

基于 EMD-TCN 的集成深度学习预测模型研究

林子达 刘菊红 任爱珍

内蒙古农业大学 理学院, 内蒙古自治区呼和浩特, 010010;

摘要: 本文提出了一种结合经验模态分解 (EMD) 和时域卷积神经网络 (TCN) 的混合模型 EMD-TCN, 旨在建立一种高效的时序数据预测模型。该模型首先通过 EMD 对原始数据进行分解, 得到一组本征模态函数 (IMFs)。接着, 采用滑动窗口归一化方法对这些 IMFs 进行数据预处理, 然后利用 TCN 对处理后的数据进行预测。针对不同特征所得到的预测值, 本文通过计算各特征预测值的均方误差进行加权平均, 以获得最终的预测结果。本文以上证指数为实例, 进行实证分析, 证明了 EMD-TCN 在时序数据预测中的有效性。

关键词: EMD; TCN; 股指预测; 滑动窗口归一化; 集成学习

DOI: 10.69979/3041-0673.24.9.024

引言

股指序列的强非线性和非平稳性导致其建模与预测面临巨大挑战。为了解决这一问题, 许多学者尝试采用短时傅里叶变换、小波变换等方法对股指进行时频分析。但它们对参数选择非常敏感, 且在处理具有强非线性和随机性的信号时, 难以完全捕捉到关键的时态特征。因此, 具有较高自适应性的经验模态分解 (Empirical Mode Decomposition, EMD) 技术成为了股指分析的有效选择。

EMD 作为一种数据驱动的信号分解方法, 无需预设基函数, 能够通过迭代筛选过程将复杂的时序信号分解为若干本征模态函数 (Intrinsic Mode Functions, IMFs) 和一个残余项。每个 IMF 代表信号在不同时间尺度上的局部振荡特征。这一特性使得 EMD 在金融时间序列的多尺度分析中具有独特优势。谢游宇等人[1]利用 EMD 对股票收盘价进行多尺度分解, 提取不同频段的时间动态特征。Shu 等人[2]提出的 EMD-CNN-LSTM 框架, 通过 EMD 实现多尺度分解, 取得了较好的预测性能。

目前, 股指预测的研究大多集中在 LSTM 和门控循环单元 (Gated Recurrent Unit, GRU) 模型上, 但这些模型在处理长序列时仍面临梯度消失的问题。为此, Lea 等人[3]提出了时域卷积神经网络 (Temporal Convolutional Network, TCN), 其因果卷积 (Causal Convolution) 和扩张卷积 (Dilated Convolution) 结构有效缓解了梯度消失问题, 提升了计算效率和训练稳定性。基于此, Deng 等人[4]提出了 KDTCN, 该模型应对市场突发变化时表现出色。严冬梅等人[5]对 LSTM、GR

U、TCN 等模型进行了对比实验, 验证了 TCN 在金融时间序列预测中的优越性。

基于上述研究, 本文将 EMD 与 TCN 相结合, 构建一种混合预测模型 EMD-TCN。本文首先介绍了 EMD-TCN 中各方法的基本原理。其次, 详细阐述了模型架构和实现过程。最后, 进行了实证分析与对比实验。实验结果表明, EMD-TCN 在预测精度和异常波动响应能力上均优于对比模型 EMD-LSTM, 且经过集成学习优化后, EMD-TCN 的预测精度进一步提升。

1 研究方法

1.1 经验模态分解

EMD 通过将复杂信号分解为多个 IMFs 和一个残余项, 从而揭示信号的内在特征。其主要步骤如下:

步骤一: 确定极值并构建包络线。在原始信号 $s(t)$ 中, 首先识别出所有的极值点, 并利用这些极值点构建上包络线 $e_u(t)$ 和下包络线 $e_d(t)$ 。

