

基于 IndRNN 的机场起飞航班延误预测模型研究

司毅洋¹ 吕娜¹
SI Yiyang LYU Na

摘要

由于没有综合考虑天气等突发状况,导致航班延误预测结果存在一定的偏差,降低用户飞行体验,对此,提出基于 IndRNN 的机场起飞航班延误预测模型研究方法。起飞航班延误划分为 4 个等级,利用 ReLU 激活函数代替 IndRNN 网络中的 sigmoid、tanh 激活函数,使得每个神经元都有其独立的时空特征;分离所有神经元,避免梯度出现消失爆炸的情况;经过数据读取、数据预处理、数据融合等一系列操作后,完成航班延误预测模型的构建。通过开展对比仿真实验,在 4 项评判指标下,所提方法均展现出了优秀的预测性能,且预测延误航班数、延误时间与实际值非常接近。

关键词

IndRNN 网络;起飞航班延误预测;ReLU 激活函数;传播梯度;数据预处理

doi: 10.3969/j.issn.1672-9528.2024.06.025

0 引言

近年来,我国航班数量出现了大幅度增长,航班运输网络也越来越庞大、繁杂,因天气、飞机故障等原因导致的航班延误现象屡见不鲜,由此产生了巨大的经济损失。因此,展开对机场离港航班的延误预测是非常有必要的,可以在一定程度上帮助航空公司、机场以及相关单位制定延误解决方案,降低延误出现的次数,降低因延误产生的经济损失。

对此,吴仁彪等人^[1]提出利用 CBAM-CondenseNet 实现对航班的延误预测。首先,分析航空系统中因航班延误产生的波及现象,得到受影响的航班链;然后,对航班链中的数据进行清洗,将其中的机场数据与航班数据进行融合;最后,对融合后的结果通过 CBAM-CondenseNet 算法完成特征提取,并结合 Softmax 分类器划分航班延误等级,完成预测。该方法仅对受影响的航班进行分析,没有考虑其他航班可能遇到的突发状况,适用范围较为狭隘。张成伟等人^[2]通过分析离港航班计划,确定某航班出现延误的情况,完成预测。

1. 新乡工程学院信息工程学院 河南新乡 453000

参考文献:

- [1] 崔宇童,牛强,王志晓.基于信号传递的半监督谱聚类社区发现算法[J].计算机工程与设计,2018,39(5):1201-1205+1213.
- [2] 张书博,任淑霞,吴涛.结合概率矩阵的改进谱聚类社区发现算法[J].西安电子科技大学学报,2019,46(3):167-172.
- [3] 张晓琴,安晓丹,曹付元.基于谱聚类的二分网络社区发现算法[J].计算机科学,2019,46(4):216-221.
- [4] 赵承志.融合节点信息的 LPA 社区检测算法的改进研究[D].沈阳:东北财经大学,2022.
- [5] 蔡威林,葛斌.基于影响度的标签传播算法[J].佳木斯大学学报(自然科学版),2022,40(1):38-40+160.
- [6] GREGORY S. Finding overlapping communities in networks by label propagation[J]. New journal of physics, 2010, 12: 103018-103043.
- [7] 孙悦.基于 GNN 的异质网络重叠社区发现算法研究[D].包头:内蒙古科技大学,2023.

- [8] 李昕泽.基于图神经网络的社区发现方法研究[D].北京:北方工业大学,2023.
- [9] YE X, SAKURAI T. Robust similarity measure for spectral clustering based on shared neighbors[J]. ETRI journal, 2016, 38(3): 540-550.
- [10] LIU Q, WANG G, LI F, et al. Preserving privacy with probabilistic indistinguishability in weighted social networks[J]. IEEE transactions on parallel and distributed systems, 2017, 28(5): 1417-1429.
- [11] CAO S, DEHMER M, SHI Y. Extremality of degree-based graph entropies[J]. Information sciences, 2014, 278(10): 22-33.

