文章编号：2095-2163(2022)12-0175-05 中图分类号：TP18 文献标志码：A

基于自适应遗传算法的随机森林模型参数优化方法

蔡 明[1,2]，孙 杰[1,2]，杨维发[1,2]，鲍 清[1]，李培德[1]
(1 湖北省气象信息与技术保障中心，武汉 430074；2 中国气象局 武汉暴雨研究所 暴雨监测预警湖北重点实验室，武汉 430074)

摘 要：随机森林模型多采用网格搜索的参数优化方法，存在搜索间隔固定、搜索效率低下的问题。为了克服以上缺陷，提出一种基于自适应遗传算法的随机森林模型参数优化方法，通过动态调节遗传操作的交叉、变异概率，在尽可能多保留优势粒子的同时更有效地产生新优势粒子，达到跳出局部最优并快速到达全局最优点的目的。利用提出的参数优化方法对随机森林算法中的决策树数目、最大树深度进行参数优化。使用 Boston house price 数据集仿真的结果表明，使用该参数优化方法优化后的随机森林模型的回归预测效果得到一定提高。

关键词：随机森林回归；自适应遗传算法；遗传操作；参数优化

Parameter optimization of Random Forest model based on adaptive genetic algorithm

CAI Ming[1,2], SUN Jie [1,2], YANG Weifa[1,2], BAO Qing[1], LI Peide[1]
(1 Hubei Meteorological Information and Technical Support Center, Wuhan 430074, China;
2 Hubei Key Laboratory for Heavy Rain Monitoring and Warning Research, Wuhan Institute of Heavy Rain, China Meteorological Administration, Wuhan 430074, China)

【Abstract】 Random Forest models mostly adopt the parameter optimization method of grid search, which has the disadvantages of fixed search interval and low search efficiency. In order to overcome the above defects, a parameter optimization method of Random Forest model based on adaptive genetic algorithm is proposed. By dynamically adjusting the crossover and mutation probability of genetic operation, the research can retain as many dominant particles as possible and generate new dominant particles more effectively, so as to jump out of the local optimization and quickly reach the global optimization. The proposed parameter optimization method is used to optimize the number of decision trees and the max depth of trees attributes in the Random Forest algorithm. The simulation results using Boston house price data set show that the regression prediction effect of the Random Forest model optimized by this parameter optimization method is improved to a certain extent.

【Key words】 Random Forest regression; adaptive genetic algorithm; genetic manipulation; parameter optimization

0 引 言

随机森林回归(Random Forest Regression)算法作为一种灵活且易于使用的机器学习算法[1-2]，其理论和方法已被作为一种替代一般线性模型(线性回归、方差分析等)和广义线性模型(逻辑斯蒂回归、泊松回归等)的方法，广泛应用于工程应用和科学领域中复杂问题的解决上。国内外学者对随机森林在回归和分类问题中的应用进行了全面研究。在国外，Kulkarni 等人[3-4]为了提高分类正确率，将决策树维度分为 2 部分。Oshiro 等人[5]证明了在随机森林性能达到最优时决策树数目存在临界值。Bernard 等人[6]研究了随机森林强度与相关性的关系。在国内，袁远等人[7]利用随机森林算法对非线性数据特征的学习能力，优化 ARIMA 模型预测残差，最终达到提高回归预测精度的目的。马景义等人[8]综合了 Adaboost 算法和随机森林算法的优势，提出了拟自适应分类随机森林算法。冯开平等人[9]将加权 K 最近邻法(KNN)与随机森林算法结合应用于表情识别，简化了计算复杂度的同时取得了不错的识别率。

自适应遗传算法是将生物进化论的自然选择和遗传机理应用于粒子滤波算法以克服其粒子多样性退化不足的一种随机化搜索方法[10-11]。其主要特点是按照优势种群遗传的原则将粒子适应度变化情况作为遗传操作中交叉和变异概率变化的依据，通

基金项目：湖北省气象局重点项目(2022Z04)；湖北省气象局年轻科研人员专项(2020Q07)。
作者简介：蔡 明(1987-)，男，硕士，工程师，主要研究方向：气象装备保障、气象信息技术；孙 杰(1981-)，男，硕士，高级工程师，主要研究方向：气象装备保障、气象信息技术；杨维发(1985-)，男，硕士，高级工程师，主要研究方向：气象装备保障、气象信息技术；鲍 清(1981-)，男，本科，助理工程师，主要研究方向：气象装备保障；李培德(1991-)，男，硕士，工程师，主要研究方向：气象装备保障。
通讯作者：孙 杰 Email：3037998@qq.com
收稿日期：2022-03-20

