吕婷, 薛琼, 闵兴华, 等. 基于经典机器学习模型的河流重点水质预测[J]. 净水技术, 2026, 45(1): 150-156.
LÜ T, XUE Q, MIN X H, et al. Prediction of key water qualities for rivers based on classical machine learning models [J]. Water Purification Technology, 2026, 45(1): 150-156.

基于经典机器学习模型的河流重点水质预测

吕　婷*,薛　琼,闵兴华,金　哲
(南京市生态环境保护科学研究院,江苏南京　210019)

摘　要　近年来,水体污染问题日益突出,给河道环境造成巨大压力,故河道水环境破坏问题亟待解决。机器学习是一种基于大量监测数据的水质预测预警方法,是河道治理的新途径。**【目的】**　本文旨在对比不同模型对不同水质指标的预测能力。**【方法】**　本文以长江中下游平原某河流断面为例,首先通过显著性分析和主成分分析筛选出主要水质影响因子,随后根据自动监测站点数据选用支持向量机(SVM)和长短期记忆网络(LSTM)、门控循环单元(GRU)、时间卷积网络(TCN)神经网络模型对水质水平进行模拟。**【结果】**　氨氮和溶解氧(DO)是筛选出的主要影响因子,4 种模型对氨氮模拟的准确度高于 DO。**【结论】**　GRU 模型对 2 种指标的模拟最具优势,SVM 模型对氨氮和 LSTM 模型对 DO 水质模拟具有相对优势,而 TCN 模型对氨氮和 DO 的预测能力均相对较弱。

关键词　水质预测　影响因子识别　机器学习　支持向量机(SVM)　人工神经网络(ANN)
中图分类号:X52　　**文献标志码**:A　　**文章编号**:1009-0177(2026)01-0150-07
DOI:10.15890/j.cnki.jsjs.2026.01.018

Prediction of Key Water Qualities for Rivers Based on Classical Machine Learning Models

LÜ Ting*, XUE Qiong, MIN Xinghua, JIN Zhe
(*Nanjing Academy of Ecological and Environmental Protection Science, Nanjing　210019, China*)

Abstract　In recent years, the problem of water body pollution has become increasingly prominent, and causing enormous pressure on the river environment. Therefore, the problem of river water environment damage urgently needs to be solved. Machine learning is a water quality prediction and warning method based on a large amount of monitoring data, which is a new approach to river management. [**Objective**]　This paper aimed to explore differences in the predictive ability of different models for different water quality indices. [**Methods**]　This experiment took a river section in the middle and lower Yangtze Valley Plain as an example. First, main water quality influencing factors were selected through significance analysis and principal component analysis. Then, neural network models, namely, support vector machine (SVM), long-term and short-term memory network (LSTM), gated cycle unit (GRU), and time convolution network (TCN) were selected according to the automatic monitoring station datas to simulate the levels of water qualities nitrogen and DO. [**Results**]　Ammonia nitrogen and dissolved oxygen (DO) were selected as the main influencing factors. And the accuracy of the four models in simulating ammonia nitrogen was higher than that of DO. [**Conclusion**]　GRU model has the most advantages in simulating the two indices. SVM models have relative advantages in simulating ammonia nitrogen and DO water quality, respectively, while TCN model has relatively weak predictive ability for ammonia nitrogen and DO.

Keywords　water quality prediction　influencing factor identification　machine learning　support vector machine (SVM)　artifical neural network (ANN)

随着水环境治理工作的不断深入,各地陆续加大投入,建立了大量的水质自动站和小微站,用于日常水质的监控与预警。与传统的手工监测方式不同,自动监测的指标较少,准确性和精度相对较低,

[收稿日期]　2024-01-10
[基金项目]　南京环保科技项目(202006)
[通信作者]　吕婷(1991—　),女,工程师,主要从事水环境管理与水污染防治等工作,E-mail:532640535@qq.com。

但是其数据量巨大,有较强的连续性,能够反映出水水质波动所存在的问题。自动站建设之后产出的海量数据如何高效利用,从而更好地为环境管理持续输出效益,成为研究热点之一。