【作者简介】

王晓娟(1998—),女,山东德州人,硕士研究生,研究方向:社交网络隐私保护。

(收稿日期:2024-03-22)

从历史航班运行数据中找出共同的内在特征，建立航班数据贝叶斯网络模型，得到不同环境下航班出现延误的概率值；利用动态贝叶斯网络推理方法，结合航班实际数据，建立隐马尔科夫延误预测模型；利用模型中的解码问题算法实现离港航班延误的预测。该算法仅考虑了历史数据，并未将天气原因、飞机故障等因素考虑在内。

基于此，本文提出基于 IndRNN 的机场起飞航班延误预测模型分析方法。增强 IndRNN 网络内各个神经元之间的联系，将两层网络堆叠在一起，构建更深、更长的网络结构，避免网络陷入消失爆炸。通过数据读取、数据预处理以及构建网络等一系列处理后，实现起飞航班延误模型的构建。通过与其他方法展开对比仿真实验，结果也验证了本文方法具有优秀的预测精度和预测性能，可有效减少因航班延误造成的经济损失，在一定程度上也提高了客户对机场的满意度。

1 起飞航班延误等级划分

根据相关规定标准，当机组得出退出指令后，地面工作人员拿走航空器^[3]最后一个轮档时间比原定计划时间晚 15 分钟，即判定该航班出现延误现象；机上延误指的是关闭舱门后至起飞前或者航班降落后至打开舱门前，旅客等待的时间较规定地面滑行时间长。由此可给出起飞航班延误的定义：航班实际起飞时间 M_s 比计划起飞时间 M_p 晚 15 min 以上，再与机场规定的地面滑行时间 T_k 进行相加。计算公式为：

$$M_p - M_s \geq 15 \text{ min} + T_k \quad (1)$$

每个城市机场对地面滑行时间的规定有所不同，本文以 $T_k=30 \text{ min}$ 为标准。为了便于分析起飞航班的延误程度，划分了如表 1 所示的航班延误等级。

表 1 机场起飞航班延误等级划分标准

航班延误等级	航班延误时长	具体描述
0	$T_k \leq 45 \text{ min}$	航班正常起飞
1	$45 \text{ min} < T_k \leq 90 \text{ min}$	航班一般延误
2	$90 \text{ min} < T_k \leq 180 \text{ min}$	航班中度延误
3	$T_k > 80 \text{ min}$	航班重度延误

地面滑行时间 T_k 的计算公式为：

$$T_k \geq M_p - M_s \quad (2)$$

2 基于 IndRNN 的机场起飞航班延误预测模型

IndRNN 在 ReLU 激活函数^[4]的基础上，将网络进行多层堆叠，加深网络层次，使得其较传统循环神经网络(recurrent neural network, RNN)相比，可有效避免梯度消失爆炸现象，取得更理想的干扰特征。同时，IndRNN 具有超强的序列特征提取能力，因此，本文在 IndRNN 中引入一维交通流数据，实现更精准的机场起飞航班延误预测。

2.1 IndRNN 算法

为了避免出现梯度弥散爆炸问题，IndRNN 将 sigmoid、tanh 激活函数利用 ReLU 激活函数做替换，以此来保证网络层内神经元具有独立的时空特征。

(1) 分离网络层内神经元

IndRNN 网络层的循环输入公式为：

$$l_t = \alpha(Qx_t + W \times l_{t-1} + m) \quad (3)$$

式中： t 表示当前干扰因素时刻； W 表示向量矩阵； $W \times l_{t-1}$ 表示矩阵点乘，也就是网络层内神经元位置元素相乘的结果； α 表示循环系数； m 表示网络层内的系数值； x_t 表示输入向量的权重矩阵； Q 表示循环向量。

将隐藏层中的干扰因素神经元在 t 时刻和 $t-1$ 时刻下的状态信息作为输入内容，且不与其他层连接。那么，第 n 个神经元可用公式 (4) 表示为：

$$h_{n,t} = \alpha(E_n x_t + W_n \times h_{n,t-1} + m_n) \quad (4)$$

式中： E_n 表示排列在矩阵 E 中的第 n 行， W_n 表示排列在矩阵 W 中的第 n 行。

(2) 避免梯度消失爆炸

将优化函数定义为 J_n ，在 t 时刻下的误差梯度反传^[5]计算结果为：

$$\begin{aligned} \frac{\partial J_n}{\partial h_{n,t}} &= \frac{\partial J_n}{\partial h_{n,T}} \frac{\partial h_{n,T}}{\partial h_{n,t}} = \frac{\partial J_n}{\partial h_{n,T}} \prod_{K=t}^{T-1} \frac{\partial h_{n,K+1}}{\partial h_{n,K}} \\ &= \frac{\partial J_n}{\partial h_{n,T}} \prod_{K=t}^{T-1} \alpha_{n,K+1}^T m_n = \frac{\partial J_n}{\partial h_{n,T}} m_n^{T-1} \prod_{K=t}^{T-1} \alpha_{n,K+1}^T \end{aligned} \quad (5)$$

