

基于核技巧改进的Informer模型的长序列时间序列预测方法

潘立群, 吴中华, 洪标

引用本文

潘立群, 吴中华, 洪标. 基于核技巧改进的Informer模型的长序列时间序列预测方法[J]. 计算机科学, 2023, 50(11A): 221100186-6.

PAN Liqun, WU Zhonghua, HONG Biao. Prediction Method of Long Series Time Series Based on Improved Informer Model with Kernel Technique [J]. Computer Science, 2023, 50(11A): 221100186-6.

相似文章推荐 (请使用火狐或 IE 浏览器查看文章)

Similar articles recommended (Please use Firefox or IE to view the article)

[基于深度学习的红外视频显著性目标检测](#)

Deep Learning Based Salient Object Detection in Infrared Video

计算机科学, 2023, 50(9): 227-234. <https://doi.org/10.11896/jsjx.220700204>

[基于度量正则化的红外与可见光跨模态行人重识别算法](#)

Metric Regularized Infrared and Visible Cross-modal Person Re-identification

计算机科学, 2023, 50(6A): 221100046-8. <https://doi.org/10.11896/jsjx.221100046>

[知识图谱嵌入模型中的损失函数研究综述](#)

Comprehensive Survey of Loss Functions in Knowledge Graph Embedding Models

计算机科学, 2023, 50(4): 149-158. <https://doi.org/10.11896/jsjx.211200175>

[基于迁移学习和多视图特征融合提高RNA碱基相互作用预测](#)

Improving RNA Base Interactions Prediction Based on Transfer Learning and Multi-view Feature Fusion

计算机科学, 2023, 50(3): 164-172. <https://doi.org/10.11896/jsjx.211200186>

[基于改进区域候选网络的场景文本检测](#)

Scene Text Detection with Improved Region Proposal Network

计算机科学, 2023, 50(2): 201-208. <https://doi.org/10.11896/jsjx.211000191>

基于核技巧改进的 Informer 模型的长序列时间序列预测方法

潘立群¹ 吴中华¹ 洪 标²

1 上海大学管理学院 上海 200444

2 上海对外经贸大学国际经贸学院 上海 201620

摘 要 如今,学者们对长序列时间序列问题的预测主要基于类 RNN 模型,且其中大部分使用的损失函数是传统的均方误差(MSE)。但类 RNN 模型在预测任务中存在只能捕捉局部信息且计算开销会随着预测序列的增多迅速提升的问题。不仅如此,MSE 损失函数无法捕捉长时间序列数据中普遍存在的非线性问题,且自身还存在对异常值敏感和鲁棒性较低的问题。基于以上背景,提出一种完全基于注意力机制的 Informer 模型,并在模型中使用基于核技巧改进的 Kernel-MSE 损失函数代替传统的 MSE 损失函数来解决长序列时间序列预测的问题。在多变量预测多变量的背景下,以 3 类数据中的 8 份数据集为例,对比改进后的 Informer 模型与经典的 Informer 模型,类 RNN 模型中的 LSTM 和 GRU 模型。结果表明,改进后的 Informer 模型预测精度更高,且精度的相对提升值随着原始数据量的增大而增大,适用于长序列时间序列预测问题。

关键词: Informer 模型;损失函数;核技巧;长序列时间序列预测

中图法分类号 TP181

Prediction Method of Long Series Time Series Based on Improved Informer Model with Kernel Technique

PAN Liqun¹, WU Zhonghua¹ and HONG Biao²

1 School of Management, Shanghai University, Shanghai 200444, China

2 School of International Business and Economics, Shanghai University of International Business and Economic, Shanghai 201620, China

Abstract Nowadays, the prediction of long sequence time series problems is mainly based on RNN like models, and most of the loss functions used are mean square error(MSE). However, MSE loss function can not capture the nonlinear problems commonly existing in long time series data. Moreover, MSE loss function itself is sensitive to outliers and has low robustness. Therefore, this paper proposes to use the improved Kernel MSE loss function based on kernel technique to replace the traditional MSE loss function in Informer model, and solve the nonlinearity in data by mapping the error from the original feature space to a higher dimensional space. Moreover, the first and second derivatives of the new loss function ensure robustness to outliers. Under the background of multivariable prediction and multivariable, this paper compares the prediction accuracy with the classical Informer model, LSTM model and GRU model, taking eight data sets in three types of data as examples. The results show that the improved Informer model has higher prediction accuracy, and the relative improvement value of accuracy increases with the increase of the original data volume, which is suitable for the prediction of long series time series.