近年来,人工智能方法(例如机器学习)迅猛发展,为流域水质模型建模提供了新的途径。Yahya等[1]针对兰加特河流域开发了一个基于支持向量机(SVM)的预测模型,选定6个主要的水质指标进行预测,获得了较为精准的试验结果,为水中污染物变化的判断提供了技术支撑。人工神经网络(ANN)是另一种广泛应用的方法,其原理是通过对人脑的组织结构进行抽象模拟,构建相关的神经模型。Li等[2]针对池塘中溶解氧(DO)水平建立递归神经网络(RNN)、长短期记忆(LSTM)和门控循环单元(GRU)3种神经网络模型,LSTM作为RNN的一个特殊变种,在处理时序数据的问题上具有一定优势,GRU在LSTM的基础上,只保留2个门控单元,参数相对较少且更容易收敛。通过对试验结果分析比较表明:GRU和LSTM的预测性能优于RNN,选取最佳预测模型,能够提前1 d预测DO的变化,为发展池塘养殖业提供了数据支撑。Liu等[3]利用LSTM模型对饮用水水质数据进行预测,试验预测值与真实值误差较小,基本反映了水质随时间变化的趋势,表明模型具有可行的预测性能。Hu等[4]提出了一种考虑皮尔逊相关系数下的LSTM模型水质预测方法,得到了高达98.56%和98.97%的预测精度。时间卷积网络(TCN)是近年提出的一种用于解决时序预测的算法,具有对多个水质指标同时预测的潜力[5]。此外,引入因子分解机的GRU可有效解决水质数据高稀疏度和高维度的问题,提高了模型对水质预测的准确性[6]。

本文以J河水系为研究对象,首先通过影响因子识别筛选出影响断面水质的主要因子,再基于机器学习算法对南京J河水质自动监测站点收集的水质数据进行处理、分析,并进行相关的预测试验验证,比较不同模型的优劣,以实现对水质数据重要指标的预测任务。

1 模型与评价

1.1 水质预测模型

1.1.1 SVM

SVM是基于核的算法中最为典型的一种,其本质是基于回归的数学统计分析。SVM可以将低维输入向量输入非线性地映射到一个非常高维的空间,从而得到一个线性的分离超平面。在这个高维空间内,数据之间存在一种线性关系,且由超平面分开的2个空间内边界上的样本点(即支持向量,到该超平面的距离最大化)尽量远离这个超平面。在试验中,水质序列由于其复杂的非线性关系难以被线性可分,当它被SVM映射到一个很高的维度时,得到一个关于水质的回归函数。此时水质预测的任务可以理解为通过这个求得的高维回归函数,输入一定的历史水质数据即可得到对应的输出,即预测下一个时间点的数据,其实现逻辑符合水质预测问题,本质是一个回归问题的判断。SVM回归的求解本质是一个经典的凸优化问题。

1.1.2 ANN

ANN也简称为神经网络(NN)或称作连接模型(CM),它是一种模仿动物神经网络行为特征,进行分布式并行信息处理的算法数学模型。这种网络依靠系统的复杂程度,通过调整内部大量节点之间相互连接的关系,从而达到处理信息的目的。

常见的神经网络模型包括卷积神经网络(CNN)、循环神经网络(RNN)、长短期记忆网络(LSTM)、GRU和TCN等,现阶段最为常用的神经网络包括CNN、LSTM、GRU,其中LSTM和GRU均由RNN变化而来。

本研究选用LSTM、GRU、TCN 3种经典神经网络模型及基于核的算法的SVM模型对河流断面水质指标进行预测。

1.2 相关评价指标

在本文之后的试验中,通过预测值和真实值在拟合曲线上的接近程度,可以观察不同模型对水质数据的预测效果,但这种通过人眼直接观察的评价方式不够精准。因此,采用一些性能评价指标,从数值上可以更精准更严格地进行评价分析,所以在本试验中将选取相关的性能评价指标。在机器学习领域,对于分类和回归两大类问题都有很多可以用作衡量误差的评价指标。因为水质预测问题解决的是一个时序数据的问题,输入变量与输出变量均为连续变量,其本质是一个回归问题,在本试验中一共用到了3种评价指标,分别是平均绝对误差(MAE)、均方根误差(RMSE)、平均绝对百分比误差(MAPE),对不同模型的预测效果进行量化的评估,

