

KTBoost 预测模型的改进及应用^①



张曼, 牟莉

(西安工程大学 计算机科学学院, 西安 710048)
通信作者: 牟莉, E-mail: 626470481@qq.com

摘要: 针对目前 KTBoost 预测模型中存在的精度低、拟合效果较差的问题, 给出一种改进的 KTBoost 预测模型. 首先提出了 OGWO 算法, 使用反正切函数对传统灰狼优化算法 (GWO) 中的收敛因子进行优化, 以解决算法中的无效迭代问题, 然后运用 OGWO 算法对 KTBoost 模型中的超参数进行优化, 从而提高模型预测的精度; 最后, 为了验证模型的可行性, 将该模型及其他预测模型应用于交通流预测场景中进行对比. 实验结果表明: 相较于 RBF 模型、随机森林模型 (RFR)、KTBoost 模型、OGWO-RBF 模型、OGWO-RFR 模型, OGWO-KTBoost 预测模型拟合效果更好, 其决定系数值达到 0.8265, 在实际应用中有较好的预测效果.

关键词: OGWO 算法; KTBoost 预测模型; 超参数; OGWO-KTBoost 预测模型; 交通流预测

引用格式: 张曼, 牟莉. KTBoost 预测模型的改进及应用. 计算机系统应用, 2022, 31(4): 369-374. <http://www.c-s-a.org.cn/1003-3254/8470.html>

Improvement and Application of KTBoost Prediction Model

ZHANG Man, MU Li

(School of Computer Science, Xi'an Polytechnic University, Xi'an 710048, China)

Abstract: An improved KTBoost prediction model is proposed to address the low accuracy and poor fitting performance of the current KTBoost prediction model. First, the OGWO algorithm is put forward to solve the invalid iteration of the traditional gray wolf optimization (GWO) algorithm by using the arctangent function to optimize its convergence factor. Then, the OGWO algorithm is employed to optimize the hyperparameters in the KTBoost model, thereby improving the prediction accuracy of the model. Finally, the improved model and other prediction models are applied to traffic flow prediction scenarios for comparison to verify the feasibility of the model. The experimental results show that compared with the RBF model, random forest regression (RFR) model, KTBoost model, OGWO-RBF model, and OGWO-RFR model, the OGWO-KTBoost prediction model has better fitting performance and a better forecasting effect in practical application with its coefficient value of determination being 0.8265.

Key words: OGWO algorithm; KTBoost prediction model; hyperparameter; OGWO-KTBoost prediction model; traffic flow prediction

随着科技的高速发展, 可靠的数学模型在解决现实问题中起着重要的作用. 预测模型作为目前使用较多的分析方法, 可以通过对海量的历史数据不断学习, 确定数据特征之间的关系, 进而科学的预测某一变量的未来走势. 即根据学习到的对象特征预测其他的特

征值. 这种方法在各个领域引起了广泛的关注, 许多国家也已致力于对预测模型的研究.

现阶段预测模型广泛应用于医疗、交通、工业、军事、农业、商业等现实生活中的方方面面, 极大地提高了生产生活的效率. 常见的预测模型有时间序列

^① 基金项目: 陕西省自然科学基金 (2019JQ-849)

收稿时间: 2021-07-06; 修改时间: 2021-08-04; 采用时间: 2021-09-01; csa 在线出版时间: 2022-03-22

预测模型、灰色预测模型、神经网络预测模型、回归预测模型等. 如刘学刚等人使用以 ARIMA(差分自回归移动平均模型) 为代表的时间序列预测模型实现了对上海市大连路交通流的静态预测^[1], 但这种模型结构单一, 局限在于需要保证时序数据的稳定性且具有很强的空间依赖性, 本质上只能获取线性变化的关系. Zeng 等人使用灰色预测模型对粮食生产情况进行了预测^[2], 这种算法具有容易运算, 所需变量个数少的优点, 但仅仅适用于数据量呈指数增长的预测. Zeng 等人结合粒子群和 BP 神经网络实现了对校园交通流量的预测^[3], 这种组合的网络模型依赖于给定的数据集样本, 没有很好的泛化能力; KTBoost 预测模型作为回归预测模型中的新算法, 是由 Sigrist 于 2019年提出的一种增强算法^[4], 它创新性的将回归树^[5,6] 和 RKHS 函数^[7] 结合, 与原始的回归树、核提升算法相比, KTBoost 预测模型的预测精度有了显著提升. 同时, 将其与 SVR^[8-10]、XGBoost^[11]、随机森林^[12-14] 等传统的回归预测模型对比, KTBoost 预测模型的预测精度更高.