过对粒子的选择、交叉和变异操作模拟生物界优胜劣汰、适者生存的过程,由于其直接对结构化的对象进行操作,故具有很好的全局寻优能力。但是由于遗传操作中的交叉和变异概率是预先设定的,参数选取不当容易使算法陷入局部最优[12-15]。

基于以上研究论述,本文提出一种基于自适应遗传算法的随机森林回归模型参数优化方法,使用 Boston house price 数据集对经过该方法优化后随机森林模型的回归预测效果进行验证。

1 相关算法介绍

1.1 随机森林回归算法

随机森林回归(Random Forest Regression,RFR)算法是一种基于决策树(Decision Tree)的引入随机特征选择的 Bagging 类集成算法,目前被广泛应用于各类回归问题。本文使用 Boston house price 数据集对随机森林回归模型进行训练和预测。随机森林回归模型的建立过程如下:

(1)从原始训练集中使用 bootstrap 方法随机有放回采样取出 m 个样本,共进行 n_tree 次采样。生成 n_tree 个训练集。

(2)对 n_tree 个训练集,分别独立训练 n_tree 个决策树模型。

(3)对于单个决策树模型,假设训练样本特征个数为 n,选择最好的特征进行切分。

(4)每棵树都按照步骤(3)来切分下去,直到该节点的所有训练样例都属于同一类。在决策树的切分过程中不需要剪枝。

(5)将生成的多棵决策树组成随机森林,模型最终预测结果为随机森林中多棵决策树预测结果的均值。

决策树的生长过程就是使用满足划分准则的特征不断将数据集划分为纯度更高、不确定性更小的子集的过程。

在步骤(3)中,当训练决策树模型时需要考虑怎样选择切分特征、切分点以及怎样衡量切分特征、切分点的好坏。针对切分特征和切分点的选择,本文采用穷举法,即遍历每个特征和每个特征的所有取值,再从中找出最好的切分变量和切分点;针对于切分特征和切分点的好坏,一般以切分后节点的不纯度来衡量,即各个子节点不纯度的加权和 $G(x_i, v_{ij})$,其计算公式如下:

$$G(x_i, v_{ij}) = \frac{n_{left}}{N_s} H(X_{left}) + \frac{n_{right}}{N_s} H(X_{right}) \tag{1}$$

其中,x_i 为节点的某一个切分特征;v_{ij} 为切分特征的一个切分值;n_{left}、n_{right}、N_s 分别为切分后左子节点训练样本个数、右子节点训练样本个数以及当前节点所有训练样本个数;X_{left}、X_{right} 分为左、右子节点的训练样本集合;$H(X)$ 为节点的不纯度函数(impurity function),回归模型一般采用均方误差(Mean Square Error, *MSE*)或平均绝对误差(Mean Absolute Error, *MAE*)作为不纯度函数,本文则选用了 *MSE* 作为模型的不纯度函数,其数学定义公式见式(2):

$$H(X_s) = \frac{1}{n_s} \sum_{i=1}^{n_s} (y - \bar{y}_s)^2 \tag{2}$$

其中,X_s 为当前节点训练样本集合;n_s 为当前节点训练样本数目;$\bar{y}_s$ 为当前节点样本目标特征的均值。

将式(2)带入式(1)后,对于任意切分点可以得到:

$$G(x,v) = \frac{1}{N_s} \left(\left(\sum_{i=1}^{n_{left}} (y_i - \bar{y}_{left}) \right)^2 + \left(\sum_{i=1}^{n_{right}} (y_i - \bar{y}_{right}) \right)^2 \right) \tag{3}$$

1.2 自适应遗传算法

以往的遗传算法常使用恒定不变的概率对粒子进行交叉和变异等遗传操作,这样会导致粒子群中适应度较大的优势粒子容易被丢弃掉,同时新的优势粒子也不容易产生,致使算法一旦陷入局部最优,就很难跳出。

