【大数据与区块链专栏】

基于CEEMDAN-LSTM的股票市场指数预测建模研究

贺毅岳[1]，李 萍[2]，韩进博[1]

（1.西北大学 经济管理学院，陕西 西安 710127；2.西安财经大学，陕西 西安 710100）

摘要：为满足主动型量化投资对股票市场指数高精度预测的要求，将自适应噪声完备集合经验模态分解（CEEMDAN）引入到股市指数预测建模中，结合长短期记忆网络（LSTM）对复杂序列中长期依赖关系高效的建模能力，采用“分解—重组—预测—集成”思路，提出一种股市指数集成预测方法CEEMDAN-LSTM。运用CEEMDAN对指数进行分解与重构，获得其高、低频分量及趋势项；分别构建各分量的LSTM预测模型并优化高频子序列IMF重组方式，进而通过加和集成各分量预测值获得指数的整体预测值。以沪深300等5个代表性的股市指数为测试数据，对CEEMDAN-LSTM与主流的金融时序机器学习建模方法的预测效果进行系统的对比实验，结果表明：CEEMDAN-LSTM的预测表现一致性地优于现有建模方法，具有更低的预测误差和滞后性。

关键词：量化投资；股市指数；预测建模；自适应噪声完备集合经验模态分解；长短期记忆网络

中图分类号：F830 文献标志码：A 文章编号：1007－3116(2020)06－0034－12

一、引 言

股票市场是上市公司筹集社会资金的重要途径之一，股票投资已成为投资者实现资产保值、增值目标的主要方式之一。在股市投资研究中，资产价格行为的分析与建模是备受研究者关注的重要课题。对主动型股票投资研究而言，价格行为研究的核心是对股票价格的走势或未来值进行有效预测，进而指导投资者的交易决策行为，以使其所持有的投资组合经风险调整后收益最优化。然而，面对信息错综复杂、状态瞬息万变的股票市场，如何透过复杂现象看本质、把握股票市场行情脉络和资产价格运动状态及趋势，进而通过持续的交易决策获得理想的投资收益甚至超额收益，是股票投资者尤其是机构投资者高度关注并深入研究的核心课题。然而，传统股票投资分析方法，包括基本面分析法和技术分析法，却因建模数据体量小、模态单一、蕴含信息量少或模型发现数据变化规律低效等缺陷，难以对股市行情和资产价格的走势或未来值进行有效预测，进而难以为股票持续交易过程中的动态最优投资决策提供足够的信息支撑[1－2]。

主动型股票量化投资利用计算机技术深入分析大量的市场交易与参与者相关数据，以捕获市场的短期非有效现象，然后运用线性或非线性优化方法构建投资策略模型，包括选股、择时、风险管控等子模型，进而应用于选股、择时和风险管理等实务操作所依赖的系列决策过程，以获取投资者期望的最优风险调整收益[1]。择时是金融投资过程中的关键环节，即投资

收稿日期：2019－12－12；修复日期：2020－02－18

基金项目：教育部人文社会科学研究青年基金项目“基于多尺度时序模式挖掘的股票在线算法交易策略研究”（16XJC630001）；中国博士后科学基金面上资助项目“大数据视角下基于深度学习预测建模的算法交易策略研究”（2017M623229）；陕西省自然科学基础研究计划项目“基于多粒度时序模式挖掘与预测的股票算法交易策略研究”（2015JQ7278）

作者简介：贺毅岳，男，湖南娄底人，博士，副教授，硕士生导师，研究方向：计算金融与风险管理；

李 萍（通讯作者），女，黑龙江哈尔滨人，博士，教授，硕士生导师，研究方向：统计理论及应用；

韩进博，男，陕西延安人，硕士生，研究方向：金融数据挖掘。

者根据其对资产价格运动趋势或未来状态的预测,在恰当的时机进入或退出市场,从而实现规避亏损、获取收益的目标。对于股票投资中的非系统风险,通常需要通过择时加以规避。股票量化择时就是运用数量化方法判断股票的走势或未来状态值,进行高抛低吸操作以获取超额收益的交易行为。择时的关键在于如何有效预测股价走势或未来值,而股票市场是一个受多种经济社会因素驱动的非线性复杂系统,其价格波动具有显著的非平稳、非线性和高噪声的复杂特性。传统的股市预测方法,包括金融计量方法、统计建模方法、浅层的机器学习方法,存在建模数据的容量较小或模型发现数据复杂模式的能力不足等重要缺陷,使得股价序列的特征提取及预测建模成为了金融数据建模领域的一个关键难题[3]。