这 3 个指标的计算如式(1)~式(3)。

$$M_{\mathrm{MAE}}=\frac{1}{n}\sum_{t=1}^{n}\left|Y_t-\hat{Y}_t\right| \tag{1}$$

$$R_{\mathrm{RMSE}}=\sqrt{\frac{\sum_{t=1}^{n}(Y_t-\hat{Y}_t)^2}{n}} \tag{2}$$

$$M_{\mathrm{MAPE}}=\frac{1}{n}\sum_{t=1}^{n}\left|\frac{Y_t-\hat{Y}_t}{Y_t}\right|\times 100\% \tag{3}$$

其中:M_{MAE}——MAE, mg/L;

R_{RMSE}——RMSE, mg/L;

M_{MAPE}——MAPE;

n——样本总数;

Y_t——真实值;

$\hat{Y}_t$——预测值。

前面指标能够反映预测值与真实值的偏离程度,任何一个指标数值越小,都表示得到的预测值越接近真实值,模型具有较高的预测精度。其中,RMSE 更多地反映真实值与预测值之间的离散程度,MAE 反映误差的平均幅值,而 MAPE 反映了真实值与预测值间误差占真实值的百分比,利用百分比的形式对误差的程度进行评价。

2　结果与讨论

2.1　河流重点断面水质

2.1.1　断面选择

J 河流域共设置了 3 个例行监测断面,分别为 A 断面(省控入江支流)、B 断面(流域补偿监测断面)、C 断面(市级考核断面)。其中,省控入江支流 A 断面和流域补偿 B 断面监测频次为一月一次,C 断面监测频次为两月一次。按照不同断面的重要性排序,且考虑对应监测频次和数据量大小,本研究以断面 A 为主要对象,以断面 B 为次要对象。

此外,为了全面准确掌握断面不同时间的水质动态变化情况,从 2019 年 5 月起,在断面 A 设置自动监测站点,监测因子为氨氮、溶解氧(DO)、电导率、化学需氧量(COD)、pH 和总磷(TP)共 6 项指标,监测频率为 4 h/次,自动监测数据用于后续机器学习水质模拟。

2.1.2　断面分析

根据 2013 年—2018 年统计 J 河 A 断面水质,按地表水Ⅴ类标准进行评价。这 6 年间,J 河 A 断面水质逐年均值为劣Ⅴ类,6 年均值为劣Ⅴ类,主要超标因子为氨氮和 TP,COD 个别值超标。6 年 COD 均值为 20.9 mg/L,未超过Ⅴ类标准水质,点次超标率为 7%;氨氮均值为 6.7 mg/L,点次超标率为 87%;TP 均值为 0.7 mg/L,点次超标率为 88%。具体而言,A 断面 COD 浓度总体可达Ⅴ类标准,逐年均值及长期均值均小于 25 mg/L;从单次监测情况来看,A 断面 COD 浓度仅有 4 次劣于Ⅴ类,且在 2016 年 1 月之后均未超过 40 mg/L,其余均达标。A 断面氨氮浓度基本无法达到Ⅴ类标准,逐年均值及长期均值均大于 5 mg/L;从单次监测情况来看,A 断面氨氮浓度在仅有 7 次达到Ⅴ类,其余点次均超标。同样,A 断面 TP 浓度也基本无法达到Ⅴ类标准,逐年均值及长期均值均大于 0.5 mg/L;从单次监测情况来看,A 断面 TP 浓度仅有 8 次达到Ⅴ类,其余点次均超标。

经过整治后,A 断面的水质得到有效改善。图 1 显示了 2020 年 J 河每月监测数据,其中 DO 和 COD 含量波动较大,其余指标如氨氮、TP 等相对稳定。同时,从数值上看,DO[(8.6±2.5)mg/L]、高锰酸盐指数[(3.2±0.4)mg/L]、COD[(11.7±2.8)mg/L]、BOD_5[(2.5±0.4)mg/L]、氨氮[(0.49±0.26)mg/L]、和 TP(0.11±0.04)mg/L]均稳定达到《地表水环境质量标准》(GB 3838—2002)中Ⅲ类水的标准,表明 A 断面水质情况良好,J 河水质治理有所成效。