针对这一问题,提出一种基于生物遗传进化思想的自适应遗传算法(AGA)。算法中,高适应度的优势个体以较高概率进行交叉操作,这样可以增大优势基因遗传到子代的可能性,更符合遗传进化规律;低适应度的个体以较高的概率进行变异操作,这样就更容易通过变异操作产生新的优势个体,避免算法陷入局部最优。通过自适应地调节遗传操作中的交叉、变异概率,从而避免遗传算法中早熟现象的出现。其中,遗传操作的交叉概率 P_c 和变异概率 P_m 可以分别表示为:

$$P_c = \begin{cases} P_{c1} - \dfrac{(P_{c1} - P_{c2})(\bar{f} - f')}{(f_{\max} - \bar{f})} & f' \leqslant \bar{f} \\ P_{c1} & f' > \bar{f} \end{cases} \tag{4}$$

$$P_m = \begin{cases} \dfrac{P_{m1}(f_{\max} - f)}{(f_{\max} - \bar{f})} & f \geqslant \bar{f} \\ P_{m2} & f < \bar{f} \end{cases} \tag{5}$$

1.3　基于自适应遗传算法的随机森林回归参数优化

以往的随机森林回归算法的参数优化多通过绘制学习曲线或网格搜索交叉验证的方法实现，实施过程中恒定不变的搜索步长使得最优参数的获取很难在速度和效果上同时达到最优。基于此，提出自适应遗传算法辅助下的随机森林回归模型参数优化方法，利用遗传算法优异的全局寻优能力，结合自适应方法动态调整的遗传操作概率，达到快速取得全局最优解的目的。

随机森林回归是基于 bagging 框架的决策树模型，因此随机森林回归模型的参数调整包括 2 部分：随机森林框架的参数调优和决策树的参数调优。使用自适应遗传算法进行随机森林回归模型参数优化的流程如图 1 所示。

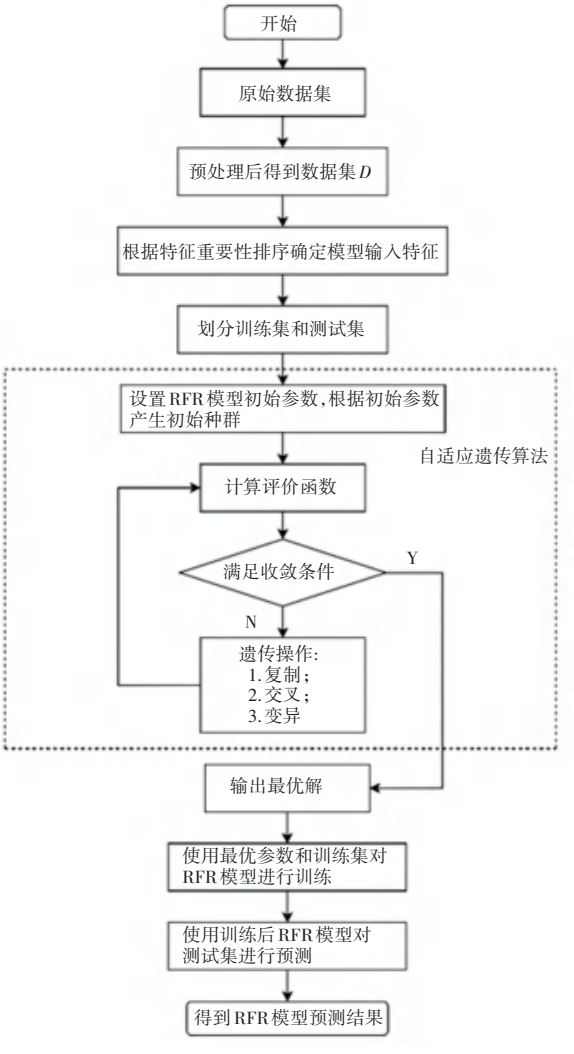

图 1　基于自适应遗传算法的随机森林回归模型流程图

Fig. 1　Flow chart of Random Forest regression model based on adaptive genetic algorithm

