基于多模型融合的数据预测方法研究

黄欣欣[1] 唐乐红[1*] 黄长泉[2] 陈海玥[1] 叶 晨[1]
1 阳光学院，福建 福州 350015
2 中科数创（厦门）智能科技研究院，福建 厦门 361015

[摘要] 针对单一预测模型难以同时刻画数据中的线性趋势、非线性关系、特征交互和长期时序依赖，且在样本波动、异常扰动与分布变化条件下预测稳定性不足的问题，提出一种基于异质基学习器和二层 Stacking 的数据预测方法。首先，对原始数据进行缺失值插补、异常值修正、标准化和滑动窗口重构；其次，选择自回归积分滑动平均模型、支持向量回归、随机森林和长短期记忆网络作为基学习器，分别提取线性、非线性、组合特征及时序依赖信息；再次，采用时序交叉验证生成折外预测值，并将基模型预测结果、关键原始特征及误差统计量共同输入 XGBoost 元学习器，实现非线性加权与残差校正；最后，在包含 8760 条样本的可控仿真数据集上进行验证。结果表明，所提模型的平均绝对误差、均方根误差和平均绝对百分比误差分别为 2.41、3.18 和 2.76%，决定系数达到 0.968，整体性能优于各单一模型及简单加权融合模型。研究表明，基于模型差异性和无泄漏训练机制的多模型融合方法能够提高复杂数据预测的精度、鲁棒性与泛化能力。

[关键词] 多模型融合；数据预测；Stacking；长短期记忆网络；XGBoost
DOI：10.64635/ja.2026.1445 中图分类号：TP181 文献标识码：A

Research on a Data Prediction Method Based on Multi-Model Fusion

Huang Xinxin[1], Tang Lehong[1]*, Huang Changquan[2], Chen Haiyue[1], Ye Chen[1]
1 Yango University, Fuzhou, Fujian 350015, China
2 Zhongke Shuchuang (Xiamen) Intelligent Technology Research Institute, Xiamen, Fujian 361015, China

Abstract: To address the difficulty of a single prediction model in simultaneously characterizing linear trends, nonlinear relationships, feature interactions, and long-term temporal dependencies in data, as well as its insufficient prediction stability under sample fluctuations, abnormal disturbances, and distribution changes, this paper proposes a data prediction method based on heterogeneous base learners and two-layer stacking. First, missing value imputation, outlier correction, standardization, and sliding-window reconstruction are performed on the original data. Second, the autoregressive integrated moving average model, support vector regression, random forest, and long short-term memory network are selected as base learners to extract linear, nonlinear, combined-feature, and temporal-dependency information, respectively. Third, time-series cross-validation is used to generate out-of-fold predictions, and the prediction results of the base models, key original features, and error statistics are jointly input into the XGBoost meta-learner to achieve nonlinear weighting and residual correction. Finally, validation is conducted on a controllable simulation dataset containing 8,760 samples. The results show that the proposed model achieves a mean absolute error, root mean square error, and mean absolute percentage error of 2.41, 3.18, and 2.76%, respectively, with a coefficient of determination reaching 0.968, and its overall performance is superior to that of individual models and simple weighted fusion models. The study indicates that a multi-model fusion method based on model diversity and a leakage-free training mechanism can improve the accuracy, robustness, and generalization ability of complex data prediction.