基于此, 为了获取 KTBoost 算法更精准的预测结果, 本文利用 OGWO 算法所具有的可寻找最优解的特性, 对 KTBoost 算法中的超参数进行调节, 并以交通流预测^[15-18] 为应用场景, 以各模型的决定系数值作为衡量标准进行对比实验. 实验结果证明相比于其他模型, OGWO-KTBoost 模型的预测精度更好, 拟合效果更佳.

1 优化 GWO 算法的 KTBoost 预测模型

1.1 GWO 算法及其改进

澳大利亚的学者们通过观察狼群的行为, 提出了灰狼优化算法 (grey wolf optimizer, GWO)^[19], 因为此算法具有设置参数较少, 易操作和实现等诸多优点, 引起了广泛的讨论, 在众多领域得到了应用^[20]. 其搜索代理群从高到低分别是 α 、 β 、 δ 和 ω 即分别代表算法中较好的 3 个解和剩余的其他解^[21]. 狼群主要通过 3 大行为完成对猎物的猎杀, 分别是追踪、狩猎和攻击.

GWO 对追踪行为定义如下:

$$\vec{D} = |\vec{C} \cdot \vec{X}_p(t) - \vec{X}(t)| \quad (1)$$

其中, t 表示更新次数, \vec{X} 表示狼群和猎物所处的位置, 为式中的系数向量, 表示狩猎者和被捕者之间的间隔和方位.

$$\vec{X}(t+1) = \vec{X}_p(t) - \vec{A} \cdot \vec{D} \quad (2)$$

其中, $\vec{X}(t+1)$ 表示对狼群的位置进行更新.

$$\vec{A} = 2\vec{a} \cdot \vec{r}_1 - \vec{a} \quad (3)$$

其中, \vec{a} 代表 $[0, 1]$ 之间的随机数, \vec{A} 为收敛因子. 收敛因子为线性递减函数, 初始值为 2, 之后随自变量的增加减小到 0^[22].

$$\vec{C} = 2 \cdot \vec{r}_2 \quad (4)$$

其中, \vec{r}_2 代表 $[0, 1]$ 之间的随机数.

$$\vec{D}_\alpha = |\vec{C}_1 \cdot \vec{X}_\alpha - \vec{X}| \quad (5)$$

$$\vec{D}_\beta = |\vec{C}_2 \cdot \vec{X}_\beta - \vec{X}| \quad (6)$$

$$\vec{D}_\delta = |\vec{C}_3 \cdot \vec{X}_\delta - \vec{X}| \quad (7)$$

其中, \vec{C}_1 , \vec{C}_2 , \vec{C}_3 分别是 $[0, 2]$ 之间的随机向量, \vec{D}_α , \vec{D}_β , \vec{D}_δ 分别表示 α 、 β 、 δ 个体狼跟踪猎物的位置向量.

$$\vec{X}_1 = \vec{X}_\alpha - A_1 \cdot \vec{D}_\alpha \quad (8)$$

$$\vec{X}_2 = \vec{X}_\beta - A_2 \cdot \vec{D}_\beta \quad (9)$$

$$\vec{X}_3 = \vec{X}_\delta - A_3 \cdot \vec{D}_\delta \quad (10)$$

其中, \vec{X}_1 , \vec{X}_2 , \vec{X}_3 分别表示 ω 狼群受到 α 、 β 、 δ 狼的影响后更新的位置向量.

$$\vec{X}(t+1) = \frac{\vec{X}_1 + \vec{X}_2 + \vec{X}_3}{3} \quad (11)$$

其中, $\vec{X}(t+1)$ 表示狼群参考 α 、 β 、 δ 的位置向量后更新的最终位置^[23].

$$|\vec{A}| > 1 \quad (12)$$

$$|\vec{A}| < 1 \quad (13)$$

由以上公式可知, 狼群主要通过与 α 、 β 、 δ 狼的距离确定狼群的位置, 因此如果收敛因子 \vec{A} 的绝对值大于 1, 狼群将分散开寻找猎物, 进行大范围的查找; 收敛因子 \vec{A} 的绝对值小于 1 表示狼群局部搜索, 对猎物进行围堵.