式中： ∂ 表示求导函数， T 表示求导周期， K 表示误差梯度系数。

通过式 (5) 可以得出，梯度反传的结果与 m 和激活函数的导数有着很大的关系。通常情况下，激活函数的取值范围设定在 $[0,1]$ 之间，所以对梯度反传结果影响最大的就是 m 。将 $m_n^{T-1} \prod_{K=t}^{T-1} \alpha_{n,K+1}^T$ 的值设置为合适的大小，保证梯度反向传播的尺度变化在合理范围内。

为了避免在 t 时刻下的梯度传递出现弥散现象^[6-8]，需要保证最小有效梯度的值始终大于 ε ，定义 W 矩阵，使得：

$$|m_n| \in^{T-t} \sqrt{\frac{\varepsilon}{\prod_{K=t}^{T-1} \alpha_{n,K+1}^T}} \quad (6)$$

与此同时，需要重新设置 W 矩阵，使得：

$$|m_n| \in \left[T-t \sqrt{\frac{\varepsilon}{\prod_{K=t}^{T-1} \alpha_{n,K+1}^T}}, T-t \sqrt{\frac{\gamma}{\prod_{K=t}^{T-1} \alpha_{n,K+1}^T}} \right] \quad (7)$$

式中： γ 表示梯度反传极值系数。

通过式 (7)，可得到梯度反传的极大值和极小值，在

一定程度上可有效避免梯度消失爆炸。

(3) 多层堆叠

根据经验已知,处于同一层内的影响航班延误的干扰神经元间独立存在,联系极少,学习时空特征也有所不同。正是由于这种特性,导致航班延误预测结果极易出现误导。因此,需要将至少两层结构堆叠在一起,以此增强各个神经元之间的联系。两层 IndRNN 网络堆叠结构图如图 1 所示。

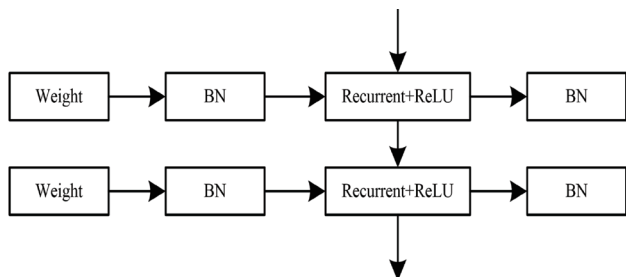


图 1 两层 IndRNN 网络堆叠结构图

图 1 中, weight 模块和 Recurrent+ReLU 模块均表示当前时刻下对输入内容的处理过程。这两个模块可以是循环模块。通过将单层的 IndRNN 网络堆叠在一起,达到加深网络层次的目的。

同时,可以在 IndRNN 网络结构中适当引入残差侧向连接,增强传播梯度的连接效率,进一步缓解 IndRNN 网络陷入梯度消失爆炸,从而构建更长、更深层次的网络结构,全方位考虑干扰影响。

2.2 起飞航班延误预测模型构建

基于 IndRNN 的机场起飞航班延误预测大体上分为 4 个步骤,分别是:数据读取^[9]、数据预处理、数据融合以及构建航班延误预测模型,接下来进行具体描述。

(1) 数据读取

将 X 看作是全部航班数据, Y 看作是与数据对应的标签,也就是真实值。 X 和 Y 的读取维度分别是 (b, q) 和 (b, l) 。其中, b 表示根据航班数据构建的样本数量, q 表示在样本中观察到的航班延误数量, l 表示不同时刻输入航班数据的数量, Y 中的行表示当前时刻下航班数量的实际值。按照规定比例,将这些数据分为训练集和验证集。

(2) 数据预处理

对所有航班数据做归一化处理^[10-12],使其均匀分布在 0 到 1 之间,以此增加 IndRNN 网络的训练速度。预处理表达式为:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (8)$$