步骤二: 计算均值曲线。计算 $e_u(t)$ 和 $e_d(t)$ 的平均值, 得到均值曲线 $m(t)$ 。

$$m(t) = \frac{e_u(t) + e_d(t)}{2} \quad (1)$$

其中, t 为时间变量。

步骤三: 初步计算 IMF 候选。将 $s(t)$ 与 $m(t)$ 进行差分, 得到第一个 IMF 候选 $h_1(t)$ 。

$$h_1(t) = s(t) - m(t) \quad (2)$$

步骤四: 检验 IMF 条件。如果候选 IMF 满足零交点与极值点的数量差不超过 1 且 $m(t)$ 接近 0, 则提取该 I

MF; 否则, 将候选 IMF 作为新信号, 重复步骤一至四, 直到满足 IMF 条件为止。

步骤五: 提取并更新残余信号。提取确认的 IMF, 并从 $s(t)$ 中减去提取的 IMF, 得到新的残余信号 $r_1(t)$ 。

$$r_1(t) = s(t) - h_1(t) \quad (3)$$

步骤六: 重复步骤一至步骤五。对新的残余信号 $r_1(t)$ 进行相同的 EMD 过程, 提取第二个 IMF 候选 $h_2(t)$ 。不断重复上述过程, 直到残余信号 $r_n(t)$ 无法进一步分解为 IMF 为止。最终信号被分解为多个 IMF 和一个最终的残余项 $r_n(t)$ 。

$$s(t) = h_1(t) + h_2(t) + \cdots + h_n(t) + r_n(t) \quad (4)$$

其中, $h_i(t)$ ($i=1,2,\dots,n$) 为第 i 个 IMF, n 为迭代次数。

1.2 滑动窗口归一化

滑动窗口归一化的基本原理是通过设定固定长度的滑动窗口, 从原始时间序列中提取多个重叠的子序列, 并对每个子序列进行归一化处理。其主要步骤如下:

步骤一: 确定窗口大小和滑动步长。根据数据的特性和分析需求, 确定适当的窗口大小和滑动步长。

步骤二: 对窗口内的数据进行归一化。本文选用极差归一化 (Min-Max Normalization), 公式如下:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (5)$$

其中, x 和 x' 为窗口内归一化前和归一化后的数据, x_{\max} 和 x_{\min} 为窗口内的最小值和最大值。

步骤三: 滑动窗口并重复归一化。按照设定的滑动步长, 窗口向前滑动至新的位置。对于新的窗口位置, 重复步骤二的归一化过程, 直到所有数据完成归一化处理。

1.3 加权集成学习

加权集成学习是一种常用的集成方法, 它通过对多个基学习器的预测结果赋予不同的权重, 然后将预测值加权平均, 以提升整体预测性能。

本文借鉴这一思想, 针对股指预测任务, 在同一模型中使用不同的价格特征进行预测, 以得到多种不同的预测结果。通过将均方误差 (Mean Squared Error, MSE) 作为权重计算指标, 可以显著降低表现较差的价格特征的权重, 从而减少它们对整体模型预测结果的负面影响。权重计算公式如下:

$$w_i = \frac{(\text{MSE}_i)^{-1}}{\sum_{j=1}^N (\text{MSE}_j)^{-1}} \quad (7)$$

其中, w_i 为第 i 个价格特征的预测结果权重, N 为价格特征数。

2 模型设计及实证分析

本节首先介绍了基于所选数据的 EMD-TCN 模型结构设计, 而后进行了实证分析与对比分析。实验在 Python 3.9 环境下进行, 使用了 TensorFlow 2.9.0、Keras-TCN 3.5.0 和 Scikit-learn 0.24.2 等库构建模型。

2.1 数据选取及 EMD-TCN 模型构成

本文选取上证指数 (000001.SH) 2021 年 1 月 1 日至 2024 年 9 月 30 日间的价格数据, 共计 908 个交易日。数据中收盘价、最高价、最低价、开盘价作为价格特征, 次日收盘价作为预测目标。选取了 2021 年 1 月 1 日至 2023 年 12 月 31 日共 727 个交易日的数据作为训练集; 其余数据作为测试集。所有数据均来源于 Choice 金融终端。

步骤一: 将训练集和测试集的价格特征及目标进行 EMD 分解。训练集分别得到了 7、7、6、6 及 7 个 IMF; 测试集分别得到了 5、5、5、6 及 5 个 IMF。针对训练集与测试集在 IMF 数量上的差异, 本文对低能量、高频 IMF 进行叠加处理。最终所有数据的 IMF 数量统一为 5。

步骤二: 对各价格特征的 IMFs 应用滑动窗口归一化处理, 窗口大小 w 为 7, 滑动步长为 1, 即每次窗口滑动一个交易日的股指数据。

步骤三: 将处理后的训练集数据输入 TCN 进行预测。以次日收盘价的 IMFs 作为目标值进行 TCN 的监督训练。然后, 利用训练好的 TCN 进行预测, 得到各价格特征 IMFs 的预测值 i ($i=1,2,\dots,n$)。