2.2　影响因子识别

2.2.1　基于每月例行监测的分析

根据 J 河 A、B 断面的每月例行监测数据绘制主成分分析(PCA)图(图 2),每个水质指标对应的箭头在主成分轴上的投影表示其影响程度。氨氮、TP、DO 和 pH 是影响两断面的主要因子。其中,根据不同因子对应箭头之间的夹角判断相关性水平,可知氨氮和 TP 呈显著正相关,而氨氮和 DO 及 pH 呈显著负相关。该方法可初步筛查出影响断面水质的主要因子,但仍须缩小选择范围。

此外,不同断面某一指标差别的显著性也能侧面反映该指标对整体水质影响的大小。同样取 J 河 A 和 B 两断面 2020 年月检测数据进行差异分析,结果表明,与 A 断面相比,B 断面的水质指标存在一定差异,但差异规律不尽相同(图 3)。例如,从 pH、

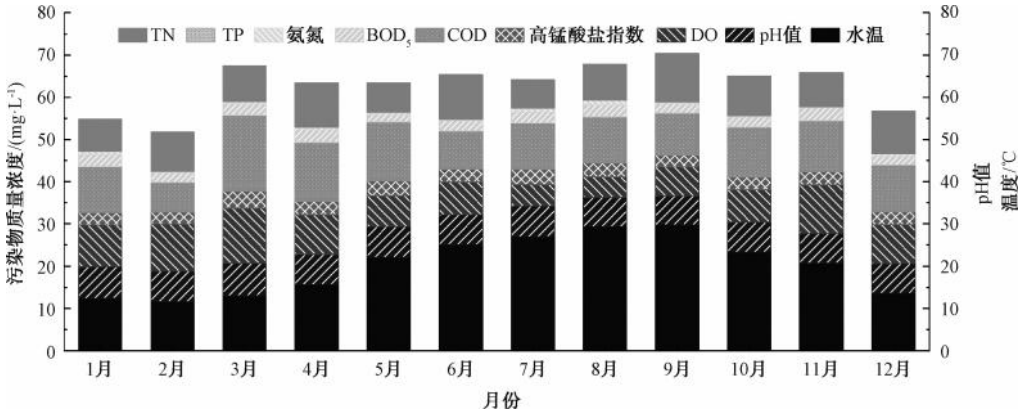


图 1　2020 年 1 月—12 月 J 河 A 断面水质指标监测值

Fig. 1　Monitoring Values of Water Quality Indices for Section A of River J from January to December in 2020

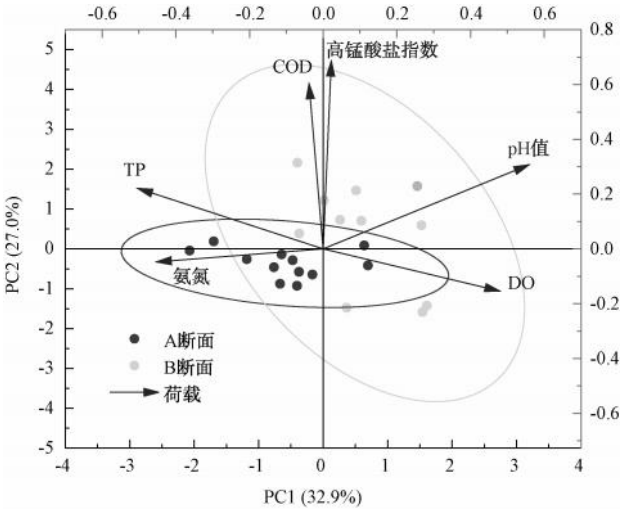
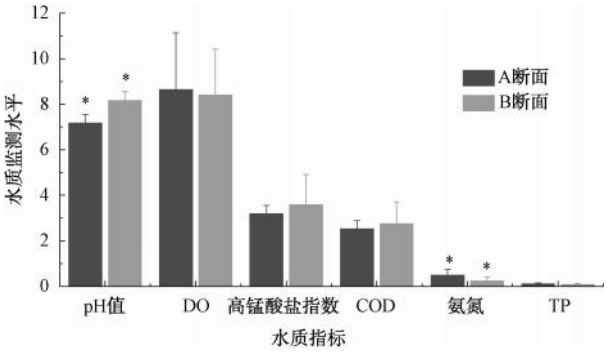