Keywords: multi-model fusion; data prediction; stacking; long short-term memory network; XGBoost

引言

随着业务系统持续产生高频、多维和强关联数据，预测任务已由规则相对稳定的单变量外推，转向具有非平稳、非线性、噪声扰动和概念漂移特征的复杂建模。自回归积分滑动平均模型等统计模型具有结构清晰、参数可解释的优点，但对非线性关系的表达能力有限；支持向量回归和树模型能够刻画复杂映射，却容易忽略序列中的长期依赖；深度时序网络具有较强的特征学习能力，但训练成本较高，且容易受到样本规模、超参数和随机初始化的影响。已有研究通过统

计模型与深度学习模型的融合，提高了时间序列预测的信息利用能力和预测精度[1]；针对非线性、非平稳时序数据，融合时间卷积、残差结构与注意力机制也表现出较好的预测精度和泛化能力[2]。

现有研究多面向特定应用领域，部分方法采用一次性训练集划分，容易在二层学习中产生信息泄漏；同时，常见融合方式主要使用基模型输出，未充分利用原始关键特征及模型误差状态。基于此，本文提出一种“异质基模型 - 时序折外预测 - 融合特征增强 -XGBoost 元学习”的两层预测方法，在保证训练过程无信息泄漏的基础上，实现不同预测模型之间的优势互补。相关研究表明，将基学习器构造特征与原始关键特征进行融合，能够进一步提高 Stacking 模型的预测精度[3]；Bi-LSTM 与 XGBoost 组合则可兼顾时序依赖建模和非线性特征提取[4]。

1 多模型融合预测的基本原理

1.1 问题定义与数据处理

设按照时间顺序获得样本集：

$$D = \{(x_t, y_t)\}_{t=1}^{N}$$

式中，x_t 为第 t 个时刻的多维输入特征向量，y_t 为对应的真实预测目标，N 为样本总数。数据预测的基本任务是学习映射函数 F，使预测值尽可能接近真实值：

$$\hat{y}_t = F(x_t)$$

$$min_F \frac{1}{N}\sum_{t=1}^{N} L(y_t, F(x_t))$$

式中，L(•) 为损失函数，可根据预测任务选用均方误差或平均绝对误差。

考虑实际业务数据中普遍存在的缺失、异常和量纲差异问题，首先采用相邻时段中位数完成缺失值插补；利用四分位距方法识别异常点，并以局部滑动中位数对异常数据进行修正；随后对连续变量实施 Z-score 标准化处理：

$$x'_{t,j} = \frac{x_{t,j} - \mu_j}{\sigma_j}$$

式中，$x'_{t,j}$ 为第 t 个时刻第 j 个特征的标准化值，$x_{t,j}$ 为原始值，μ_j 和 σ_j 分别为第 j 个特征在当前训练集或训练折上的均值与标准差。上述统计量仅由训练数据计算，并应用于相应验证集和测试集，以避免预处理阶段产生信息泄漏。对于序列型特征，利用长度为 L 的滑动窗口构造输入序列，以保留数据的近期变化、周期波动和滞后信息。多尺度窗口与膨胀卷积结构能够进一步增强不同时间尺度特征的提取能力。

1.2 Stacking 融合机制

设第 m 个基模型对第 t 个样本的预测结果为 $p_{t,m}$，由 M 个基模型得到的预测结果构成向量：

$$P_t = [p_{t,1}, p_{t,2}, \ldots, p_{t,M}]^T$$

传统加权融合模型可以表示为：

$$\hat{y}_t^w = \sum^{M} w_m p$$

其中，w_m 为第 m 个基模型的权重。为保证各权重具有可解释性，通常设置如下约束：

$$\sum_{m=1}^{M} w_m = 1,\ w_m \geq 0,\ m = 1, 2, \ldots, M$$

静态权重难以反映不同数据状态下各模型预测性能的变化。本文采用 XGBoost 作为元学习器，建立非线性融合函数。已有 Stacking 应用研究表明，两层集成框架可通过元模型学习多个基模型的互补信息，提升预测泛化能力[5]。

$$\hat{y}_t = G(P_t, x_t^*, e_t)$$

式中，x_t* 表示经过筛选的关键原始特征向量，e_t 表示各基模型在最近观测窗口内形成的误差统计向量，包括平均绝对误差、残差均值和残差方差。元学习器可依据当前数据状态自动调整各基模型的预测贡献，并对基模型产生的系统性残差进行二次修正。