近年来,机器学习在计算机视觉和语音识别等领域取得了系列突破性进展,特别是谷歌 Alpha Go 的出现,激起了众多行业、领域展开人工智能研究与应用的热潮,对数据密集型的金融投资行业产生了尤其深远的影响。国内外机构投资者正深入研究如何将最新的机器学习与人工智能技术引入到量化投资策略建模过程中,并已逐步形成了新的结合智能方法的主动型量化投资模式[4]。目前,国内已出现了一些该类投资模式的成功案例,如广发证券金融工程团队证实了深度学习在多因子选股、量化择时和 CTA 策略构建等方面的优异表现[5]。深度学习是一种新型的多隐藏层神经网络,通过模拟人类大脑在学习过程中的多层抽象机制,建立从底层信号到高层语义的非线性可逆映射关系,在对复杂输入样本本质特征的抽取方面表现出了强大的能力。在计算机视觉、自然语言处理和金融数据建模等众多应用领域中,基于深度学习构建的模型性能及泛化能力优异,多数应用效果取得了历史性突破[6]。因此,将深度学习中的最新方法拓展应用于股市指数序列的预测建模,可为复杂金融时序数据的建模研究提供有益的参考,同时有利于提升量化择时研究方法的科学性与实用性,这也正是当前股票量化投资研究的一个热点[1,3]。

二、文献综述

股票价格预测建模即建立股价走势或未来值的预测模型,是量化择时策略建模过程中的核心环节,也是量化投资理论和实务界形成共识的重要研究课题[3]。国内外学者针对股票价格及市场指数的预测建模展开了系统的研究,提出了三类预测建模方法。(1)技术分析法以道氏理论为基础,认为股价基本走势与市场波动趋势趋同,包括短、中、长三种走势,三者同时存在相辅相成。典型的股票技术分析研究大多以择时方法或策略的构造为应用背景。Mabu 等运用一种基于图的进化计算方法——遗传网络规划方法,提取大量的技术指标规则创建规则池,并构建了适合日本股票市场的基于多技术指标规则组合的量化择时模型,实证研究结果表明:其所构建的多指标组合择时策略的收益比传统的单指标择时策略更高[7]。Wang 等将技术指标规则组合应用于 NASDAQ100 指数成分股,构建了一个复杂的绩效奖励交易策略,其中使用时变粒子群算法获得策略的最优参数集,实证结果表明技术规则组合择时表现胜过基于单个指标规则的择时[8]。梁淇俊等以技术指标为择时策略依据,根据指标 MACD、RSI 和 OBV 构建交易信号以及信号有效性的择优体系,并以中信证券收盘价数据为例,对基于三个技术指标的单策略、联合策略有效性进行了量化分析,得到了 MACD 指标择时相对最优的结论[9]。(2)统计建模方法依据严谨的统计学理论对股价序列进行预测建模。国内外学者对 ARIMA、GARCH 和 HMM 等代表性方法进行了系列研究。Hassan 提出一种新的 HMM 与模糊模型相结合的股价预测方法,使用 HMM 识别股价变化模式并用模糊逻辑进行预测,得到了比 ARIMA、ANN 等模型精度更高的预测效果[10]。张超提出基于误差校正的 ARMA-GARCH 股价预测方法,并将其应用于上证指数,显著提升了预测精度[11]。张蓓利用高斯混合 GHMM 模型对 IBM 的股价进行预测,并验证了其预测效果优于 HMM 模型[12]。(3)利用机器学习方法对金融时序进行预测建模是近年来金融数据分析领域的研究热点。Tay 等从结构风险最小化角度深入分析了 SVM 的最小化泛化误差优势,首次利用 SVM 对标普 500 指数进行预测,验证了 SVM 的金融预测性能优于传统神经网络[13]。Chen 等提出了一个基于信息增益的特征加权 SVM 和 KNN 结合的预测模型,并对沪深股市指数进行预测实验,获得了比现有模型更好的预测效果[14]。Bao 等实证证明了长短期记忆网络(LSTM)对金融时序的预测性能优于传统的 RNN[15]。Thomas 等利用 LSTM 对标普 500 指数的变化方向进行预测,发现 LSTM 比随机森林、深度神经网络与 logistic 回归的分类效果好[16]。杨青等构造深层 LSTM 神经网络并对全球 30 种股票指数的 3 种不同期限进行预测,结果表明 LSTM 泛化能力强,对全部指数在不同期限下的预测效果稳定,比 ARIMA、MLP 和 SVR 预测精度更高,并能有效控制