图 2　不同断面水质指标 PCA

Fig. 2　PCA of Water Quality Indices in Different Sections

高锰酸盐指数、BOD_5 和 DO 来看，A 断面水质优于 B；而 B 断面的氨氮和 TP 含量更低，表明该断面这 2 种指标更优。同时，由方差分析（ANOVA）结果可知（表 1），两断面 6 种指标中，仅 pH 和氨氮存在显著性差异（$P<0.05$），表明两者是不同断面水质的主要影响因素。



注：* 表示存在显著差异。

图 3　2020 年 J 河不同断面水质指标监测值比较

Fig. 3　Monitoring Values Comparison of Water Qualities Indices for Different Sections of River J in 2020

表 1　2020 年 J 河不同断面水质指标监测值差异显著性

Tab. 1　Significance of Differences in Monitoring Values of Water Quality Indices in Different Sections of River J in 2020

水质指标	pH	DO	高锰酸盐指数	BOD_5	氨氮	TP
显著性	0.000*	0.804	0.337	0.443	0.008*	0.055

注：* 表示存在显著差异

2.2.2　基于断面自动监测的分析

基于断面每月例行监测的数据值有限，可能降低分析结果的精确程度，故选取 J 河 A 断面自动监测（每 4 h 1 次）数据中的 6 种水质指标（COD、TP、氨氮、DO、电导率和 pH）进行主成分分析，并采用最大化正态方差法进行旋转，最终得到主成分贡献率和因子载荷如表 2 所示。结果呈现 3 个主成分，其中第一主成分贡献率最高，为 25.408%，在第 1 主成分中，明显看出氨氮的因子载荷数值最大（0.873），可以认为第 1 主成分和氨氮指标密切相关，因此氨氮是 J 河水环境影响因子中相对最为主要的一种评价因子。此外，第一主成分 DO 的因子载荷绝对值为 0.831，仅次于氨氮，则可认为 DO 为除氨氮外另一主要评价因子。

表 2　主成分贡献率及因子载荷

Tab. 2　Principal Component Contribution Rates and Factor Loading

指标	第 1 主成分	第 2 主成分	第 3 主成分
COD	-0.042	0.028	0.727
TP	0.075	-0.051	0.647
氨氮	0.873	0.182	-0.111
DO	-0.831	0.256	-0.178
电导率	0.014	0.772	0.160
pH 值	0.050	-0.756	0.199
特征值	1.524	1.256	1.009
贡献率	25.408%	20.936%	16.819%
累计贡献率	25.408%	46.344%	63.163%

综合上述 3 种评价因子分析比选,考虑到不同水质的主成分荷载值及对不同断面影响的显著性,以下选择氨氮和 DO 两指标对水质情况进行模拟预测。

2.3　水质预测预警

2.3.1　水质预测结果

氨氮是首要影响因子。图 4 是 4 种模型对于氨氮水平预测的结果,其中实线代表真实数据,虚线为机器学习模型预测到的水质数据。

观察氨氮的数据可以发现,其数值一般为 0~4 mg/L,对于少部分的时间点出现较大波动情况,会有接近 10 mg/L 出现。考虑现实状况可能是因为污水排放导致氨氮数据突增,具有一定的合理性。如图 4(a)~图 4(d)所示,真实值和预测值曲线均十分接近,甚至重合,但因其比例尺原因,较小的误差显示不够直观。在测试集中稳定变化的数据段上,可以发现相较于其他模型 LSTM 对于稳定变化的数据预测存在较明显的误差,但总体来说这种误差也相对比较平稳且在能够接受的范围内。对于异常点数据的预测,GRU、SVM 模型中存在对这些点预测值与真实值相差较大的情况,LSTM 在这方面的预测表现却明显优于其余 3 个模型,而 SVM 对于异常值的预测效果相对最差。这表明 LSTM 模型可以学习到数据中异常数据的发生,对其具有较好的预测能力。