Keywords Informer, Loss function, Kernal trick, Long series time series prediction

1 引言

长序列时间序列预测是指利用大量关于过去行为的时间序列数据来进行长期预测。现有的时间序列预测模型大多是在短期问题设置下设计的,预测长度为 48 个单位或更少。预测长度的增加使模型的预测能力不断变差,给数据管理和分析问题带来新的挑战。经典的类 RNN 模型 LSTM(Long Short-term Memory)在预测长度大于 48 个单位后, MSE 会大幅提升,并且模型的推理速度也会迅速下降。为了满足长时间序列预测需要的远程比对能力和输入和输出的有效操作,有学者注意到 Transformer 模型。Transformer 的本质是一个编码器-解码器(Encoder-Decoder)的结构,其与神经网络模型最大的不同是其完全基于注意力(attention)机制,注意力机制可以解决 LSTM 及其变体存在的长距离依赖问题,

可以有更好的记忆力,记住更长距离的信息;另外,注意力机制也支持并行化计算,相比 LSTM 模型其运算效率有大幅提升。Transformer 模型的自注意力机制和并行的性质显示出了解决长序列时间序列预测问题的巨大潜力^[1]。Li 等^[2]通过压缩其子层(Transformer 的基本构建块)来简化架构,实现了更高的并行性。Zhou 等^[3]提出了基于 Transformer 模型轴承 RUL 预测方法,利用了自注意力机制与编码器-解码器的结构优势,解决了预测序列过长的记忆力退化问题。Wu 等^[4]通过集合经验模式分解(EEMD)将一维原始风速序列转换为 16 维序列,并使用 Transformer 直接对多维风速数据进行建模,得到了更精确的预测结果。Zhou 等^[5]提出了一种考虑时空关联的多通道交通流预测方法(MCST-Transformer),在 Transformer 模型中引入空间关联模块对交通量进行预测,得到了精度更高的结果。但在以上的研究中,并没有学者

对 Transformer 模型在时间序列预测中的应用瓶颈进行深入的解决,这些瓶颈包括二次时间复杂度过高、高内存使用率以及编码器-解码器架构的固有限制。基于以上问题,Zhou 在 2021 年提出了一种基于 Transformer 的 Informer 模型。Informer 提出了一种稀疏的自注意力机制并实现了解码器的一次性输出,大大减少了模型运行时间并提升了模型预测的精度。结果证明,Informer 模型能有效提升长时间序列预测问题上的精度,预测精度远远高于当时所有的主流长序列时间序列预测模型^[1]。

在涉及深度学习的方法中,损失函数反映了优化过程中预测值与实际值之间的差异,并涉及深度学习模型的泛化能力和准确性,从而成为深度学习的基础^[6-7]。神经网络常用的损失函数分别为 L1(MAE)和 L2(MSE)损失函数。L1 损失函数可用于缓解时间序列数据中的异常,更具有鲁棒性。但 L1 损失函数在 0 点处的导数不连续,使得求解效率低下,导致收敛速度慢;不仅如此,L1 损失函数的梯度并不会随着损失值的大小而变化,所以不利于网络的学习^[8]。L2 损失函数的收敛速度快,能够对梯度给予合适的惩罚权重,使梯度更新的方向可以更加精确。但 L2 函数对异常值十分敏感,梯度更新的方向很容易受离群点所主导,不具备鲁棒性^[9]。不仅如此,传统的 L1 和 L2 损失函数仅反映了低维空间中数据的性质,无法准确评估数据中的非线性。在长时间时序预测中,序列中数据普遍存在的非线性特征通常被研究人员们忽略。为了捕获长时间序列中的非线性因素,可以在传统方法中利用核技巧。Chen 提出在使用神经网络模型预测风速时使用具有核技巧的损失函数。结果证明,新的损失函数能有效提高风速预测的精度^[10]。

在长时间序列预测问题中,随着收集的数据数量不断增加,原始数据的非线性几乎是普遍存在的。因此,在模型中有必要构建一种能衡量数据非线性的损失函数,以增强模型的鲁棒性。本文将在 Informer 模型中使用基于核技巧的 Kernel-MSE 损失函数代替传统的 MSE 损失函数。相比其他损失函数,Kernel-MSE 损失函数能更准确地衡量 Informer 模型中误差的非线性;不仅如此,Kernel-MSE 损失函数也解决了 MSE 损失函数对异常值过于敏感和不够鲁棒的问题。为了验证新损失函数在 Informer 中的有效性,本文还使用了传统的 Informer 模型、LSTM 和 GRU 模型来评估新模型的预测性能。文章的主要贡献如下:

(1)在长时序预测模型中针对数据普遍存在的非线性特征,提出将 Kernel-MSE 损失函数嵌入 Informer 模型中。Kernel-MSE 损失函数能通过将解映射到高维平滑空间来迭代更新,以更新最优解。此外,Kernel-MSE 损失函数对异常值的鲁棒性更强,增强了结果的稳定性。以前对长序列时间序列预测模型的研究主要集中在模型结构的改进和更好的特征提取上,新的损失函数可以为未来的长序列时间序列预测模型提供基础。

(2)在长序列时间预测中多变量预测多变量的背景下,本文以 3 类不同领域的不同大小的 8 套数据集为例,以考虑均方误差(MSE)和平均绝对误差(MAE),分别使用改进后的 Informer 模型、Informer 模型、LSTM 和 GRU 模型为基础进行实验。实验结果表明,基于核技巧改进的 Informer 模型的预测结果普遍更好,且该模型的相对预测精度会随着数据量

的增多不断提高。

2 相关概念

2.1 KERNEL-MSE 损失函数

2.1.1 传统的 MSE 损失函数

经典的 MSE 损失函数如下所示:

$$L_{mse} = \sum_{t=1}^N (y_t - \hat{y}_t)^2 \quad (1)$$

其中, y_t 为真实值, \hat{y}_t 为预测值。MSE 值越高,通常表示模型的拟合能力越差。误差值越高,预测性能越差。该损失函数处理线性空间中的误差,无法有效估计误差的非线性。此外,MSE 损失函数同等对待所有误差。因此,来自离群值的大误差会主导优化,误导优化方向。

2.1.2 KERNEL-MSE 损失函数

由于长时间序列预测的数据普遍存在的非线性问题,传统的深度学习模型无法在线性空间中直接评估非线性造成的误差。一个替代的方法是找到一个函数,该函数可以将误差从原始特征空间映射到更高维的空间,在非线性空间中量化非线性数据的预测误差。本文使用核函数的相关技术解决该问题。

假设 Φ 是从原始特征空间 R_n 到再生核希尔伯特空间 (RKHS) H 的非线性映射函数,形式为: $\Phi: R^n \rightarrow H$ 。内核空间中的内积可以定义如下:

$$K(y_i, \hat{y}_i) = (\Phi(y_i), \Phi(\hat{y}_i)) \quad (2)$$

RKHS 中数据点之间距离的平方可以用 Mercer 核表示,如下所示:

$$\begin{aligned} d_{H^2}^2(x, y) &= k(y_i, y_i) - 2k(y_i, \hat{y}_i) + k(\hat{y}_i, \hat{y}_i) \\ &= \Phi(y_i)\Phi(y_i) - 2\Phi(y_i)\Phi(\hat{y}_i) + \Phi(\hat{y}_i)\Phi(\hat{y}_i) \end{aligned} \quad (3)$$

常用的核函数包括高斯核函数、多项式核函数等。高斯核函数是欧几里德测度的径向展开,不改变相对数据位置。因此,我们在本研究中使用了高斯核函数。高斯核函数如下:

$$k(y_i, \hat{y}_i) = \exp\{-(y_i - \hat{y}_i)^2 / 2\sigma^2\} \quad (4)$$

由此,RKHS 中数据点之间距离的平方可以表示为:

$$\begin{aligned} d_{H^2}^2(x, y) &= 2 - 2k(y_i, \hat{y}_i) \\ &= 2 - 2\exp\{-(y_i - \hat{y}_i)^2 / 2\sigma^2\} \end{aligned} \quad (5)$$

因此,Kernel MSE 损失函数如下表示:

$$L_{Kernel-mse} = \sum_{t=1}^N [1 - \exp(-(y_t - \hat{y}_t)^2 / 2\sigma^2)] \quad (6)$$

在本文中, σ 的值为 $\sqrt{2}/2$ 。

2.2 经典神经网络模型

2.2.1 LSTM

LSTM 模型最初是由 Schmidhuber 于 1997 年提出的^[11]。LSTM 神经网络本质上是一个可以识别长短期依赖关系的 RNN(循环神经网络)。相比 RNN,LSTM 引入了 3 个门函数即输入门、遗忘门和输出门来控制输入值、记忆值和输出值。LSTM 的门控单元和记忆机制能够选择性地增加或删除神经元中的信息,进而大大缓解了 RNN 训练中梯度爆炸的问题^[12]。