2 基于多模型融合的数据预测方法

2.1 总体框架

本文方法包括数据预处理、基模型训练、折外预测生成、融合特征构造、元学习器训练和结果评估六个环节，具体流程见图 1。

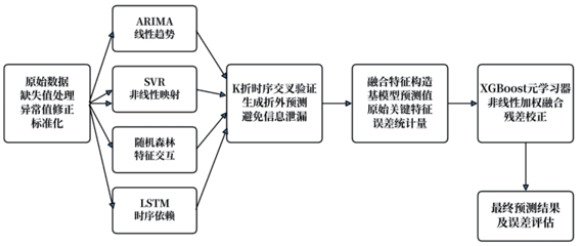

图 1 基于异质基学习器与二层 Stacking 的多模型融合预测流程

在第一层中，ARIMA、SVR、随机森林和 LSTM 分别从不同角度学习数据特征；在第二层中，利用 XGBoost 综合基模型预测结果、原始关键特征和误差

统计量，生成最终预测结果。

2.2 无泄漏折外预测

若直接使用基模型在训练集上的拟合结果训练元学习器，二层模型将接收到过度乐观的预测数据，容易形成信息泄漏，导致模型在测试集或新业务数据上的性能下降。

为解决该问题，本文采用 5 折扩展窗口时序交叉验证：第 k 折仅使用当前验证区间之前的数据训练基模型，并对当前验证区间进行预测。完成全部折次后，将各验证区间的预测结果按时间顺序拼接，形成完整的折外预测矩阵。已有研究也采用扩展窗口 K 折时序交叉验证，以保持样本时间先后关系并提高预测稳定性和泛化能力[6]。

测试阶段使用全部训练数据重新训练各基模型，再对测试数据进行预测。该方法保持了时间序列的先后关系，能够避免未来数据进入历史训练过程，使元学习器学习到更加接近真实应用环境的基模型输出。

2.3 模型训练与结果输出

模型训练过程中，ARIMA 的阶数依据赤池信息准则确定；SVR 通过网格搜索选择惩罚系数和径向基核参数；随机森林设置 200 棵回归树，并限制单棵树的最大深度；LSTM 采用两层网络结构，隐藏单元数分别设置为 64 和 32，同时使用 Dropout 和早停策略抑制过拟合。

元学习器采用最大深度为 3 的 XGBoost 回归树，学习率设置为0.05,并根据验证集误差确定迭代次数。最终模型除输出目标变量的点预测值外，还持续计算预测残差。当融合模型连续多个时点的预测误差超过训练期误差上界时，触发滚动训练或模型参数更新，以适应数据分布变化。

3 实验与结果分析

3.1 数据与实验设置

为检验方法有效性，构建包含 8760 条小时级样本的可控仿真数据集。目标变量由线性趋势、日周期、周周期、非线性特征交互、历史滞后项和随机噪声共同生成；输入变量包含时间编码、历史目标值、外部影响变量及类别状态等 12 项特征。

在仿真数据中随机设置 2% 的缺失值和 1% 的异常点，用于检验数据预处理及模型抗干扰能力。数据按照时间顺序划分为训练集、验证集和测试集，占比分别为 70%、15% 和 15%。需要说明的是，该数据仅用于方法有效性验证，不代表特定企业或行业的真实经营数据。

模型评价采用平均绝对误差（MAE）、均方根误差（RMSE）、平均绝对百分比误差（MAPE）和决定系数（R^2）。设测试样本数为 n，真实值为 y_i，预测值为 $\hat{y}_i$，真实值均值为 $\bar{y}$，则各指标计算如下：

$$MAE = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$$

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

其中，MAE 反映平均绝对偏差，RMSE 对较大误差更敏感，MAPE 衡量相对误差，R^2 评价模型对数据波动的解释程度。当真实值包含 0 时，MAPE 计算中应使用极小正数 ε 替代零分母。