式中: X_{min} 、 X_{max} 分别表示航班数据中的最小值和最大值。

为了满足 IndRNN 网络的训练要求,本文将一维时间序列进行扩展,使其转换为三维张量,形状为 $[b, q, l]$ 。

(3) 数据融合

用 N 来表示所有航班数据,包括航班计划起飞日期和时刻、航班实际起飞日期和时刻、始发地机场和目的地机场、航线距离等。用 H 来表示气象数据,包括航班起飞当天的气温、气压、风速、云高、风向、能见度以及一些特殊天气数据等。

将所有航班数据和气象数据融合在一起,并将航班日期与当日天气对应上,可以更加具体地得到航班延误的情况。

(4) 起飞航班延误预测模型

对经过融合处理后的数据,建立起飞航班延误预测模型表达式为:

$$g(N, H) = \begin{cases} 1, & \text{航班出现延误} \\ 0, & \text{航班未出现延误} \end{cases} \quad (9)$$

式中: $N=(N_1, N_2, \dots, N_{14})^T$ 表示始发地机场和目的地机场、航线距离等 14 项航班数据; $H=(H_1, H_2, \dots, H_{18})^T$ 表示气压、风速、云高等 18 项气象数据。

利用 IndRNN 网络预测起飞航班延误问题,就是通过网络内层层结构的强大学习能力,来挖掘航班数据与气象数据之间存在的关联。

对于 IndRNN 网络隐藏层的激活函数,本文选择 tanh 函数,表达式为:

$$f(x) = \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (10)$$

式中: e^x 、 e^{-x} 分别表示激活函数的正向系数和负向系数。

对于 IndRNN 网络输出层的激活函数,本文选择的是 sigmoid 函数:

$$r(x) = \frac{1}{1 + e^x} \quad (11)$$

航班延误预测的根本是二分类问题^[13],由于交叉信息熵函数在这方面有着非常出色的表现,所以本文将其作为预测模型的损失函数,完整的预测模型公式为:

$$S = \frac{1}{L} \sum_{i=1}^L \sum_{j=1}^l d_i \ln p_i \quad (12)$$

式中: p_i 表示起飞航班延误时间, d_i 表示航班实际起飞时间, L 表示损失函数系数。

3 仿真实验

为了验证本文方法在实际应用中是否同样合理有效,与引言中提到的 CBAM-CondenseNet 算法和贝叶斯网络模型进行了仿真对比实验。实验平台由 64 位的 Windows 专业版系统下的 Keras 深度学习框架完成搭建,处理器采用的是

Inter(R) Coer(TM) i7。

3.1 实验数据

实验数据选用某机场 2020 年 6—8 月和 2021 年 1—8 月的起飞航班离港数据和气象情况, 合计共 152 362 条数据。

首先对实验数据进行预处理, 过程如下所示。

(1) 将所获得的起飞航班离港数据和气象数据进行筛选, 找出其中有用的数据, 同时将无用数据做删除处理。

(2) 在所有航班离港数据中, 将航班取消的数据以及当天的气象数据剔除掉。

(3) 将所获得的气象数据转换为数值数据, 按照存在该气象类型用数字 1 表示, 不存在用数字 0 表示的规则进行转换。

(4) 对所有数据进行离散化处理, 并转换为二进制字符数据。

(5) 将转换后的离港数据和气象数据利用标准差标准化方法做标准化处理, 其公式为:

$$D^* = \frac{D - \bar{D}}{\theta} \quad (13)$$

式中: $\bar{D}$ 、 θ 分别表示原始数据的均值和标准差。

(6) 经过式 (9) 的标准化处理后, 将离港数据与气象数据进行融合, 得到机场起飞航班延误数据集。

(7) 经过一系列处理后的数据集数量较少, 无法满足实验需要。经过反复的复制后, 数据集数量扩充为 485 232 条, 其中训练集、验证集以及测试集的数据数量分别为 263 514 条、114 526 条以及 107 192 条。

3.2 实验指标

本文选取平方相关系数 R^2 、均方根误差 RMSE、最小信息准则 AIC 以及平均绝对误差 MAE 作为算法的性能评判指标, 表达式为:

$$R^2 = \frac{\sum_{i=1}^n (\chi_i - \bar{\chi}_i)^2 - \sum_{i=1}^n (\hat{\chi}_i - \bar{\chi}_i)^2}{\sum_{i=1}^n (\chi_i - \bar{\chi}_i)^2} \quad (14)$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (\chi_i - \hat{\chi}_i)^2} \quad (15)$$

$$\text{AIC} = -2 \ln(L) + 2f \quad (16)$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |\chi_i - \hat{\chi}_i| \quad (17)$$