随后,进行 DO 水平预测试验。图 5 是 4 种模型对于 DO 水质预测的结果,分别对真实值和预测值数据进行绘制。

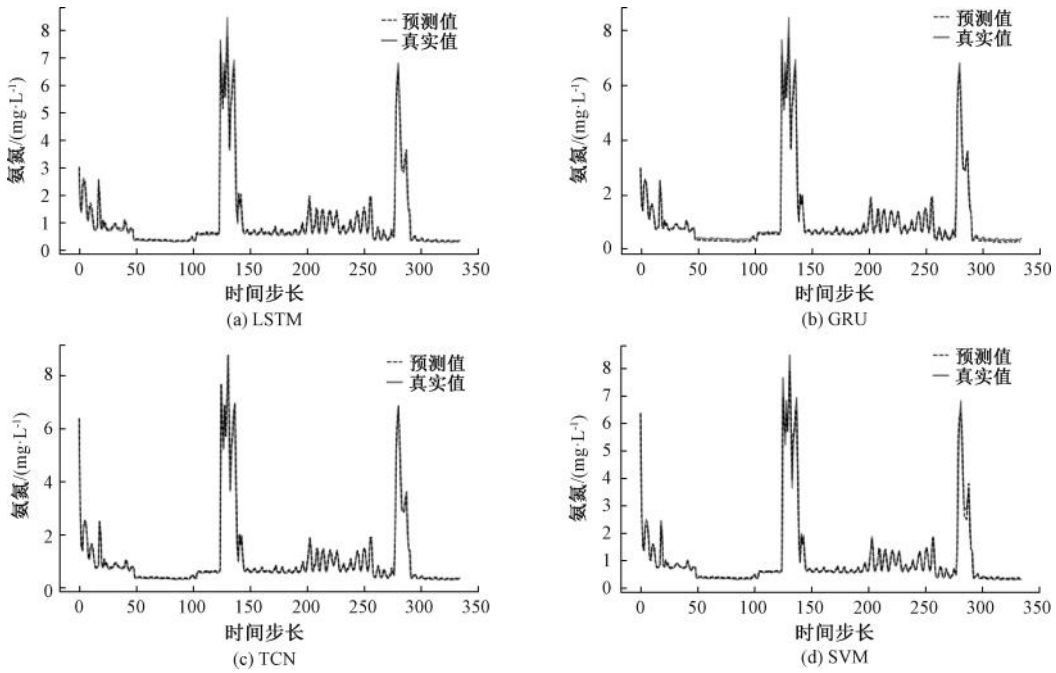


图 4　4 种模型下氨氮水平预测

Fig. 4　Prediction of Ammonia Nitrogen Levels under Four Models

观察上面的拟合曲线图,DO 水平呈现出频繁上下振荡变化,数值波动较为显著。且总体数值呈下降趋势,在前半时间段较高,在后半段较低,这符合测试集的数据从 2 月—5 月的变化,即因气温升高,可能会使水中 DO 含量有所降低。DO 作为化学水质指标,是指溶解在水中的氧气,也是水中植物动