误差波动，提高不同期限下指数预测的稳定度[5]。

上述三类方法在金融时序预测问题上大多取得了较好的实证效果，但依然存在一定的理论或实用性缺陷：技术分析法直观，但其时效性较弱、所产生的买卖信号不确定性过高，易导致预测偏误；统计建模方法的预测结果在统计意义上可靠，但通常假定所预测序列线性或近似线性，难以实现对非线性、低信噪比金融时序的高精度预测；机器学习方法避免了统计建模方法中数据分布假设过于严格的问题，并具有更强的非线性关系抽象能力，能显著提升股价预测的准确性[3]。然而，股票市场是一个以多种方式对外部环境变化进行响应的复杂系统，随机性很强且各种现象之间存在复杂的非线性内在关系，而现有的金融时序预测建模通常依靠单一方法直接对序列模式进行挖掘，无法充分提取复杂的序列变化模式，故即便通过 SVM、RNN 和 LSTM 等机器学习方法，依然难以获得股票投资决策所需的高精度股价预测信息。

随着对金融市场微观结构与交易行为心理等方面研究的不断深入，学者们逐渐认识到单个技术难以高效地挖掘并刻画复杂金融市场中的多维量价变化规律，进而实现高精度预测，而融合金融计量、信号处理和机器学习等多学科方法的混合或集成模型，则能通过其不同子模块识别数据的不同模式，进而汇总获得其中蕴含的完整变化规律，实现金融时序的高精度预测[5]。美国工程院院士 Huang 等创造性地提出了经验模态分解(EMD)方法，将时序信号中不同尺度的趋势或波动逐级分解，生成一系列具有不同特征尺度的本征模函数(IMF)，理论上可实现对非平稳、非线性时序信号的分解[17]。针对 EMD 分解不彻底、产生虚假分量和模态混叠的问题，Wu 等通过引入频率分布均匀的辅助噪声改进 EMD 方法，提出了集成经验模态分解(EEMD)方法，解决了模态混叠问题，但处理过程中加入的高斯白噪声很难完全去除[18]。Torres 等通过加入自适应白噪声进一步改进 EEMD，提出了 CEEMDAN 方法，有效克服了 EEMD 分解不完备和重构误差过大的问题。CEEMDAN 分解获得的各 IMF 相对简单且相互独立，为充分提取 IMF 子序列的波动特征提供了有利条件，从而显著降低了金融时序预测建模的难度[19]。EMD 早期主要应用于信号去噪与气象科学领域，近年被引入到经济与金融等领域，其中与机器学习方法相结合的典型研究有：Yang 等将汇率序列经 EMD 分解获得的 IMF 输入极限学习机，实现了对汇率预测精度的提升[20]；贺毅岳等提出了 EMD 分解下基于 SVR 的股价集成预测方法 EMD-SVRF，实证结果表明该方法比 EMD-Elman 和 ARMA-GARCH 等已有方法具有更小的预测误差[21]；李合龙等运用 EEMD 方法对投资者情绪和股指价格序列进行分解和重构，并结合计量模型分析两者在不同时间尺度下的波动关联性[22]；Zhang 等在对地表温度的预测研究中提出构建 EEMD 与 LSTM 混合的预测模型，其实证结果表明该模型的预测效果优于 RNN、LSTM 和 EMD-RNN 等机器学习预测模型[23]。上述研究表明：CEEMDAN 克服了模态混叠问题并具有自适应分解完备和重构误差低的优点，在提取复杂时间序列的波动模式进而提升预测建模精度方面具有突出的优势，是金融时间序列分析领域极具应用前景的新方法；另一方面，LSTM 通过引入门控单元系统，解决了传统 RNN 模型训练中梯度爆炸和梯度消失问题，在提取序列数据中的长期依赖关系方面极具优势，可利用前期“记忆”为当期决策提供支持，是当前复杂高维时序数据分析中最成功的非线性建模方法之一，也是近年来金融数据建模领域的研究热点[3,5]。