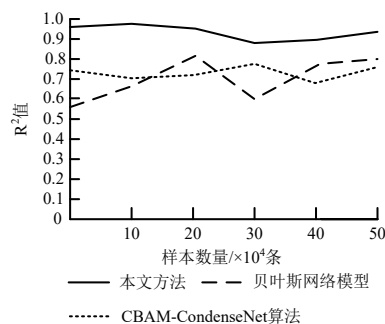
式中: χ_i 表示响应变量, $\hat{\chi}_i$ 表示 χ_i 的估计值, $\bar{\chi}_i$ 表示 χ_i 的均值, L 表示响应值的极大似然函数估计, f 表示算法中具有独立参数的数据数量。

R^2 值越接近 1, 说明算法的预测性能越优。RMSE 可以反映算法预测的误差大小。RMSE 的值越小, 说明算法的预

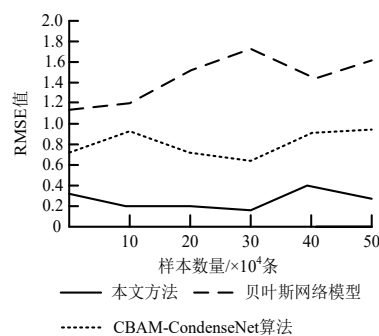
测精度越高。AIC 可以反映出算法的复杂度和拟合数据的有效性。AIC 的值越小, 说明算法的数据拟合效果越好。MAE 是实际值与算法平均值偏差的绝对值的平均值, 可以很好地反映出算法预测结果与实际值之间的偏差。MAE 的值越小, 说明算法的预测误差越小。

3.3 实验结果及分析

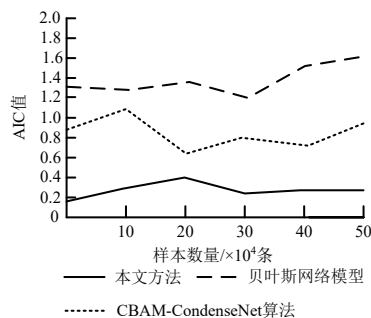
首先, 利用上述四种评判指标对三种算法的性能展开测试, 结果如图 3 所示。通过观察图 3 可以很明显的看出, 利用本文方法取得的预测结果 RMSE、AIC、MAE 值都是最小的, R^2 值与 1 最为接近。同时, 4 幅图像中, 本文方法的曲线变化最为平缓, 没有出现较大的波动, 而其他两种方法曲线均出现了不同程度的波动, 贝叶斯网络模型最为突出。这说明了本文方法对于数据的拟合效果较为优秀, 同时保证了较高的预测精度。这是由于本文方法对 IndRNN 网络进行了堆叠处理, 增加了网络深度和层次, 在一定程度上提高了预测精度。



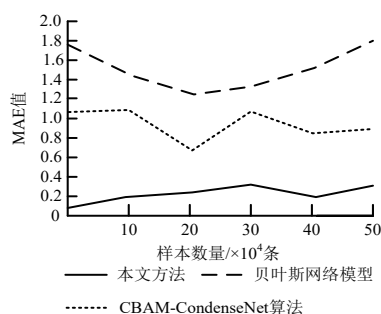
(a) 三种算法 R^2 值对比



(b) 三种算法 RMSE 值对比



(c) 三种算法 AIC 值对比



(d) 三种算法 MAE 值对比

图3 三种算法预测性能对比

接下来，将三种算法的预测结果与实际情况进行对比，图4为三种方法的迭代速率对比，图5为三种算法预测的起飞航班离港延误总时间与实际值的对比。从图4中可以看出，本文方法在第35次迭代时误差曲线就达到了最优，而贝叶斯网络模型和CBAM-CondenseNet算法分别在第40次和第63次迭代时曲线逐渐趋于平缓。这说明本文方法可在最短迭代周期内完成航班的延误预测。再观察图5，本文方法的预测延误时间与实际值之间的平均误差保持在30分钟之内，其他两种方法的预测结果与实际值偏离较大。由此可以说明，本文方法在起飞航班延误预测方面具有一定的有效性和可行性。