研究发现在有效的迭代范围内随着迭代次数的增加, \vec{a} 随之减小, 之后决定系数将不再变化, 即此时的 \vec{a} 变化已无法影响到决定系数的变化. 所以为了进一步对灰狼优化算法的勘探和搜索能力进行改进, 本文对 GWO 算法进行优化, 在此采取的方法是对收敛因子 \vec{a} 和 \vec{A} 进行调节: 依据反正切函数是单调递增函数的特

性,将反正切函数用于收敛因子的公式中,改进公式如下所示:

$$a = \begin{cases} 2, & i = 0 \\ 2.53 \cdot \arctan(\max(\text{sub}a_i, \text{sub}a_{i-1})) + \text{random}, & i > 0 \end{cases} \quad (14)$$

其中, i 表示收敛因子的迭代次数,为了避免陷入局部最优值,本文设置 random 的初始值为 0.01, 当 $i \% 3 = 0$ 时, random 值为 1.

1.2 OGWO-KTBoost 模型

KTBoost 模型以回归树和 RKHS 函数作为基学习器,将内核和树结合,获得一种新的增强算法,其中 RKHS 岭回归采用高斯核.相比于经典的树与核,不同类型基学习器的互补既具有树的学习不连续函数的优势,又具有 RKHS 函数适合学习平滑函数的特点,泛化误差更小,可以实现接近最优的测试误差. KTBoost 默认依赖所有特征,先采用梯度下降法^[24,25]并行学习回归树和惩罚 PKHS 回归函数,每次增强迭代时,从回归树和惩罚 PKHS 回归函数中选择其中一个作为基学习器,添加到集成中,以结构风险最小化学习规格,相比于其他算法,KTBoost 具有更高的泛化能力和非线性建模精度,其优势在小样本数据和有噪数据中表现明显.

尽管 KTBoost 模型优势突出,但底层思想还是通过对多个弱学习器的集成,使之成为一个强学习器,从而建立最终的预测模型.因此,同 Boosting 模型类似,KTBoost 模型的超参数对预测精度影响较大,而灰狼优化算法具有较强的参数寻优能力,利用灰狼优化算法可以选择 KTBoost 模型中超参数的最佳值,因此,本文采用优化后的灰狼优化算法(OGWO)优化 KTBoost 模型中弱学习器的个数 $n_estimators$ 和弱学习器的最大深度 max_depth 这两个重要参数,从而提出了 OGWO-KTBoost 模型,其算法步骤如下所示:

(1) 构建 OGWO 算法,对狼群数目 N 、最大迭代次数 Max_iteration 等相关参数进行初始化.

(2) 根据 OGWO 算法初始化 KTBoost 训练模型,并通过计算获取结果.

(3) OGWO 算法根据优化后的超参数,在下次迭代时重新设置超参数并训练模型.

1.3 特征工程

原始数据集中为零的数据已提前通过取缺失点前后数据均值的方式进行了数据修复.本文主要对每一

个传感器的时间戳进行操作,从时间戳中提取出 week , day , hour 和 minute 的时间特征,与历史的车辆数量结合起来作为预测因子构建 OGWO-KTBoost 模型进行学习.

2 实验与结果分析

2.1 数据来源及其环境设置

本文中的实验数据均来自经过实测的交通数据集 PeMs 数据集,其数据的时间始于 2014 年 5 月 1 日,结束于 2014 年 6 月 30 日,共计 61 天数据,包含 24 个传感器,每隔 5 分钟获取一次检测器数据,即每个传感器含有 17568 组数据,数据量大且全面.将数据集中 70% 的数据用于模型的训练,20% 的数据用于在验证中调节参数,剩余数据用于测试模型效果.