为此，本文提出一种 CEEMDAN 与 LSTM 结合的股市指数预测建模方法 CEEMDAN-LSTM：首先，运用 CEEMDAN 方法对市场指数序列进行分解与重构，获得高频分量、低频分量与趋势项 3 个子序列；然后，分别构建各子序列 LSTM 预测模型，并依据模型获得各子序列的预测值，进而通过加和集成处理获得市场指数的整体预测值。最后，以沪深 300 和中证 500 等 5 个代表性的国内股市指数为测试数据集，对本文预测建模方法和现有主流的金融时序机器学习预测建模方法的市场指数预测效果进行对比实验，以分析、验证本文方法的有效性和实用性。

本文旨在提出高精度的股市指数预测建模方法，为主动型量化投资研究与实践者把握股市动态趋势、规避市场风险进而增强超额收益能力提供更有效的工具。本文的主要创新在于：(1)将具有自适应分解能力的 CEEMDAN 方法引入到股市指数的预测建模过程中，从而获得波动特征相对简单且相互独立的高频、低频分量和趋势项 3 个子序列，为进一步对各子序列的高精度预测建模创造了有利条件，避免了现有建模方法直接从指数时序数据中提取波动模式的技术难题，显著降低了指数时序预测建模的难度。(2)针对指数 CEEMDAN 分解所产生的多个子序列，运用 LSTM 构建各子序列的预测模型，克服了传统统计建模方法对适用数据的分布假设过于严格的局限性，且能更高效地提取序列中

蕴含的长期动态依赖关系,可为复杂金融时序的非线性预测建模提供有益参考。(3)将 CEEMDAN 的自适应分解功能与 LSTM 的长期依赖关系提取能力有效结合,构建股市指数的高精度混合预测模型,对提升量化择时信号的准确度与有效性具有较强的应用参考价值,有利于拓宽基于机器学习建模的量化投资策略设计的研究思路。

三、预测建模的理论基础

(一)CEEMDAN 分解和重构

1. CEEMDAN 原理。EEMDAN 是针对经验模态分解(EMD)和集成经验模态分解(EEMD)的不足而提出的一种噪声辅助数据分析方法。EMD 作为自适应信号时频处理方法可用于非线性、非平稳信号的分析处理,其特征是将信号平稳化,提取出信号中不同尺度的波动模式,生成一系列具有不同时间尺度局部特征的数据序列,每一个序列即为一个本征模态函数(IMF)。EMD 分解的基本思路是用上、下包络的平均值去确定"瞬时平衡位置",进而提取 IMF,具体包括如下四个步骤:

(1)识别 $S(t)$ 中所有极大值点 max 和极小值点 min,用三次样条插值方法分别绘制出上、下包络线。其中,$S(t)$ 表示当前待分解序列,本文中其取值为市场指数收盘价序列。

(2)计算每一时刻上、下包络线的局部瞬时均值,从而获得平均包络线 $m(t)$,按照式(1)计算新序列 $d(t)$。

$$d(t) = S(t) - m(t) \tag{1}$$

然后,按照式(2)计算出 S_d 值来判断 $d(t)$ 是否为本征模函数。

$$S_d = \frac{\sum_{t}^{T} |d_i(t) - d_{i-1}(t)|^2}{\sum_{t}^{T} d_i^2(t)} \tag{2}$$

其中,$d_i(t)$ 为第 i 次筛分的结果,S_d 的阈值通常设定为 0.2 ~ 0.3。若 S_d 值小于阈值,则筛分处理停止;否则,将 $d(t)$ 当作新的待分解序列 $S(t)$,重新执行上述迭代处理过程。

(3)若 $d(t)$ 满足 IMF 成立所需要的两个条件,则 $d(t)$ 为一个 IMF,将 $d(t)$ 从 $S(t)$ 中分离,得到余项 $r(t) = S(t) - d(t)$。

(4)若余项 $r(t)$ 已成为一个单调函数或常数,或振幅低于既定阈值而无法进一步提取 IMF,则整个分解过程结束。否则,将 $r(t)$ 当作待分解序列 $S(t)$,返回步骤(1),重新执行上述迭代处理过程。