表 1 不同预测模型的实验结果

模型	MAE	RMSE	MAPE/%	R^2
ARIMA	4.68	6.15	5.34	0.884
SVR	3.76	4.92	4.21	0.923
随机森林	3.29	4.37	3.74	0.939
LSTM	2.96	3.91	3.38	0.950
简单加权融合	2.74	3.61	3.11	0.957
本文融合模型	2.41	3.18	2.76	0.968

3.2 实验结果分析

由表 1 可知，ARIMA 对数据周期和总体趋势具有一定拟合能力，但无法充分描述复杂的非线性交互关系，因此各项误差相对较大。SVR 和随机森林的预测精度明显提高，说明非线性映射与特征组合对复杂数据预测具有重要作用。

LSTM 在四个基模型中表现最好，其 MAE 为 2.96，RMSE 为 3.91，说明序列中的滞后关系和周期依赖是影响预测结果的重要因素。但 LSTM 仍可能受到局部异常、样本规模和训练随机性的影响，单独使用时难

以保证不同数据状态下的稳定性。

简单加权融合模型优于全部单一模型，但其 MAPE 仍为 3.11%。这说明固定权重虽然能够降低依赖单一模型造成的风险，却不能根据数据状态动态调整不同模型的贡献。

本文融合模型的 MAE 和 RMSE 分别较 LSTM 降低 18.58% 和 18.67%，MAPE 下降 0.62 个百分点，决定系数由 0.950 提高至 0.968。其性能改善主要来自三个方面。

一是 ARIMA、SVR、随机森林和 LSTM 分别学习线性趋势、非线性映射、特征交互和长期时序依赖，降低了不同模型预测偏差的同质性。

二是时序折外预测避免了元学习器直接使用基模型训练集拟合结果，降低了信息泄漏和过拟合风险。

三是融合阶段引入关键原始特征和近期误差统计量，使元学习器能够识别不同运行状态下更加可靠的基模型，并对基模型的系统性偏差进行修正。

4 结论

本文针对单一预测模型难以充分适应复杂数据结构的问题，提出一种基于 ARIMA、SVR、随机森林、LSTM 和 XGBoost 的二层 Stacking 预测方法。该方法利用异质基学习器提取不同类型的信息，通过时序交叉验证构建无泄漏的折外预测结果，并将基模型输出、关键原始特征和误差统计量共同输入元学习器。

可控仿真实验结果表明，本文模型在 MAE、RMSE、MAPE 和决定系数等指标上均优于单一模型及简单加权融合模型，能够有效提高预测精度、稳定性和泛化能力。研究结果可为实时业务监控、市场需求预测、设备运行调度和风险预警等数据预测任务提供方法参考。

[参考文献]

[1] 张建勋，胡少杰，芦丽旭，等．多模型融合的时间序列数据预测方法 [J]．西安邮电大学学报，2025,30(1):115-122.

[2] 周卓辉，杨欢，刘小芳．基于时间模式注意力机制和改进 TCN 的 PM2.5 浓度预测方法 [J]．无线电工程，2024,54(10):2315-2324.

[3] 李浩，卢朝阳，谈翌平，等． 基于 MIC-iAFF-Stacking 集成学习的航空器滑出时间预测 [J]．交通运输工程与信息学报，2024,22(4):142-153.

[4] 代业明，周琼．基于改进 Bi-LSTM 和 XGBoost 的电力负荷组合预测方法 [J]．上海理工大学学报，2022,44(2):138-147.

[5] 康文豪，徐天奇，王阳光，等．基于 CEEMDAN-精细复合多尺度熵和 Stacking 集成学习的短期风电功率预测 [J]．水利水电技术（中英文），2022,53(2):163-172.

[6] 仲浩帆，黎雅红，朱恩豪．基于模型组合的电力负荷精准预测 [J]．自动化应用，2024,65(7):223-229.