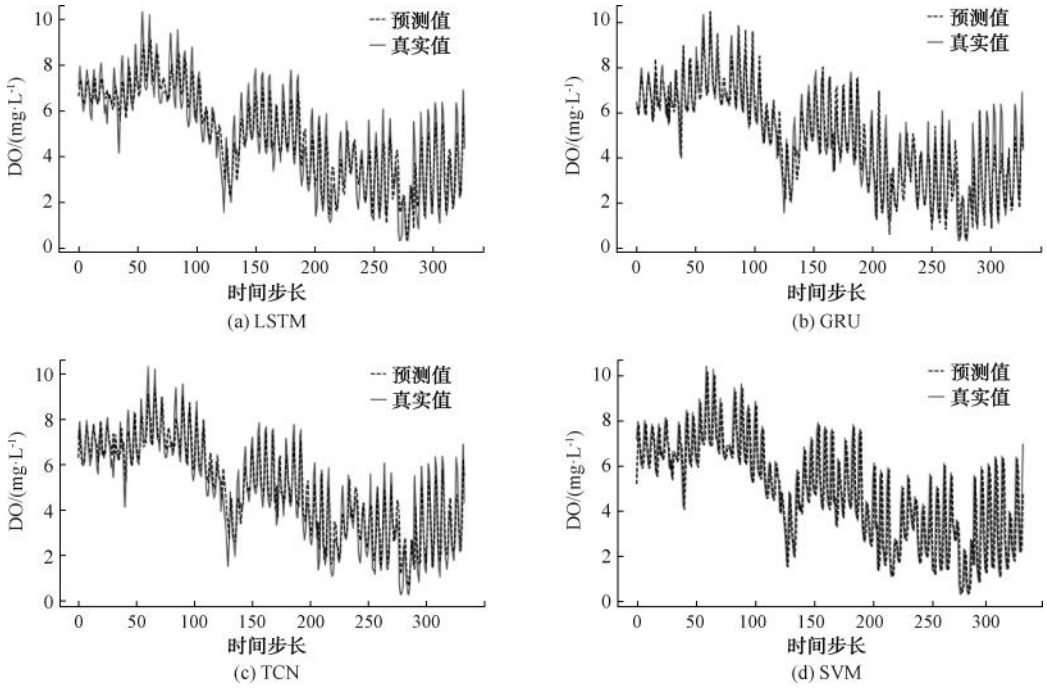


图 5　4 种模型下 DO 水平预测

Fig. 5　Prediction of DO Levels under Four Models

物生长必不可少的要素,它也反映水质的污染程度[7]。在曲线中,预测值和真实值两曲线拟合效果并不理想,曲线重合程度小于对氨氮水平的预测。综合数据曲线图和评价指标统计图来看,GRU 模型在 DO 水平预测上占有绝对的优势,在拟合图中虽然真实值和预测值较为参差,但数据的走势预测基本吻合,且能够对一些较大变化的数据点进行预测,说明 GRU 模型可以有效利用数据的时序性,获得较好的预测效果。而 SVM 模型的表现不是很理想,预测得到的值都略低于真实值,不适合应用于 DO 水平预测。

2.3.2　模型可行性分析

通过不同评价指标的具体数值可直观、准确地评估各模型的准确性及可行性(表 3)。在 4 种模型中,GRU 模型在 MAE、MAPE 指标上都获得了最好的表现,且 RMSE 数值也较低;而对于 SVM 模型,其 RMSE 指标值最低,且具有较低的 MAE 和 MAPE 值。

如图 6(a)所示,在氨氮水质数据集上,GRU 模型和 SVM 模型预测效果较佳。从 MAE 和 RMSE 指标来看,GRU 和 SVM 模型效果更好;在 MAPE 指标的表现上,GRU 模型相较于其他模型特别是 LSTM 优势明显,表明 GRU 模型在数据较小的氨氮数据集上,预测的效果较好。LSTM 模型除 MAPE 较大之外,另 2 个评价指标也不占优势,相对来说其在氨氮水平上的预测能力较弱。此外,TCN 模型的表现比较平均,具有较好的 MAPE 值,虽然 RMSE 是最大值,表明模型在个别值上预测表现相对较差,但其值也相对合理,同样具有不错的预测能力。

表 3　模型评价指标综合统计

Tab. 3　Comprehensive Evaluation of Models

指标	模型	MAE/(mg·L^{-1})	MAPE	RMSE/(mg·L^{-1})
氨氮	LSTM	0.225 77	26.122%	0.574 05
	GRU	0.185 31	14.148%	0.564 85
	TCN	0.204 00	18.131%	0.588 70
	SVM	0.187 41	19.987%	0.533 28
DO	LSTM	0.717 27	16.256%	0.912 41
	GRU	0.623 38	13.620%	0.830 30
	TCN	0.742 13	19.840%	1.040 42
	SVM	0.897 84	18.428%	1.194 38

由图 6(b)所示,在 DO 水质数据集上,GRU 模型和 LSTM 模型预测效果较佳。GRU 和 LSTM 模型的 MAE、MAPE、RMSE 3 种评价指标数值最低,表明 GRU 和 LSTM 模型对 DO 的模拟效果优于其他两种模型,且 GRU 模型优势明显。LSTM 模型在该数据

集上获得了相较于 GRU 较弱,但略胜于其他两个模型的预测能力。而不同于氨氮预测结果,SVM 模型的 MAE 和 RMSE 数值均最高,而 MAPE 数值也高于 GRU 和 LSTM 模型,说明 SVM 对 DO 水平预测能力较弱,预测精度相对较低。此外,TCN 模型的 MAPE 值在 4 种模型中最大,且其 MAE 和 RMSE 的值也都较大,说明其预测能力在 DO 数据上较为一般。