经 EMD 分解原序列 $S(t)$ 被迭代分解为 n 个彼此正交的 IMF,记为 $c_i(t), i = 1,2,\cdots,n$,以及表示原时序信号 $S(t)$ 趋势的最终残差项 $r_n(t)$。如式(3)所示,其中 $c_i(t)$ 依次取为步骤(3)所得到的本征模函数 $d(t)$。

$$S(t) = \sum_{i=1}^{n} c_i(t) + r_n(t) \tag{3}$$

为解决 EMD 分解中存在的模态混叠问题,Wu 等在 EMD 分解中引入频率分布均匀的辅助噪声,提出了 EEMD 方法:每次将不同的频率均匀分布的辅助白噪声加入目标信号,然后利用 EMD 分解含有附加白噪声的信号,重复执行上述过程 N 次,最后对分解获得的 IMFs 和趋势项分别进行集成平均,得到原信号的最终分解结果[18]。虽然 EEMD 显著改进了 EMD 的不足,但 EEMD 对原序列所添加的白噪声仍可能在多次平均后影响分解产生的子序列,进而影响子序列的预测精度。CEEMDAN 进一步改进 EEMD 算法,在每次分解中都添加自适应白噪声来平滑干扰脉冲,进一步提升了 EEMD 分解的完整性,降低了重构误差[19]。

2. IMF 重组方法。原序列 $S(t)$ 进行 CEEMDAN 分解获得的各本征模函数 $c_i(t)$,按如下三个步骤进行重组[21-22],可获得 $S(t)$ 的高频分量、低频分量和趋势项 3 个子序列:

(1)分别计算各本征模函数 $c_i(t)$ 的均值,$i = 1,2,\cdots,n$;

(2)给定显著性水平为 0.05,按 $i = 1,2,\cdots,n$ 的顺序,依次对 $c_i(t)$ 执行均值不为 0 的 t 检验;

(3)若 $c_k(t)$ 为第一个均值显著非零的 IMF,则将 $c_1(t)$ 至 $c_{k-1}(t)$ 加和得到 $S(t)$ 的高频子序列,将 $c_k(t)$ 至 $c_n(t)$ 加和得到 $S(t)$ 的低频子序列,而将 $r_n(t)$ 作为 $S(t)$ 的趋势项。

(二)LSTM 内部结构与工作原理

图 1 给出了循环神经网络(RNN)按时间展开的结构,其中主体结构 A 在 t 时刻读取输入信息,包括来自输入层的 x_t 以及模型的上一时刻状态 h_{t-1},以此更新其自身状态为 h_t 并产生输出 o_t[16]。RNN 凭借其在不同时刻隐含节点具有连接的结构,可实现对历史信息的记忆并应用于当前输出的计算,因而适用于时序信息的挖掘问题,已被广泛应用于包括语音识别等多领域中序列数据的建模过程。然而,RNN 参数优化时面临梯度消失和梯度爆炸的问题,致使其参数难以训练达到最优值,进而使得 RNN 网络无法有效处理长期时序依赖关系。

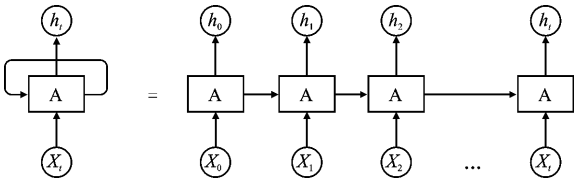


图 1　RNN 按时间展开的结构

LSTM 是通过引入由输入门、遗忘门和输出门构成的门控单元系统而产生的一种 RNN 变体[3]。“门”是一种能对信息的通过进行选择性控制的结构，通过一个 sigmoid 层和一个逐点相乘操作来实现，其输出值在 0～1 之间，0 表示完全不通过，1 表示完全通过。LSTM 用内部记忆单元即细胞的状态保存历史信息，并利用不同的“门”动态地让网络学习适时遗忘历史信息、依据新信息更新细胞状态，以解决 RNN 中梯度消失与梯度爆炸的问题。LSTM 神经网络记忆单元的基本结构如图 2 所示[5]。

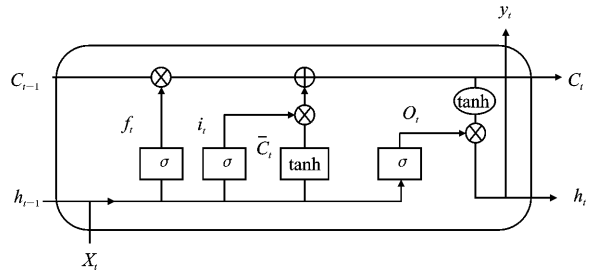


图 2　LSTM 单元的内部结构

LSTM 通过遗忘门控制从当前状态中移除哪些信息，输入门控制哪些信息传递到当前状态中，输出门控制当前状态中的哪些信息用作输出，三个“门”共同作用、处理信息，完成时间序列的预测。遗忘门决定哪些信息被细胞状态丢弃，它读取 h_{t-1} 和 x_t，按照式(4) 计算遗忘门的输出 f_t：