步骤四: 通过滑动窗口反归一化处理获得各价格特征 IMFs 的原始预测值, 并进行数据摘取。最终的预测序列由初始窗口的完整预测值和后续窗口的最后一个预测值组成, 确保每个交易日仅有一个预测值, 并保持预测结果的时间顺序。

步骤五: 对各价格特征的 IMFs 分别求和, 得到各价格特征对于次日收盘价的预测结果。最后根据 2.4 节的加权法, 基于各价格特征值的预测结果的 MSE^{-1} 分配权重, 对其进行加权平均, 得到最终的预测值。

2.2 实证分析和对比试验

各价格特征 IMF 的预测值的评价指标如表 2 所示。本文选取了 MSE、均方根误差 (RMSE) 和平均绝对误差 (MAE) 作为评价指标。

如表 2 所示, 对于收盘价 IMF1 的 MSE, EMD-TCN 比

EMD-LSTM 减少了 62.44%, RMSE 减少了 38.72%, MAE 减少了 49.39%。这一结果表明, EMD-TCN 在捕捉短期内收盘价的细微变化和波动特征方面表现更为出色。

然而, 在最高价 IMF2 及最低价 IMF2 的这种中期预测值评价指标中, EMD-TCN 的表现稍显逊色, 这可能是由于卷积核设置较小导致。

表 2 两模型各价格特征的 IMF 预测值的评价指标对比表

特征	EMD-TCN			EMD-LSTM		
	MSE	RMSE	MAE	MSE	RMSE	MAE
最高价 IMF1	0.059716	0.244369	0.172819	0.159009	0.398759	0.341478
最高价 IMF2	0.054961	0.234438	0.173552	0.060884	0.246746	0.194369
最高价 IMF3	0.067992	0.260753	0.186352	0.080266	0.283313	0.195491
最高价 IMF4	0.137398	0.370672	0.260858	0.139361	0.37331	0.264735
最高价 IMF5	0.143071	0.378247	0.220609	0.169261	0.411413	0.299051
最高价 IMF1	0.132731	0.364323	0.296225	0.13838	0.371995	0.320876
最高价 IMF2	0.154452	0.393004	0.327623	0.093622	0.305977	0.266879
最高价 IMF3	0.087616	0.296	0.242407	0.070189	0.264932	0.218934
最高价 IMF4	0.070625	0.265754	0.222882	0.063017	0.251032	0.217948
最高价 IMF5	0.361757	0.601462	0.518516	0.32008	0.565756	0.494974
最低价 IMF1	0.134791	0.367139	0.292637	0.141048	0.375563	0.325198
最低价 IMF2	0.199404	0.446546	0.35182	0.10933	0.330651	0.27924
最低价 IMF3	0.081724	0.285874	0.223133	0.076425	0.276451	0.222963
最低价 IMF4	0.059738	0.244413	0.137922	0.031955	0.178758	0.119566
最低价 IMF5	0.002267	0.047611	0.040956	0.016377	0.127972	0.099998
开盘价 IMF1	0.121847	0.349066	0.275779	0.13048	0.36122	0.312291
开盘价 IMF2	0.149215	0.386284	0.317357	0.130065	0.360646	0.309442
开盘价 IMF3	0.107	0.327109	0.227839	0.103252	0.321328	0.231008
开盘价 IMF4	0.034329	0.18528	0.11069	0.034884	0.186773	0.109839
开盘价 IMF5	0.425944	0.652644	0.560292	0.441802	0.664682	0.572986

EMD-TCN 和 EMD-LSTM 的最终预测值及真实值折线图如图 1 所示, 展示了这两种模型预测值与真实值的对比情况。可以从中清楚地看到 EMD-TCN 模型在极值预测的显著优势, 它能够更准确地捕捉价格的急剧波动, 在捕捉市场价格的快速变化和细微波动方面表现出色。