2.2.2 GRU

Cho 等在 2014 年首次提出了 GRU 模型^[13]。GRU 是

LSTM 网络的一种效果很好的变体。GRU 模型中只有两个门:更新门和重置门。“更新门”确定来自先前存储器状态的多少信息包括在当前存储器中,“重置门”允许模型忽略在隐藏状态中不相关的信息状态。GRU 相比于 LSTM 少了输出门,其参数比 LSTM 少,这使得 GRU 模型在较小数据集上的性能往往比 LSTM 更好。

2.3 Informer 模型

Informer 是基于注意力机制的监督学习模型,整体上由编码器(Encoder)和解码器(Decoder)2 个部分构成(见图 1)。其中,编码器以全局的视野获得原始输入,保证时间序列的鲁棒性;解码器可进一步地实现序列预测。

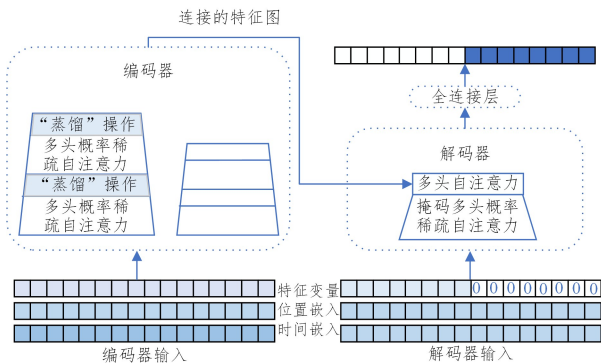


图 1 Informer 模型的基本架构

Fig.1 Basic architecture of Informer model

2.3.1 模型的输入与输出

模型输入的主要内容是前 t 个时刻的 F 种标准化后的影响因子数据 $X = (X_1, X_2, \dots, X_t)^T \in R^{F \times t}$, 在输入模型前,影响因子数据首先从 F 扩张到 d_{model} 的维度。对于时间序列预测问题,数据间的顺序关系至关重要。为了使序列数据输入模型后,其位置关系(顺序结构)不丢失,Informer 为每个输入数据的位置信息进行编码(位置嵌入),具体操作如下所示:

$$PE_{(pos, 2j)} = \sin \frac{pos}{(2L_x)^{2j/d_{\text{model}}}} \quad (7)$$

$$PE_{(pos, 2j+1)} = \cos \frac{pos}{(2L_x)^{2j/d_{\text{model}}}} \quad (8)$$

其中, pos 为位置(序列次序); $j = 1, 2, \dots, d_{\text{model}}/2$ 表示维度。另外,考虑到时间信息对时间序列预测问题的重要性,如工作日、节假日、每天不同时段等对人类活动会产生影响,Informer 将序列数据的月份、年份、时刻等信息进行编码(时间嵌入),这些全局时间信息的输入有助于提高模型捕获长期依赖的能力。最终,位置和时间嵌入信息加在变换维度后的影响因子数据上,一同进入模型。

模型的编码器和解码器均可接收数据输入,但具体内容有所差别。编码器的输入全部来自于历史数据的长序列(如前 24h),而解码器的输入由 1 段短序列(如前 12h)和 1 段与预测步长相等的零值组合而成。其中,解码器输入中的 0 值用作预测值的占位符。

2.3.2 Informer 的多头概率稀疏自注意力机制

编码器内部由多头概率稀疏自注意力(Multi-head ProbSparse Self-attention)模块和“蒸馏”(Distilling)机制模块堆叠而成。

传统的自注意力机制如式(9)所示:

$$A(Q, K, V) = \text{Softmax} \left(\frac{QK^T}{\sqrt{d}} \right) V \quad (9)$$

其中, Q, K, V 分别为由输入特征变量线性变换得到的 3 个同尺寸的矩阵。在自注意力机制运算的过程中,作者发现自注意力的结构分布为典型长尾式的分布,即少数的点积对主要的 attention 的贡献比重大,所以即使其他的点积的贡献忽略不计,模型最终的结果也不会产生明显的偏差。这意味稀疏自注意力可以在不明显损失预测精度的情况下,将模型计算的时间和空间复杂度从 $O(L^2)$ 下降至 $O(L \ln L)$ ^[1]。