2　实验准备

2.1　数据集准备

为了验证经自适应遗传算法优化后的随机森林回归模型的有效性，使用 Kaggle Boston house price 数据集进行仿真验证。数据集中的每一行数据都是对波士顿周边或城镇房价的情况描述，数据集共有 14 个特征，分别为：城镇人均犯罪率（*CRIM*）、住宅用地所占比例（*ZN*）、城镇中非住宅用地所占比例（*INDUS*）、虚拟变量（*CHAS*），用于回归分析；环保指数（*NOX*）、每栋住宅的房间数（*RM*）、1940 年以前建成的自住单位的比例（*AGE*）、距离 5 个波士顿就业中心的加权距离（*DIS*）、距离高速公路的便利指数（*RAD*）、每一万美元的不动产税率（*TAX*）、城镇中的教师 / 学生比例（*PTRATIO*）、城镇中的黑人比例（*B*）、地区中有多少房东属于低收入人群（*LSTAT*）、自住房屋房价（*PRICE*）。其中，*PRICE* 为目标变量，其他 13 个特征为模型的输入自变量特征。各自变量特征的重要性见表 1。由表 1 可以发现，不论是 RFR 模型、还是自适应遗传算法优化后的 AGA - RFR 算法模型，各自变量的重要性程度都是相近的，且 *RM* 和 *LASTAT* 都是对模型最重要的变量。

模型初始特征集中各项特征之间的相关性热力图如图 2 所示。图 2 中，部分特征间呈现负相关性，部分呈现正相关性。将 Kaggle Boston house price 数据集按照 7 ∶ 3 的比列划分为训练集和测试集。

2.2　评价指标

为了对比自适应遗传算法参数优化方法的应用对随机森林回归模型预测精度的影响，需要对随机森林回归模型的预测精度进行评价。本文使用均方根误差、决定系数和平均绝对误差这 3 个指标对模型的预测精度进行评价。对此拟给出研究分述如下。

（1）均方根误差（Root Mean Squared Error，*RMSE*），也叫回归系统的拟合标准差。由于均方根误差对一组测量值中的特大或特小误差反映非常敏感，所以，均方根误差能够很好地反映出测量的精密度。具体数学公式可写为：

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_o^i - y_e^i)^2} \qquad (6)$$

（2）决定系数（Coefficient of Determination，R^2）。表示对模型进行线性回归后，评价回归模型系数的拟合优度。R^2 反映了模型因变量的全部变异能通过回归模型被自变量解释的比例。R^2 越大，线性回归模型解释的变异越大。具体数学公式可写为：

表 1 模型特征变量重要性

Tab. 1 Importance of model characteristic variables

特征变量	*CRIM*	*ZN*	*INDUS*	*CHAS*	*NOX*	*RM*	*AGE*	*DIS*	*RAD*	*TAX*	*PTRATIO*	*B*	*LSTAT*
RFR	0.055 6	0.000 6	0.005 9	0.002 4	0.013 8	0.486 0	0.012 8	0.063 3	0.002 8	0.019 0	0.013 7	0.010 4	0.313 6
AGA-RFR	0.056 2	0.000 8	0.006 2	0.002 4	0.014 0	0.487 5	0.013 5	0.057 3	0.003 1	0.018 7	0.013 4	0.011 2	0.315 8

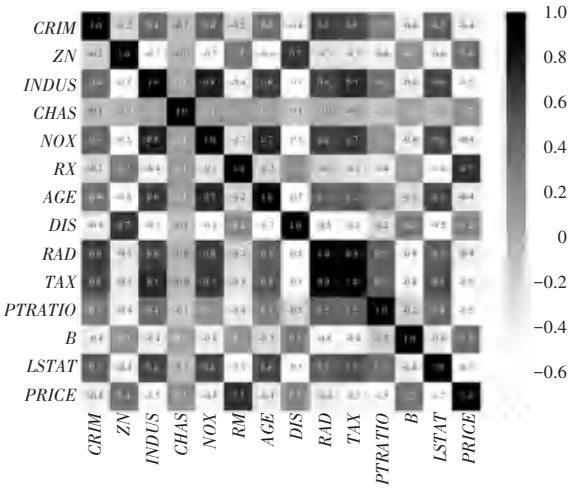

图 2 模型特征相关性热力图

Fig. 2 Model characteristic correlation thermodynamic diagram

$$R^2 = \left(\frac{\sum_{i=1}^{n} ((y_e^i - \bar{y_e}) \cdot (y_o^i - \bar{y_o}))}{\sqrt{\sum_{i=1}^{n} (y_e^i - \bar{y_e})^2} \cdot \sqrt{\sum_{i=1}^{n} (y_o^i - \bar{y_o})^2}} \right)^2 \tag{7}$$