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \tag{4}$$

其中，激活函数 σ 表示 sigmoid 函数。确定哪些信息被存放于细胞状态中涉及两步过程：输入层即 sigmoid 决定需要更新的值，它通过线性变换和激活函数 σ 得到 i_t；tanh 层创建新的候选向量 $\widetilde{C}_t$ 并加入到细胞状态：

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \tag{5}$$

$$\widetilde{C}_t = \sigma(W_c[h_{t-1}, x_t] + b_c) \tag{6}$$

将旧状态 C_{t-1} 与遗忘门向量 f_t 相乘并丢弃部分信息，再加上 $i_t \widetilde{C}_t$，得到新的细胞状态值 C_t：

$$C_t = f_t C_{t-1} + i_t \widetilde{C}_t \tag{7}$$

最后，通过“输出门”来确定输出什么信息。通过 sigmoid 层确定细胞状态的输出部分，然后使用 tanh 层对细胞状态 C_t 进行处理并与 sigmoid 门的输出相乘，确定最后的输出 h_t：

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{8}$$

$$h_t = o_t \tanh(C_t) \tag{9}$$

在式(4) ～ (9) 中，x_t、h_t 和 C_t 分别表示细胞的输入、输出和状态向量；f_t、i_t 和 o_t 分别表示遗忘门输出、输入门输出和输出门输出向量；W 和 b 表示权重向量和偏置项。用 LSTM 单元替代标准 RNN 中的隐状态节点可以构建出 LSTM 网络。基于门控单元系统的结构特征使得 LSTM 网络可以高效地处理复杂的长期时序动态依赖关系，特别适用于复杂金融时间序列建模。

四、面向股市指数预测的 CEEMDAN-LSTM 模型构建

(一)CEEMDAN-LSTM 的建模思路

股市指数经 CEEMDAN 分解与重组产生的子序列波动特征相对简单，为进一步构建预测建模以充分提取子序列的波动模式创造了有利条件，可显著降低对指数序列高精度预测建模的难度。为此，本文将 CEEMDAN 的时序分解与 LSTM 的时序预测的两个优势功能进行结合，提出一个高精度的市场指数预测方法 CEEMDAN-LSTM。图 3 是 CEEMDAN-LSTM 的建模流程：以股市指数收盘价序列为输入数据，通过 CEEMDAN 分解、本征模函数 IMF 的重组、高/低频分量及趋势项的 LSTM 建模及各分量预测值的加和集成四个处理阶段，最终获得高精度的指数序列预测值。

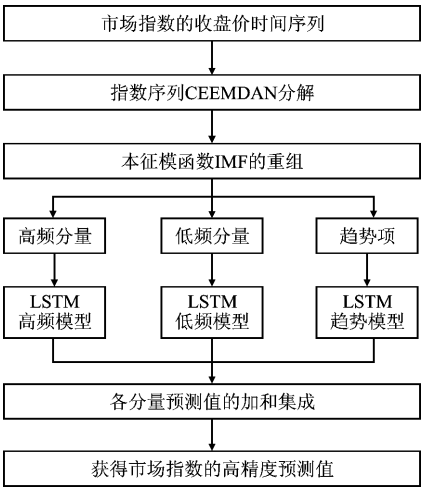


图 3　CEEMDAN-LSTM 预测建模的流程

步骤 1：运用 CEEMDAN 方法将市场指数收盘价序列分解为 n 个本征模函数 $c_i(t)$，$i=1,2,\cdots,n$，以及一个趋势项 $r_n(t)$。

步骤 2：按照前文所述的基于均值 t 检验的

IMF 重组方法，将各本征模函数 $c_i(t)$ 重组为原指数序列的高频分量、低频分量以及趋势项 3 个子序列。

步骤 3：针对高、低频分量和趋势项 3 个子序列，分别建立对应的 LSTM 预测模型，并对高频子序列重组中 IMF 组合方式进行优化，以使高频预测模型达到最优预测效果。