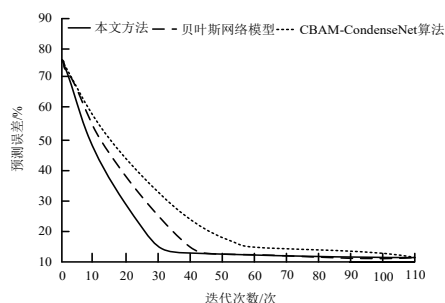


图4 三种算法航班延误预测迭代速率对比

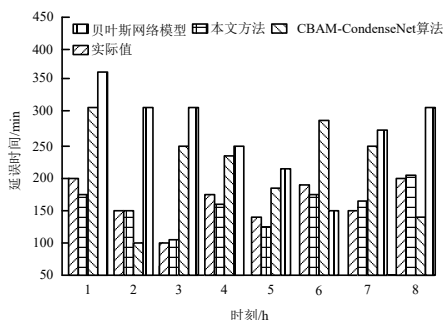


图5 起飞航班离港延误总时间

4 结论

针对现有方法难以实现精准的航班延误预测，提出基于IndRNN的起飞航班延误预测模型研究方法。通过将IndRNN网络堆叠，加深网络层次和深度，从而保证算法具有理想的预测精度。通过与其他算法展开对比仿真实验，结果表明，

所提方法具有较高的预测精度和数据拟合度，同时，预测的延误航班数和延误时间与实际值非常接近，可以将其应用在机场，帮助工作人员预测未来航班的延误情况以及监测机场人员拥挤程度，从而帮助机场降低因延误产生的经济损失，提升客户满意度。在未来的研究工作中，考虑将多种方法组合在一起，通过组合预测的方式实现多方面预测，在现有基础上再次提高预测精度。在条件允许的情况下，可针对具体机场和航班制定延误预测方法，实现更精准的预测效果。

参考文献：

- [1] 吴仁彪,赵娅倩,屈景怡,等.基于CBAM-CondenseNet的航班延误波及预测模型[J].电子与信息学报,2021,43(1):187-195.
- [2] 张成伟,罗凤城,代毅.基于数据挖掘的指定航班计划延误预测方法[J].计算机科学,2020,47(S2):464-470+485.
- [3] 屈景怡,董樑,曹烨琇,等.基于图簇随机连接的CliqueNet航班延误预测模型[J].计算机应用,2020,40(8):2420-2427.
- [4] 谷润平,来靖晗,时统宇,等.基于小波分解与ARMA-RBF模型的航班延误时间短期预测[J].飞行力学,2021,39(4):88-94.
- [5] 唐红,褚文奎,何林远,等.基于非线性赋权XGBoost算法的航班延误分类预测[J].系统仿真学报,2021,33(9):2261-2269.
- [6] 丁建立,孙玥.基于LightGBM的航班延误多分类预测[J].南京航空航天大学学报,2021,53(6):847-854.
- [7] 刘继新,杨光.基于KNN的机场航班短期延误风险预测[J].重庆交通大学学报(自然科学版),2021,40(12):12-18.
- [8] 杨俊,刘芳,张雄威,等.基于Skip-LSTM的机场群延误预测模型[J].信号处理,2020,36(4):584-592.
- [9] 王春政,胡明华,杨磊,等.基于Agent模型的机场网络延误预测[J].航空学报,2021,42(7):452-465.
- [10] 王丹,王萌,王晓曦,等.用于航班延误预测的集成式增量学习算法[J].北京工业大学学报,2020,46(11):1239-1245.
- [11] 张亚东,葛晓程,郭进,等.基于GBDT的列车晚点时长预测模型研究[J].铁道标准设计,2021,65(8):149-154+176.
- [12] 卢宇红,宋佳丽,王萌,等.基于深度神经网络融合稀疏分组lasso的预测模型研究[J].中国卫生统计,2021,38(6):821-827.
- [13] 杨英,唐平.VAE_LSTM算法在时间序列预测模型中的研究[J].湖南科技大学学报(自然科学版),2020,35(3):93-101.

【作者简介】

司毅洋(1991—),女,河南郑州人,硕士,助教,研究方向:计算机应用。

吕娜(1993—),女,河南新乡人,硕士,助教,研究方向:软件工程。

(收稿日期:2024-04-12)