为了度量 Query 的稀疏性,作者用到了 KL 散度。其中第 i 个 Query 的稀疏性的评价公式是:

$$M(Q_i, K) = \ln \sum_{j=1}^{K_K} e^{\frac{q_i k_j^T}{\sqrt{d}}} - \frac{1}{L_K} \sum_{j=1}^{L_K} q_i k_j^T \quad (10)$$

因此,可以通过以上公式得到由 Q 概率稀疏化后得到的 \bar{Q} ,进而得到概率稀疏自注意力的公式,即:

$$A(Q, K, V) = \text{Softmax} \left(\frac{\bar{Q}K^T}{\sqrt{d}} \right) V \quad (11)$$

多头是指使用不同的权重系数并行地进行概率稀疏自注意力机制,最后将所有结果集成。这样做有利于提升最终结果的精度和鲁棒性。

2.3.3 Informer 模型的编码器

编码器是为了提取长序列输入的远期的依赖性而设置的结构。在编码器内部,带着位置特征和时间特征的输入数据在经过多次多头概率稀疏自注意力模块和“蒸馏”机制模块运算后输出 1 个中间结果。“蒸馏”操作使得每层的解码器都将输入序列的长度缩短一半,从而极大地节约了编码器的内存开销和计算的时间。编码器输出的结果将会作为输入参与解码器多头自注意力的运算。

2.3.4 Informer 模型的解码器

在解码器内部,带着位置特征和时间特征的输入数据先经过带掩码的多头概率稀疏自注意力操作。与先前多头概率稀疏自注意力不同,带掩码的多头概率稀疏自注意力在预测 t 时刻的值时仅依照 t 时刻之前的特征进行预测,不像多头概率稀疏自注意力在得到 t 时刻的值时会对全局进行观察。输入数据经过带掩码的多头概率稀疏自注意力操作后,再与编码器输出中间结果进行多头自注意力操作,最终经全连接层调整数据输出维度,得到预测结果 $Y = (Y_{t+1}, Y_{t+2}, \dots, Y_{t+n})^T$ 。输出的预测结果计算损失函数后进行反向的梯度传播,不断优化模型。

时序预测模型模型总共有 3 种预测目标,分别是多变量预测多变量、单变量预测单变量和多变量预测单变量。其中,多变量预测多变量的预测目标因其输入输出的多样性以及数据间关联度的复杂性,导致该预测目标的预测难度通常是 3 类预测目标中最大的。先前的研究已经表明,在长序列时间序列预测问题下,传统的统计学模型和神经网络模型在多变量预测多变量的预测目标中的表现并不好^[15-16]。基于以上分析,本文的实证分析以多变量预测多变量的预测目标为例,证明本文提出的新模型为解决长序列时序预测问题上所做出的贡献。

3 实证分析

3.1 实验环境与数据

实验环境为 python 3.8, pytorch 1.12 框架, Windows 10, 12.0 GB, Intel (R) Core (TM) i7-8750H CPU @ 2.20 GHz,

NVIDIA GeForce GTX1060 显卡。

3.1.1 气象数据

本文使用的第一类实验数据来自于中国气象局官网下的中国地面气候资料日值数据集。选取了上海地区附近的 5 个气象基站(见图 2)1995—2020 年逐日的气象数据集(WEA)。一份实验数据集包括该气象观测点的检测日期、气压(0.1 hPa)、气温(0.1℃)、日降水量(0.1 mm)等 10 个指标,拥有 9500 条数据(数据示例见图 3)。该模型以 7:2:1 的比例划分训练集、验证集和测试集。