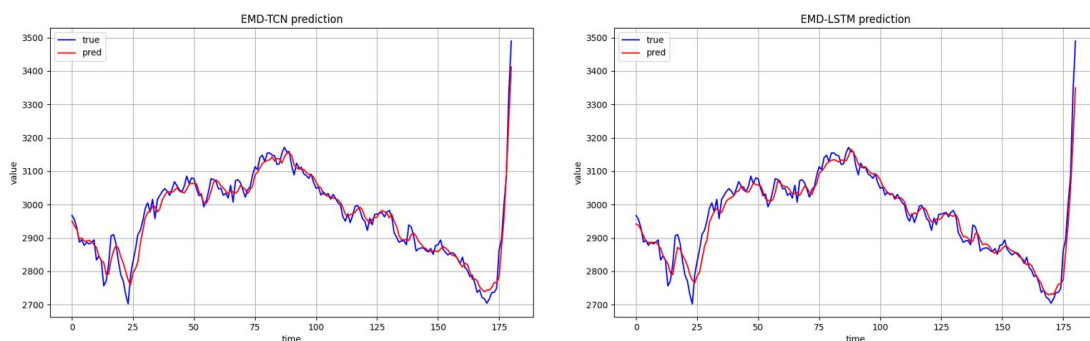


图 1 两模型最终预测值与真实值对比折线图

次日收盘价预测值的评价指标如表 3 所示。与 IMF 的预测值评价指标不同, 两种模型展现出了更显著的性能差异。具体而言, 采用开盘价预测次日收盘价时, EMD-TCN 在各项评价指标上均显著优于 EMD-LSTM。从数据特征的角度分析两模型差异的原因, 测试集中开盘价

被分解出 6 个 IMF, 而其他价格特征均分解出 5 个 IMF。这表明开盘价数据更为复杂, 其包含更高的波动频率和噪声。这就对模型的处理能力提出了更高要求, 而 TCN 模型在处理此类数据时展现出明显优势。

此外, 表 3 中还展示了集成学习的预测结果。集成

学习后,EMD-TCN 的各项评价指标均优于 EMD-LSTM 模型,且其各项评价指标也均优于单价格特征预测值的评价指标。进一步验证了 EMD-TCN 在综合预测性能上的优越性。

表 3 两模型各价格特征及集成学习预测值的评价指标对比表

模型	特征	MSE	MAE	RMSE
EMD-TCN	收盘价预测	711.7926	20.38387	26.67944
	最高价预测	1635.556	27.31478	40.44201
	最低价预测	1071.086	25.13097	32.72745
	开盘价预测	947.0387	22.01765	30.77399
	集成学习预测	681.6535	19.63832	26.1085
EMD-LSTM	收盘价预测	678.6263	19.40347	26.05046
	最高价预测	1514.594	25.19593	38.91778
	最低价预测	1095.989	21.66371	33.10572
	开盘价预测	1261.722	24.29934	35.52073
	集成学习预测	834.52	20.0409	28.88806

3 总结

综上所述,EMD-TCN 在金融时间序列预测中展现出了显著的优势。更强的理解和分析能力,使其在处理复杂、高频和噪声较多的数据时,能够更为精准地捕捉价格波动中的波峰和波谷特征并预测结果,显示了其在金融市场快速变化情境中的适应性和稳健性。这为金融市场中股指预测模型的选择提供了重要的参考依据。

参考文献

[1]谢游宇,王万雄.基于 EMD 和 SSA 的股票预测模型

[J]. 计算机工程与应用,2023,59(18):285-292.

[2]Shu, Wangwei, and Qiang Gao. "Forecasting stock price based on frequency components by EMD and neural networks." Ieee Access 8 (2020): 206388-206395.

[3]Lea C, Flynn M D, Vidal R, et al. Temporal convolutional networks for action segmentation and detection[C]//proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 156-165.

[4]Deng, Shumin, et al. "Knowledge-driven stock trend prediction and explanation via temporal convolutional network." Companion proceedings of the 2019 world wide web conference. 2019.

[5]严冬梅,何雯馨,陈智.融合情感特征的基于 RoBERTa-TCN 的股价预测研究[J]. 数据分析与知识发现,2022, 6(12):123-133.

课题来源:国家自然科学基金,课题名称:短花针茅荒漠草原种间竞争及其空间发生尺度对载畜率的响应,(课题编号:32160258).

作者简介:林子达(1998-),男,吉林长春人,研究生在读,研究方向:经济统计分析。

通讯作者简介:任爱珍(1974-),女,内蒙古呼和浩特人,蒙古族,副教授,研究方向:经济统计分析、机器学习。