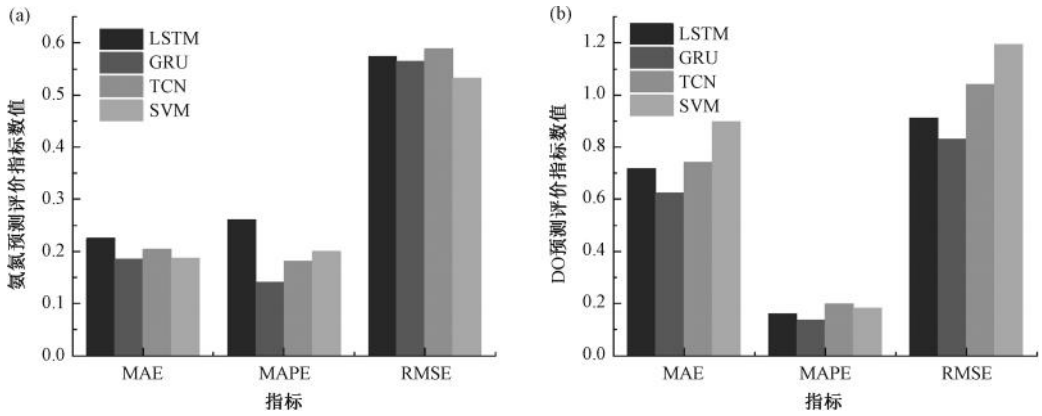


图 6　4 种模型对(a)氨氮和(b)DO 的预测评价指标结果

Fig. 6　Results of Predictive Evaluation Indices of Four Models for (a) Ammonia Nitrogen; (b) DO

3　结论

本文首先基于 J 河 A、B 断面的每月例行监测数据(pH、DO、高锰酸盐指数、BOD、氨氮、TP)进行分析,筛选出主要影响因子,随后基于 J 河 A 断面的自动监测数据进行主成分分析,最终确定影响断面水质的主要因子氨氮和 DO 进行后续水质预测。选取 LSTM、GRU、TCN 和 SVM 4 种经典模型进行水质模拟试验,结果表明 4 种模型对于氨氮和 DO 的预测都有较好的效果,其中 GRU 模型对氨氮和 DO 的预测最具优势,此外 SVM 模型对氨氮的预测和 LSTM 模型对 DO 的预测也具有相对优势。总之,经典的机器学习模型对于水质指标预测能力较强,且精度较高,故机器学习在水质指标预测领域具有可行性。

参考文献

[1] YAHYA A S A, AHMED A N, OTHMAN F B, et al. Water quality prediction model based support vector machine model for ungauged river catchment under dual scenarios [J]. Water, 2019, 11(6): 1231. DOI: 10.3390/w11061231.

[2] LI W Y, WU H, ZHU N Y, et al. Prediction of dissolved oxygen in a fishery pond based on gated recurrent unit (GRU) [J]. Information Processing in Agriculture, 2021, 8(1): 185-193.

[3] LIU P, WANG J, SANGAIAH A K, et al. Analysis and prediction of water quality using LSTM deep neural networks in IoT environment [J]. Sustainability, 2019, 11(7): 2058.

[4] HU Z H, ZHANG Y R, ZHAO Y C, et al. A water quality prediction method based on the deep LSTM Network Considering Correlation in Smart mariculture [J]. Sensors, 2019, 19(6): 1420. DOI: 10.3390/s19061420.

[5] ZHANG Y F, THORBURN P J, FITCH P. Multi-task temporal convolutional network for predicting water quality sensor data [C]//International Conference on Neural Information Processing, Springer, 2019: 122-130.

[6] XU J, WANG K, LIN C, et al. FM-GRU: A time series prediction method for water quality based on seq2seq framework [J]. Water, 2021, 13(8): 1031. DOI:10.3390/w13081031.

[7] 陈前, 唐文忠, 许妍, 等. 基于溶解氧和耗氧污染物变化的长江流域水质改善过程分析(2008—2018 年)[J]. 环境工程学报, 2023, 17(1): 279-287.
CHEN Q, TANG W Z, XU Y, et al. Recovery process analysis of water quality in the Yangtze River Basin based on changes of dissolved oxygen and oxygen-consuming substances (2008—2018) [J]. Chinese Journal of Environmental Engineering, 2023, 17(1): 279-287.