R^2 为 1 时,表明模型预测值和真实值观测值没有任何误差,表示回归分析中自变量对因变量的解释越好; R^2 为 0 时,模型中样本的每项预测值都等于均值; R^2 接近于 0 时,表明模型预测能力差,预测效果接近于“使用观测值的平均值作为模型预测值”。这就表示可能用了错误模型,或者模型假设不合理。

(3)平均绝对误差(Mean Absolute Error, *MAE*)计算公式如下:

$$MAE = \frac{1}{n} \cdot \sum_{i=1}^{n} |y_o^i - y_e^i| \tag{8}$$

其中, *MAE* 的取值范围为 [0, +¥), 当预测值与真实值完全吻合时等于 0,即完美模型;误差越大,该值越大。

3 结果与分析

研究使用 Kaggle Boston house price 训练数据集对经过自适应遗传算法优化得到的随机森林回归模型进行训练,训练后的模型对测试集进行预测。对比未经过参数优化的 RFR 模型与经过参数优化的 AGA-RFR模型的预测结果,预测效果对比见表 2。

表 2 模型预测精度对比

Tab. 2 Comparison of prediction accuracy of models

要素	*RMSE*	R^2	*MAE*
RFR	4.174	0.833	2.515
AGA-RFR	4.111	0.868	2.503

观察表 2 可以发现,经过自适应遗传算法优化参数后的 AGA-RFR 模型的回归预测结果中, *RMSE* 为 4.111,优于 RFR 的 4.174;AGA-RFR 的 R^2 为0.868,同样优于 RFR 的 0.833;对比 2 种模型的 *MAE* 也是同样的情况。综上可知,经过参数优化后的 AGA-RFR 模型的 *MAE* 要优于 RFR 模型。这就说明通过使用自适应遗传算法对随机森林回归模型的参数进行优化,使得随机森林回归模型的预测效果得到了提高。

以 *Prices* 为横坐标, *Predicted prices* 为纵坐标,绘制出的模型预测价格与实际价格的对比结果图如图 3 所示。

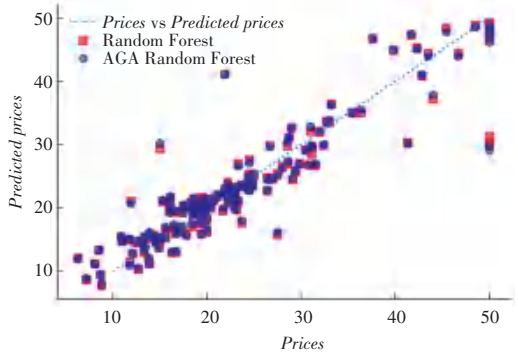

图 3 模型预测价格与实际价格对比图

Fig. 3 Comparison between model predicted prices and actual prices

由图 3 可知,相比于方形所代表的 RFR 模型预测结果,圆形所代表的 AGA-RFR 模型的预测结果总体上更接近于代表模型预测价格与实际价格相等的虚直线。由此说明,AGA-RFR 模型的预测结果比 RFR 模型的预测结果更接近于真实价格。

模型预测值残差与实际价格对比如图 4 所示。图 4 中,相比于方形所代表的 RFR 模型预测结果,圆形所代表的 AGA-RFR 模型的预测结果总体上更接近于代表预测残差为 0 的虚直线。这也说明, AGA-RFR 模型的预测结果比 RFR 模型的预测结果具有更小的预测残差。

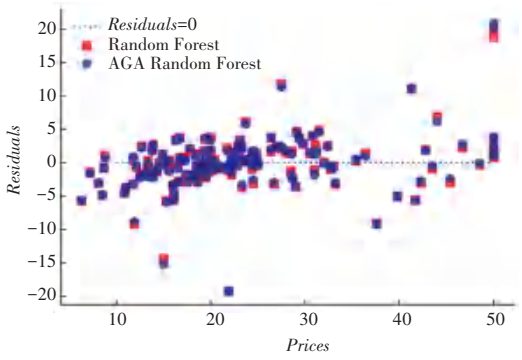

图 4　模型预测值残差与实际价格对比图

Fig. 4　Comparison between residual error of model predicted value and actual prices