所有实验都在配置为 RAM 16.0 GB, Intel(R) Core(TM) i5-10210U CPU @ 1.60 GHz 2.11 GHz 的处理器,64 位操作系统的笔记本上完成的,仿真软件采用 PyCharm 3.0.

2.2 模型处理和参数调优

在回归预测问题中,判断模型好坏的标准有残留标准误差(RSE)、均方根误差(RMSE)、决定系数(R^2_score)等,其中是一个归一化的度量标准,它既能体现出预测值和真实值之间的差值,也能体现出真实值之间的差值.总而言之,通过决定系数可以体现出拟合效果的好坏,所以本文中所有实验都将 R^2_score 的结果作为判断模型优劣的依据,其结果保留小数点后 5 位.评价公式如下所示:

$$R^2 = 1 - \frac{\sum_i (\hat{y}^{(i)} - y^{(i)})^2}{\sum_i (\bar{y} - y^{(i)})^2} \quad (15)$$

其中, $y^{(i)}$ 、 \bar{y} 、 $\hat{y}^{(i)}$ 表示真实值、平均值、预测值. GWO 算法中当种群规模过大会影响收敛速度,但全局解整体朝向稳定的进化方向,即使出现极个别最优解,也影响较小;规模过小易出现局部最优解,但收敛速度快,因此需要不断调整默认参数-灰狼数量 SearchAgents_no 以确定最佳值.

如表 1 所示,为了呈现改进算法的效果,在 OGWO-KTBoost 模型上运行,通过控制变量法确保除灰狼数量之外的其他参数全部相同,经实验验证,当灰狼数量赋值 30 时效果最好.

表 1 不同灰狼数量的决定系数

超参名称	赋值	决定系数
SearchAgents_no	10	0.81728
	20	0.81713
	30	0.82650
	40	0.82187
	50	0.81764

灰狼优化算法主要通过调节 a 值完成勘探和开发的平衡, 分别将未改进的和改进后的 a 值在基于灰狼优化算法的 KTBoost 模型上运行. 在对传统的 GWO-KTBoost 模型运行后发现, 当迭代次数达到 6 后, 模型的决定系数基本保持在 0.78 左右, 后续 a 值的变化将不再影响决定系数, 即后续的迭代皆为无效迭代, 通过对参数 a 的改进, R^2_score 在原有的基础上有了一定的提升, 达到了 0.82650. 具体的 a 值对比结果如表 2 所示.

表 2 不同 a 值的决定系数对比表

名称	决定系数
传统 a 值	0.78281
改进后的 a 值	0.82650

2.3 多模型对比

为了更直观的体现出本文提出的 OGWO-KTBoost 预测模型性能的优势, 基于上述数据集和评价指标, 将 OGWO-KTBoost 与现有模型 RBF、RFR、KTBoost、OGWO-RBF、OGWO-RFR 分别进行对比分析不同

模型的参数设置如下: RBF 神经网络的隐藏层神经元核函数为 Gaussian 函数, 基函数中心采用前人已经验证过的成熟算法 K-means 算法进行训练, 权值参数采用梯度下降法. OGWO-RBF 模型中除原有的默认参数外, 运用 OGWO 优化 RBF 模型中的 c 和 γ . RFR 模型中的分类器个数 ($n_estimators$) 为 10, 最大特征数目 ($max_features$) 设置为 auto, 最大深度 (max_depth) 默认为 2, 采用五折交叉验证法. OGWO-RFR 模型中除 RFR 模型中的默认参数外, 运用 OGWO 优化 RFR 模型中的分类器的个数 $n_estimators$ 和最大特征数目 $max_features$. KTBoost 模型的默认参数如下: Alpha 为 1, $n_neighbors$ 为 5, $n_components$ 为 100, 分类器个数 ($n_estimators$) 取值 500, 最大特征数 ($max_features$) 取值 15. 各个模型的预测结果图如图 1-图 6.

为了清晰的呈现出预测效果, 在此通过实验预测

其中一天的车流量. 如图 1-图 6 所示, 以一天中的分钟值为横坐标, 实时车流量为纵坐标, 黑色折线表示实际的车流量走势, 红色折线为模型的拟合效果. 其中 RBF 模型的预测效果最差, 出现了欠拟合现象; RFR 模型预测效果不佳, 尤其在 300 分钟后预测结果明显小于实际结果; KTBoost 模型基本捕捉到了交通流预测的特征; OGWO-RBF 模型的前期预测效果差, 后期结果良好; OGWO-RFR 模型基本满足预测要求, 拟合效果良好. 而 OGWO-KTBoost 模型的拟合精度更佳, 预测效果最好. 各模型的预测数据如表 3 所示.