步骤 4：利用步骤 3 构建的 3 个子序列 LSTM 预测模型，计算获得各子序列的预测值，进而通过加和集成处理获得指数的高精度预测值。

（二）建模数据的选取及检验

本文在阐述股市指数的预测建模过程中选取沪深 300 指数作为建模的数据基础，原因包括：首先，沪深 300 指数是以沪深两市具有很强代表性的 300 只股票为基础编制而成，覆盖了 A 股市场中大多数蓝筹股，覆盖的行业较均衡合理，其市值约占 A 股市场的六成，具有很强的市场代表性，能较准确地反映沪深两市股价变化的整体行情及趋势。其次，该指数收益率是评价股票组合投资业绩的重要基准之一，可为市场中的指数化投资、指数衍生产品的创新提供基础条件，因而研究沪深 300 指数预测建模对衍生品市场的投资研究也具有重要意义。

利用 Python 从聚宽量化平台在线提取了 2006 年 1 月 1 日至 2018 年 2 月 1 日之间沪深 300 指数的收盘价，剔除节假日等因素的影响，共计2 955个数据作为指数预测建模的原始时序数据。在图 4 所示的建模时间区间内，指数先后两次大致经历了上涨、下跌和横盘震荡三种行情阶段，构成了两个完整的股指运行周期，这使得本文所建立的模型对股市指数变化规律的表达更充分、对行情变化的适应性更强，从而能增强本文研究结论的说服力。

图 4　沪深 300 指数序列

对沪深 300 指数序列数据进行 ADF 检验，结果显示在 1%显著性水平下指数非平稳；对指数的对数收益率序列进行 Jarque-Bera 检验，偏度为－0.587，峰度为 3.639，具有尖峰厚尾特征，p 值近似为 0，指数收益率分布显著非正态。同时，利用 Ljung-Box 统计量检验指数收益率序列的 ARCH 效应，结果显示滞后阶数超过 4 以后，p 值远远小于 0.05，表明收益率序列有显著的波动聚集性。沪深 300 指数序列非平稳且包含大量的噪声，而传统的 ARIMA、GARCH 等计量模型，在未进行高效的降噪处理情况下，很难对这种复杂金融时序进行高精度的预测建模。因此，本文引入 CEEMDAN 对指数进行自适应分解、去噪与重构，然后运用非线性时序建模方法 LSTM 对指数进行预测建模是合理且必要的。

（三）指数序列的 CEEMDAN 分解和重组

1. 指数序列的 CEEMDAN 分解。按照前文所述 CEEMDAN 分解过程，对沪深 300 指数序列进行自适应分解，结果如图 5 所示，得到从上往下依次排列的 10 个 IMF 和 1 个残余项，其中横轴表示指数的时间序号，纵轴表示各 IMF 的频率，从 IMF1～IMF10 到残余项频率逐步下降，变化模式也较原序列更简单。