4　结束语

本文提出一种用于随机森林回归模型参数优化的方法，利用自适应遗传算法在求解全局最优解的研究时不易陷入局部最优的优势，通过对粒子的选择、交叉和变异操作模拟生物界优胜劣汰、适者生存的过程。通过使用 Boston house price 数据集对经过该方法优化后随机森林模型的回归预测效果进行验证，试验结果表明，经过该方法参数优化后的 AGA-RFR 模型的回归预测效果要优于未经过参数优化的 RFR 模型的预测效果。

本文提出的基于自适应遗传算法的随机森林模型参数优化方法可以作为随机森林回归模型参数优化的一种有效手段。

参考文献

[1] 任利强，张立民，王海鹏，等. 基于优化聚类的 IXGBoost 短期电力负荷预测[J]. 计算机与数字工程，2020，48(04)：741-747.

[2] 张克锐，李庆领，王传伟，等. 基于遗传算法的高速列车头型多目标优化[J]. 计算机与数字工程，2021，49(07)：1330-1336.

[3] KULKARNI V Y，PRADEEP K S. Efficient learning of random forest classifier using disjoint partitioning approach [C]// Proceeding of the Word Congress on Engineering. London：IAENG，2013：1-5.

[4] KULKARNI V Y，PRADEEP K S. Random forest classifiers：A survey and future research directions[J]. International Journal of Advanced Computing，2011，36(1)：1144-1153.

[5] OSHIRO T M，PEREZ P S，BARANAUSKAS J A. How many trees in a random forest[M]//PERNER P. Machine Learning and Data Mining in Pattern Recognition. MLDM 2012. Lecture Notes in Computer Science ( ). Berlin/ Heidelberg：Springer，2012，7376：154-168.

[6] BERNARD S，HEUTTE L，ADAM S. Towards a better understanding of random forests through the study of strength and correlation[C]//HUANG D S，JO K H，LEE H H，et al. Emerging Intelligent Computing Technology and Applications. With Aspects of Artificial Intelligence. ICIC 2009. Lecture Notes in Computer Science ( ). Berlin/Heidelberg：Springer，2009，5755：536-545.

[7] 袁远，郭天添. ARIMA-RF 组合模型的销售预测研究[J]. 软件导刊，2021，20(09)：33-38.

[8] 马景义，吴喜之，谢邦昌. 拟自适应分类随机森林算法[J].数理统计与管理，2010，29(05)：805-811.

[9] 冯开平，赖思渊. 基于加权 KNN 与随机森林的表情识别方法[J]. 软件导刊，2018，17(11)：30-33.

[10] 张馨予，孙宏宇，逯洋，等. 遗传算法优化的支持向量机回归计算老龄化人口方法[J]. 智能计算机与应用，2021，11(08)：143-145，150.

[11] 祁翔，张心光. 基于遗传算法优化 BP 神经网络的预测建模[J]. 智能计算机与应用，2021，11(05)：160-162，169.

[12] 刘鹏程，李新利. 基于多种群遗传算法的含分布式电源的配电网故障区段定位算法[J]. 电力系统保护与控制，2016，44(02)：36-41.

[13] 唐国新，陈雄. 基于改进遗传算法的机器人路径规划[J]. 计算机工程与设计，2007，28(18)：4446-4449.

[14] 颜晓娟，龚仁喜，张千锋. 优化遗传算法寻优的 SVM 在短期风速预测中的应用[J]. 电力系统保护与控制，2016，44(09)：38-42.

[15] 郭宁明，杨飞，覃剑，等. 基于遗传算法及信号谱分析的电网故障定位方法[J]. 电力系统自动化，2016，40(15)：79-85.

(上接第 174 页)

[12] 侯贺平，王靓，任婉倩，等. 基于数字足迹的河南省 A 级景区旅游流网络特征研究[J]. 地域研究与开发，2022，41(01)：91-97.

[13] 王海江，苏景轩，苗长虹，等. 中国中心城市旅游出行的空间分布规律与结构图谱研究[J]. 地理科学，2021，41(11)：1907-1916.

[14] 陈倩，邓敏. 世界自然遗产景区网络关注度与客流量相关性—以四川省四姑娘山为例[J]. 乐山师范学院学报，2020，35(06)：47-52.

[15] 孙烨，张宏磊，刘培学，等. 基于旅游者网络关注度的旅游景区日游客量预测研究—以不同客户端百度指数为例[J]. 人文地理，2017，32(03)：152-160.