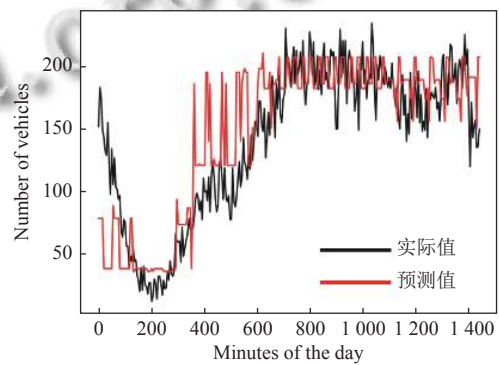


图 1 RBF 模型图

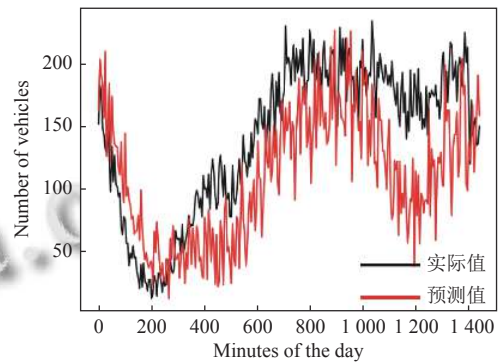


图 2 RFR 模型图

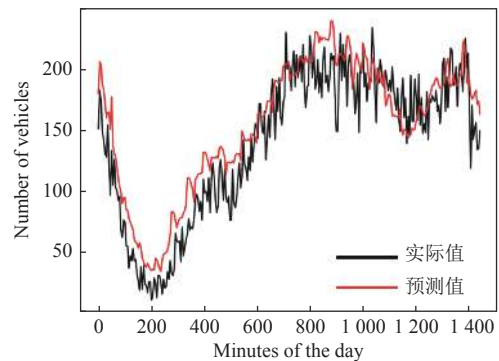


图 3 KTBoost 模型图

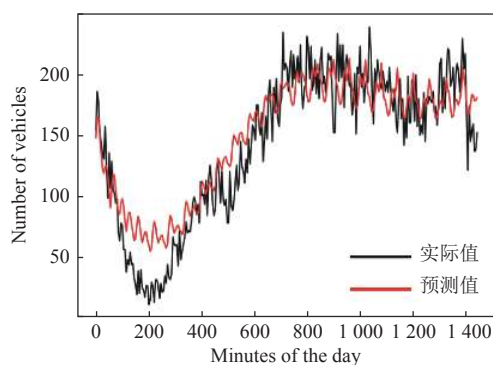


图4 OGWO-RBF模型图

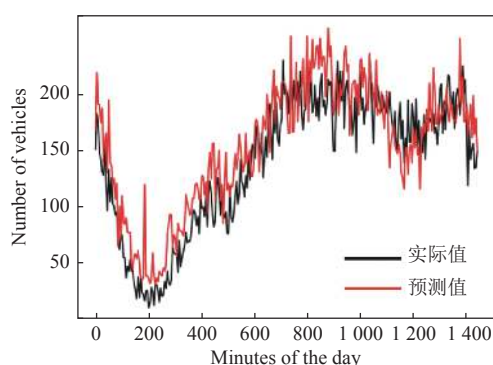


图5 OGWO-RFR模型图

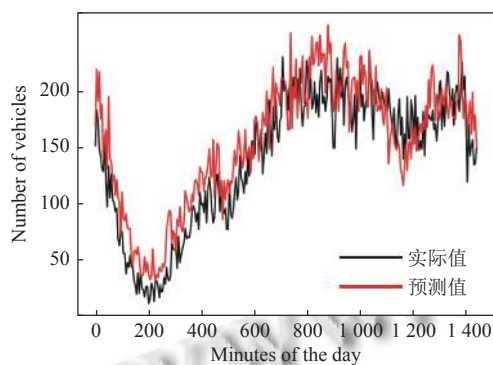


图6 OGWO-KTBoost模型图

如表3所示,预测最差的是RBF神经网络,基本不能满足预测要求,预测最好的是OGWO-KTBoost模型。KTBoost模型虽然与预期目标有所差距,但与RBF模型、RFR模型相比,可以看出它在预测方面具有显著优势。通过将OGWO-RBF模型与RBF模型的决定系数进行对比,可明显看出OGWO-RBF模型的决定系数已有了大幅度提升,从而侧面证明了OGWO算法适用于对模型的优化。

表3 各模型决定系数值

预测模型	系数
RBF	0.52124
RFR	0.63981
KTBoost	0.76586
OGWO-RBF	0.76041
OGWO-RFR	0.81837
OGWO-KTBoost	0.82650

3 结论

本文提出了一种新的预测模型,首先针对传统灰狼优化算法中存在的无效迭代问题,通过对参数 a 进行改进,提高了灰狼优化算法在有效迭代范围内的寻优速度。利用改进后的算法优化KTBoost模型的弱学习器的个数 $n_estimators$ 和避免过拟合参数 max_depth ,得到了OGWO-KTBoost预测模型。然后将该模型放在交通流预测的场景中,采用PeMs数据集进行实验,实验结果表明,OGWO-KTBoost模型预测精度约为0.8265,已达到预期目标值。同时将该模型与RBF、RFR、KTBoost、OGWO-RBF、OGWO-RFR模型进行了对比,由实验结果知OGWO-KTBoost预测模型优于其他预测模型,实际预测精度最佳。

参考文献

- 1 刘学刚,张腾飞,韩印.基于ARIMA模型的短时交通流预测研究.物流科技,2019,42(12):91-94,102. [doi: 10.3969/j.issn.1002-3100.2019.12.024]
- 2 Zeng B, Li H, Ma X. A novel multi-variable grey forecasting model and its application in forecasting the grain production in China. Computers & Industrial Engineering, 2020, 150: 106915.
- 3 Zeng C, Hua CY, Lei TY, et al. Short-term traffic flow prediction on campus based on modified PSOBP neural network. Journal of Physics: Conference Series, 2020, 1592: 012071.
- 4 Sigrist F. KTBoost: Combined kernel and tree boosting. Neural Processing Letters, 2021, 53(2): 1147-1160. [doi: 10.1007/s11063-021-10434-9]
- 5 Horiguchi A, Pratola MT, Santner TJ. Assessing variable activity for Bayesian regression trees. Reliability Engineering & System Safety, 2021, 207: 107391.
- 6 Liu QS, Wang XY, Huang X, et al. Prediction model of rock mass class using classification and regression tree integrated AdaBoost algorithm based on TBM driving data. Tunnelling