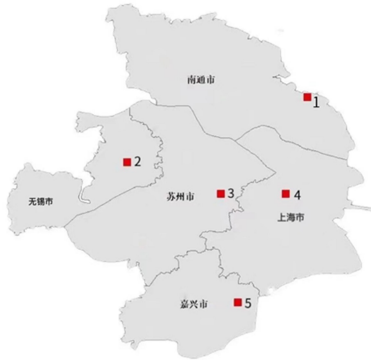


图 2 气象观测站地理位置示意图

Fig. 2 Geographical location of meteorological observation station

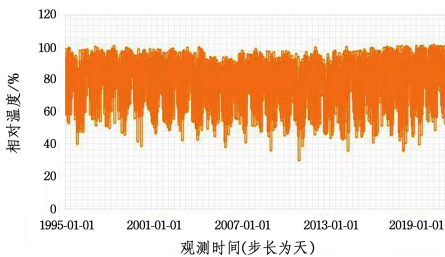


图 3 气象数据集 WEA1 中相对湿度数据的分布

Fig. 3 Distribution of relative humidity data in data set WEA₁

3.1.2 温度数据

ETT 是电力长期部署的关键指标。本文使用中国两个县的两年数据作为实验模型^[1]。文章将两个县的数据划分成了单独的数据集,1h 水平的 ETT_{h1} 和 ETT_{h2}。每个数据点由目标值“油温”和 6 个功率负载特征组成,分别含有 17 420 条数据(数据示例见图 4)。该模型以 4:1:1 的比例划分训练集、验证集和测试集。

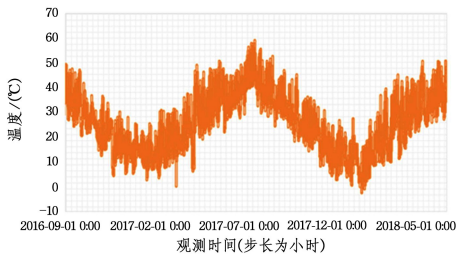


图 4 温度数据集 ETT_{h1} 中温度数据的分布

Fig. 4 Distribution of temperature data in data set ETT_{h1}

3.1.3 用电数据

该类数据来自于美国国家可再生能源实验室(NREL)提供的用电数据(ECL)^[11]。用于预测的数据集收集自某个发电站 320 个客户 2012—2015 年两年内每小时的电能消耗(kw/h),含有 26 304 条数据(数据示例见图 5)。该模型以

15:3:4 的比例划分训练集、验证集和测试集。

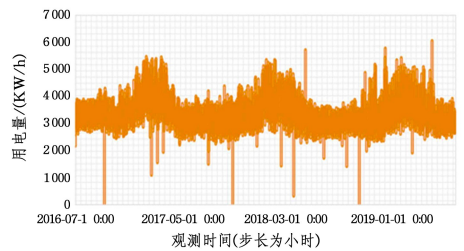


图 5 用电数据集 ECL 中用电量数据的分布

Fig. 5 Distribution of power consumption data in data set ECL

第一份气象数据为中国政府官方提供的长序列时序数据集,时间跨度较长且数据较多,有较好的验证性能。而后两份数据是被广泛认可的,被各类模型用来验证预测准确性的基准数据集^[1,11]。

3.2 数据预处理

考虑到原始数据中缺失值并不多,在原始数据预处理部分,选择使用线性插值法处理异常值与缺失值。与其他插值方式相比,线性插值法具有简单、方便的特点。

$$\frac{y_1 - y_0}{x_1 - x_0} = \frac{x - x_0}{x_1 - x_0} \quad (12)$$

其中, $(X_0, Y_0), (X_1, Y_1)$ 是缺失数据相邻的确定量, X 和 Y 分别为缺失值和缺失值所在数据中对应的真实值。 $\{X, X_0, X_1\}$ 和 $\{Y, Y_0, Y_1\}$ 都应属于同一类变量。

数据的归一化可以使得所有数据的度量单位统一,并且有利于提高模型的精度。进行归一化的公式如式(13)所示:

$$X^* = \frac{X - \mu}{\sigma} \quad (13)$$

其中, μ 为所有样本数据的均值, σ 为所有样本数据的标准差。

3.3 实验评价标准

本文选取 MAE 和 MSE 作为几种预测模型性能的评价指标。

平均绝对误差 MAE:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (14)$$

均方误差 MSE:

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (15)$$

其中, n 为测试数据集的大小; y_i 表示股价的第 i 个真实值; \hat{y}_i 表示模型的第 i 个预测值。

4 结果与分析

本文使用 3 种不同的数据类型,共计 8 套数据集进行实验。对 3 种数据集,选择采用 48, 96, 168 和 336 这 4 种不同的预测步长测试模型的预测精度。

8 套数据集实验结果如表 1 和表 2 所列。对比表 1 的结果可以看出,首先,在所选 8 套数据集中,基于注意力机制的 Informer 模型在长时序的预测精度上远远好于类 RNN 模型,且相对精度会随着预测步长的增大而增加。不仅如此,基于核技巧的 Informer 模型(KERNAL-Informer)在预测中的误差普遍小于经典的 Informer 模型以及 LSTM 和 GRU 模型。在 32 次预测比较中,基于核技巧的 Informer 模型的评价指标都低于其他 3 类深度学习模型。不仅如此,随着预测指标和预测条数的增多,基于核技巧的 Informer 模型的预测

