

图 7 随机项预测结果

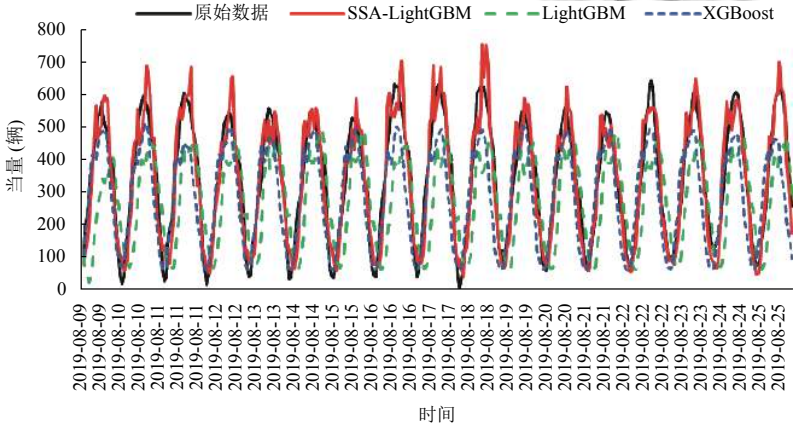

图 8 不同模型预测结果对比图

表 6 各模型评价指数

评价指标/模型	SSA-LightGBM	XGBoost	LightGBM
MAE	49.49	81.0685	122.4568
RMSE	60.25	99.0780	143.1784
R^2	0.8875	0.6958	0.3647

表 7 评价指标

模型评价指标	公式	意义
MAE (平均绝对误差)	$MAE=\frac{1}{n}\sum_{i=1}^{n}\lvert x_i-x_p\rvert$	计算值与真实值之间的平均绝对误差, *MAE*越接近0, 模型越准确
RMSE (均方根误差)	$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-x_p)^2}$	计算值与真实值之间的均方根误差,*RMSE*越接近0, 模型越准确
R^2 (线性相关系数)	$R^2=1-\frac{RSS}{TSS}=1-\frac{\sum_{i}^{n}(x_i-x_p)^2}{\sum_{i}^{n}(x_i-\bar{x}_i)^2}$	因变量与自变量之间的线性相关系数, R^2越接近1, 模型越准确

综上, 无论是直观的曲线对比 (图 8), 还是数学统计方法 (表 6), 均证明 SSA-LightGBM 模型可以高精度、高稳定性地预测 2019 年 8 月韩城高速交调数据当量变化趋势. 我们可以使用该模型对全国的高速公路交调数据进行预测, 这对我国高速公路更好的休整和养护以及服务水平的提升具有重要的参考价值.

5 结论

由于传统的时间序列预测模型和单独的机器学习模型对周期性比较强的时间序列预测存在一定弊端, 本文采用 SSA-LightGBM 模型进行预测, 以韩城高速 2019 年 7 月到 8 月两个月 1344 条交调数据为例, 以前 70% 的数据作为模型的训练集, 以后 30% 的数据作为验证集. 将结果与单纯的机器学习模型 XGBoost 和 LightGBM 模型进行对比, 发现本文提出的模型预测结果更接近真实值, 同时该模型的 *MAE*、R^2 和 *RMSE* 均优于其他两个模型, 表明本文的模型可以很好地预测韩城高速交调当量数据, 这对该地区高速公路的整修扩建、养护具有一定的指导意义.

参考文献

1 Zhou H, Ding DY, Li YL, *et al*. Research on inventory demand forecast based on BP neural network. Information Technology, 2016, 40(11): 38–41.

2 成云, 成孝刚, 谈苗苗, 等. 基于 ARIMA 和小波神经网络组合模型的交通流预测. 计算机技术与发展, 2017, 27(1): 169–172.

3 Zhang M, Fei X, Liu ZH. Short-term traffic flow prediction based on combination model of XGboost-LightGBM. Proceedings of 2018 International Conference on Sensor Networks and Signal Processing. Xi'an, China. 2018. 322–327.

4 Wu RL, Xu WH, Shen WZ. Research on forecast model of freight in Yunnan Province based on gray neural network. Logistics Sci-Tech, 2019, 42(8): 13–16, 19.

5 Yang ZY. Rural logistics demand forecast based on gray neural network model. Proceedings of 2018 4th World Conference on Control, Electronics and Computer Engineering. Jinan, China. 2018. 322–327.

6 Li L, Chen SW, Yang CF, *et al*. Prediction of plant transpiration from environmental parameters and relative leaf area index using the random forest regression algorithm. Journal of Cleaner Production, 2020, 261: 121136. [doi: 10.1016/j.jclepro.2020.121136]

7 张清亮. 基于奇异值分解与深度递归神经网络的齿轮剩余寿命预测方法研究 [硕士学位论文]. 重庆: 重庆大学, 2018.

8 Ibrahim SJA, Thangamani M. Enhanced singular value decomposition for prediction of drugs and diseases with hepatocellular carcinoma based on multi-source bat algorithm based random walk. Measurement, 2019, 141: 176–183. [doi: 10.1016/j.measurement.2019.02.056]

9 Montesinos-López OA, Montesinos-López A, Crossa J, *et al*. A singular value decomposition Bayesian multiple-trait and multiple-environment genomic model. Heredity, 2019, 122(4): 381–401. [doi: 10.1038/s41437-018-0109-7]

10 Zhang SH, Dong XG, Xing Y, *et al*. Analysis of influencing factors of transmission line loss based on GBDT algorithm. Proceedings of 2019 International Conference on Communications, Information System and Computer Engineering. Haikou, China. 2019. 179–182.

11 张潇, 韦增欣, 杨天山. GBDT 组合模型在股票预测中的应用. 海南师范大学学报 (自然科学版), 2018, 31(1): 73–80.

12 余登武, 罗永平. 基于 GBDT-Ridge 的短期电力负荷预测. 新型工业化, 2019, 9(6): 23–26, 33.

13 Zhang J, Mucs D, Norinder U, *et al*. LightGBM: An effective and scalable algorithm for prediction of chemical toxicity—Application to the Tox21 and mutagenicity data sets. Journal of Chemical Information and Modeling, 2019, 59(10): 4150–4158. [doi: 10.1021/acs.jcim.9b00633]

14 叶志宇, 冯爱民, 高航. 基于深度 LightGBM 集成学习模型的谷歌商店顾客购买力预测. 计算机应用, 2019, 39(12): 3434–3439.

15 张振, 曾献辉. 基于 CNN-LightGBM 模型的高速公路交通量预测. 信息技术与网络安全, 2020, 39(2): 34–39.