- and Underground Space Technology, 2020, 106: 103595. [doi: [10.1016/j.tust.2020.103595](https://doi.org/10.1016/j.tust.2020.103595)]
- 7 李立婷, 庄虹莉, 林雨婷, 等. 基于 Logistic 回归惩罚函数的企业财务危机预警. 科技和产业, 2017, 17(6): 62–65, 136. [doi: [10.3969/j.issn.1671-1807.2017.06.012](https://doi.org/10.3969/j.issn.1671-1807.2017.06.012)]
- 8 Zhang Y, Qiu JB, Zhang YG, *et al.* The adoption of a support vector machine optimized by GWO to the prediction of soil liquefaction. Environmental Earth Sciences, 2021, 80(9): 360. [doi: [10.1007/s12665-021-09648-w](https://doi.org/10.1007/s12665-021-09648-w)]
- 9 Shang S, He KN, Wang ZB, *et al.* Sea clutter suppression method of hfswr based on RBF neural network model optimized by improved GWO algorithm. Computational Intelligence and Neuroscience, 2020, 2020: 8842390.
- 10 张天逸, 孙毅然, 刘凡琪, 等. 基于 K 均值聚类算法与 RBF 神经网络的交通流预测方法. 智能计算机与应用, 2020, 10(8): 148–151. [doi: [10.3969/j.issn.2095-2163.2020.08.035](https://doi.org/10.3969/j.issn.2095-2163.2020.08.035)]
- 11 Mosavi A, Hosseini FS, Choubin B, *et al.* Ensemble boosting and bagging based machine learning models for groundwater potential prediction. Water Resources Management, 2021, 35(1): 23–37. [doi: [10.1007/s11269-020-02704-3](https://doi.org/10.1007/s11269-020-02704-3)]
- 12 Zhang YP, Yao XY, Wu Q, *et al.* Turbidity prediction of lake-type raw water using random forest model based on meteorological data: A case study of Tai lake, China. Journal of Environmental Management, 2021, 290: 112657. [doi: [10.1016/j.jenvman.2021.112657](https://doi.org/10.1016/j.jenvman.2021.112657)]
- 13 延会波, 张玮, 李立毅, 等. 滴流床反应器床层液体分布 GWO-RFR 模型的研究与优化. 天然气化工—C1 化学与化工, 2021, 46(2): 98–103.
- 14 王晓晖, 张亮, 李俊清, 等. 基于遗传算法与随机森林的 XGBoost 改进方法研究. 计算机科学, 2020, 47(S2): 454–458, 463.
- 15 Li T, Ni AN, Zhang CQ, *et al.* Short-term traffic congestion prediction with Conv-BiLSTM considering spatio-temporal features. IET Intelligent Transport Systems, 2020, 14(14): 1978–1986. [doi: [10.1049/iet-its.2020.0406](https://doi.org/10.1049/iet-its.2020.0406)]
- 16 Culita J, Caramihai SI, Dumitrache I, *et al.* An hybrid approach for urban traffic prediction and control in smart cities. Sensors, 2020, 20(24): 7209. [doi: [10.3390/s20247209](https://doi.org/10.3390/s20247209)]
- 17 张文胜, 郝孜奇, 朱冀军, 等. 基于改进灰狼算法优化 BP 神经网络的短时交通流预测模型. 交通运输系统工程与信息, 2020, 20(2): 196–203.
- 18 陈佳维. 基于 K 近邻非参数回归方法的短时交通流预测 [硕士学位论文]. 成都: 西南交通大学, 2017.
- 19 张晓凤, 王秀英. 灰狼优化算法研究综述. 计算机科学, 2019, 46(3): 30–38. [doi: [10.11896/j.issn.1002-137X.2019.03.004](https://doi.org/10.11896/j.issn.1002-137X.2019.03.004)]
- 20 凌颖, 杨春燕, 黎新, 等. 改进灰狼优化算法的研究. 科学技术创新, 2020, (16): 61–63. [doi: [10.3969/j.issn.1673-1328.2020.16.033](https://doi.org/10.3969/j.issn.1673-1328.2020.16.033)]
- 21 Zhao J, Gao ZM. An improved grey wolf optimization algorithm with multiple tunnels for updating. Journal of Physics: Conference Series, 2020, 1678: 012096. [doi: [10.1088/1742-6596/1678/1/012096](https://doi.org/10.1088/1742-6596/1678/1/012096)]
- 22 Lakhlef N, Oudira H, Dumond C. Optimal pattern synthesis of linear antenna arrays using modified grey wolf optimization algorithm. Instrumentation Measure Métrologie, 2020, 19(4): 255–261.
- 23 Hu P, Pan JS, Chu SC. Improved binary grey wolf optimizer and its application for feature selection. Knowledge-Based Systems, 2020, 195: 105746. [doi: [10.1016/j.knosys.2020.105746](https://doi.org/10.1016/j.knosys.2020.105746)]
- 24 Zhou BC, Han CY, Guo TD. Convergence of stochastic gradient descent in deep neural network. Acta Mathematicae Applicatae Sinica, English Series, 2021, 37(1): 126–136. [doi: [10.1007/s10255-021-0991-2](https://doi.org/10.1007/s10255-021-0991-2)]
- 25 Chandrashekar A, Desai U, Abhilash P. Cost prediction using gradient descent algorithm. Journal of Physics: Conference Series, 2020, 1706: 012038. [doi: [10.1088/1742-6596/1706/1/012038](https://doi.org/10.1088/1742-6596/1706/1/012038)]