精度也会相应提升。对于第一类数据集(10个预测指标,9499条数据),使用基于核技巧的 Informer 模型可以使得 MSE 平均提高 6.18%,MAE 平均提高 8.49%。对于第二类数据集(6个预测指标,17420条数据),使用基于核技巧的 In-

former 模型可以使得 MSE 平均提高 16.87%,MAE 平均提高 13.48%。对于第三类数据集(319个,预测指标,26304条数据),使用基于核技巧的 Informer 模型可以使得 MSE 平均提高 64.19%,MAE 平均提高 21.01%。

表1 各模型在不同数据集上的预测结果

Table 1 Prediction results of each model on different data sets

预测模型 衡量指标	Kernal-Informer		Informer		LSTM		GRU		
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	
WEA ₁	48	0.915	0.545	0.932	0.561	1.220	0.896	1.169	0.875
	96	0.930	0.549	0.945	0.590	1.267	0.903	1.207	0.885
	168	0.929	0.542	0.957	0.583	1.302	0.919	1.234	0.891
	336	0.986	0.555	1.009	0.598	1.314	0.911	1.303	0.917
WEA ₂	48	0.669	0.554	0.725	0.579	1.146	0.875	1.115	0.866
	96	0.715	0.562	0.737	0.585	1.213	0.884	1.179	0.878
	168	0.712	0.558	0.734	0.585	1.245	0.897	1.214	0.893
	336	0.722	0.569	0.760	0.595	1.265	0.908	1.296	0.917
WEA ₃	48	0.739	0.555	0.758	0.581	1.246	0.938	1.252	0.930
	96	0.741	0.559	0.761	0.578	1.342	0.972	1.297	0.961
	168	0.766	0.559	0.782	0.581	1.353	0.970	1.387	0.991
	336	0.801	0.574	0.834	0.605	1.363	0.971	1.415	0.987
WEA ₄	48	0.593	0.519	0.626	0.556	1.033	0.817	1.060	0.825
	96	0.601	0.522	0.624	0.555	1.045	0.820	1.009	0.809
	168	0.614	0.526	0.641	0.562	1.119	0.850	1.102	0.842
	336	0.609	0.526	0.629	0.557	1.343	0.934	1.290	0.914
WEA ₅	48	0.580	0.527	0.602	0.549	1.613	1.050	1.556	1.024
	96	0.579	0.531	0.599	0.553	1.632	1.052	1.661	1.060
	168	0.581	0.531	0.602	0.555	1.638	1.049	1.674	1.068
	336	0.567	0.526	0.588	0.550	1.911	1.134	1.801	1.101
ETTh ₁	48	0.623	0.584	0.678	0.619	0.702	0.675	0.672	0.665
	96	0.981	0.753	1.022	0.793	1.212	0.867	1.234	0.912
	168	0.950	0.747	1.077	0.849	1.424	0.994	1.564	1.034
	336	1.123	0.833	1.154	0.872	1.960	1.322	2.031	1.386
ETTh ₂	48	1.194	0.824	1.457	1.001	1.671	1.221	1.684	1.286
	96	2.964	1.371	3.489	1.515	4.117	1.674	3.966	1.547
	168	2.591	1.286	2.723	1.340	3.434	1.549	3.543	1.621
	336	3.024	1.340	3.467	1.473	3.963	1.788	2.019	1.388
ECT	48	0.286	0.370	0.297	0.386	0.375	0.437	0.397	0.534
	96	0.288	0.376	0.315	0.403	0.412	0.455	0.491	0.541
	168	0.306	0.388	0.343	0.431	0.442	0.473	0.492	0.557
	336	0.305	0.388	0.326	0.408	0.439	0.473	0.507	0.566

表2 新模型下不同数据集的精度平均提升值

Table 2 Average precision improvement of different data sets with new model

	(单位: %)		
	气象数据 (数据总量= 10 * 9499)	温度数据 (数据总量= 6 * 17420)	用电数据 (数据总量= 319 * 26304)
MSE 平均相对提升值	6.18	13.22	14.16
MAE 平均相对提升值	8.49	11.27	11.84

结束语 针对长时序预测任务中数据普遍存在的非线性和类 RNN 模型对全局信息捕捉能力较弱而不适用于长时序预测任务等问题,本文提出在基于注意力机制的 Informer 模型中使用一种基于核技巧改进的 Kernal-MSE 损失函数代替传统的 MSE 损失函数。Kernal-MSE 损失函数通过非线性空间中量化非线性数据的预测误差,提高了预测精度。

本文使用来自不同行业,数据量大小不同的 3 数据共 8 套数据集类型来验证新模型的预测有效性。实验结果表明,相比较经典的 Informer 模型、LSTM 模型和 GRU 模型,基于核技巧改进的 Informer 模型在长序列时间序列预测上的精度上更高。该模型适用于解决长序列时间序列预测问题。

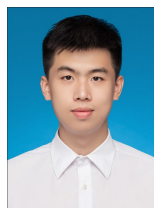
参考文献

[1] ZHOU H,ZHANG S,PENG J, et al. Informer: Beyond efficient

transformer for long sequence time-series forecasting[C]// Proceedings of the AAAI Conference on Artificial Intelligence. AAAI Press, 2021: 11106-11115.

- [2] ZHOU Z T, LIU L, SONG X, et al. Remaining useful life prediction method of rolling bearing based on Transformer model[J/OL]. [2021-08-14]. <https://doi.org/10.13700/j.bh.1001-5965.2021.0247>.
- [3] LI Y, LIN Y, XIAO T, et al. An efficient transformer decoder with compressed sub-layers[C]// Proceedings of the AAAI Conference on Artificial Intelligence. AAAI Press, 2021: 13315-13323.
- [4] WU H, MENG K, FAN D, et al. Multistep short-term wind speed forecasting using transformer[J]. Energy, 2022, 261: 125231.
- [5] ZHOU C H, LIN P Q. Traffic flow prediction method based on multi-channel Transformer[J/OL]. [2022-10-28]. <https://doi.org/10.19734/j.issn.1001-3695.2022.06.0306>.
- [6] NEYSHABUR B, BHOJANAPALLI S, MCAL-LESTER D, et al. Exploring generalization in deep learning[C]// 31st Conference on Neural Information Processing Systems. Curran Associates, Inc., 2017: 5949-5958.
- [7] CHEN L, QU H, ZHAO J. Generalized correntropy induced loss function for deep learning[C]// 2016 International Joint Conference on Neural Networks. IEEE, 2016: 1428-1433.

- [8] LAI G, CHANG W C, YANG Y, et al. Modeling long-and short-term temporal patterns with deep neural networks[C]// The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. Association for Computing Machinery, 2018:95-104.
- [9] BANDARA K, BERGMEIR C, HEWAMA-LAGE H. LSTM-MSNet: Leveraging forecasts on sets of related time series with multiple seasonal patterns[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 32(4):1586-1599.
- [10] CHEN X, YU R, ULLAH S, et al. A novel loss function of deep learning in wind speed forecasting [J]. Energy, 2022, 238: 121808.
- [11] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computer, 1997, 9(8):1735e80.
- [12] LIU H, MI X W, LI Y F. Wind speed forecasting method based on deep learning strategy using empirical wavelet transform, long short term memory neural network and Elman neural network[J]. Energy Convers Manag, 2018, 156:498e514.
- [13] CHO K, VAN MERRIENBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[C]// Conference Learning Phrase Representations Using RNN Encoder-decoder for Statistical Machine Translation. 1724-1734.
- [14] LI S, JIN X, XUAN Y, et al. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting [C]// 33rd Conference on Neural Information Processing Systems. Curran Associates, Inc., 2019:5244-5254.
- [15] LI H A, ZHOU X F, FANG L S, et al. Multivariable time series prediction method based on space-time map convolution network [J]. Computer Application Research, 2022, 39(12):1-7.
- [16] WAN C, LI W Z, DING W X, et al. A multivariable time series prediction algorithm based on self evolutionary pre training[J]. Journal of Computer Science, 2022, 45(3):513-525.
- [17] LIU H X, XIANG M, ZHOU B T, et al. Power load forecasting for long sequence time-series based on informer[J]. Journal of Hubei Minzu University(Natural Science Edition), 2021, 39(3): 326-331.
- [18] MA J W, YAN J H, SUN R W, et al. Prediction model of PM_{2.5} concentration based on LSTM-GCN[J]. Environmental Monitoring in China, 2022, 38(5):153-160.



PAN Liqun, born in 2000, postgraduate. His main research interests include deep learning and demand forecasting.