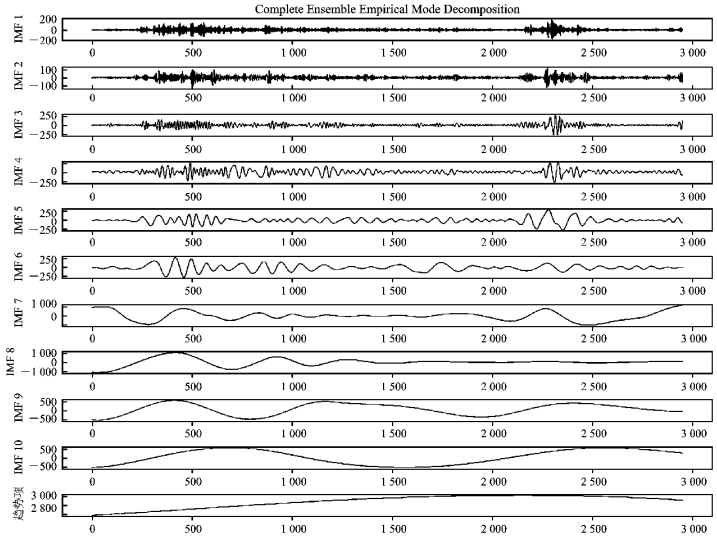


图 5　沪深 300 指数 CEEMDAN 分解结果

2. IMF 重组处理。为了适当降低 LSTM 预测建模的复杂度和避免模型过拟合，参考李合龙等的研究，按照前文所述 IMF 重组方法，对沪深 300 指数经 CEEMDAN 分解所产生的 10 个 IMF 进行重组[22]。依次对 IMF1～IMF10 进行均值为 0 的 t 检验，检验结果显示其中 IMF5 是首个 P 值小于 0.05 的本征模函数，即 IMF5 的均值显著不等于 0。因此，本文将 IMF1～IMF4 重组成为指数的高频分量，IMF5～IMF10 重组为指数的低频分量，将残余项作为指数的趋势项 $r(t)$，从而获得图 6 所示从不同频率视角下刻画原指数序列变化模式的 3 个子序列。子序列变化模式相对简单、有规律性，便于进一步充分提取各子序列的波动特征。

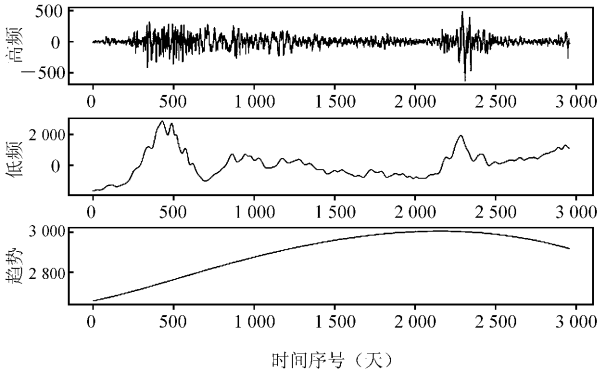


图 6　IMF 重组得到的 3 个子序列

（四）子序列 LSTM 建模及效果评价

针对 CEEMDAN 分解与重组获得的 3 个分量子序列，包括高频分量、低频分量和趋势项，分别构建各子序列的 LSTM 预测模型，进而利用模型对预测区间内各子序列进行滚动预测，并采用确定系数 R^2、可解释方差 EVS、均方根误差 RMSE 和平均绝对误差 MAE 四个评价指标，对各模型的预测效果进行评估。

1. 子序列 LSTM 预测建模。采用滚动预测建模方式，以最近 30 天的指数值为输入来预测下一天的指数值[5,21]。从建模的原始时序数据中选取 2006 年 1 月 1 日至 2016 年 2 月 1 日共 2 450 个数据构建模型训练集，并采用 Python 库 Pandas 中的 DataFrame 对象来表示，大小为(2420×30)，以剩余的 505 个数据构建测试集，对应的 DataFrame 对象大小为(505×30)。然后依次建立高、低频分量和趋势项对应的 LSTM 预测模型。

本文所构建的深层 LSTM 网络具有图 7 所示的计算图结构，虚线方框内表示深层网络的结构。在建模过程中，为消除数据间的量纲影响并提升模型的运算速度，对数据进行 Z-score 标准化处理[4]。模型参数设置方面参照了杨青等的研究，考虑到金融时序的非线性复杂特征及模型的运算效率，将隐藏层个数设置为 2 层，且每次投入模型的样例个数即 batch_size 设置为 41，迭代次数设置为 100 次，同时增设 Dropout 层以优化神经网络，失活率设置为 0.2。为使模型快速收敛时损失函数取全局最小值，选取优化器为 Adagrad，设置动态学习率的初值设定为 0.1，并根据经验公式 $0.1\times(0.96^{epoch})$ 动态调整，其中 epoch 为迭代次数，以使学习率随模型迭代次数的增加而均匀下降[5]。

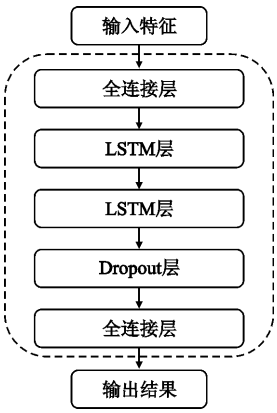


图 7　LSTM 网络的计算图结构

表 1 给出了不同神经元个数组合条件下高频子序列 LSTM 预测模型的多指标评价结果，其中(10，10)对应的实验结果整体最佳，故将高频模型中两个隐藏层的神经元个数设定为(10，10)。按照同样的方法，将低频、趋势子序列预测模型的两个隐藏层神经元个数均设定为(6，6)。在确定上述参数条件下，建立各子序列的 LSTM 预测模型，模型均在 100 次迭代后损失函数均能收敛到平稳状态，故本文选取 100 次迭代后的训练模型作为最优预测模型。按照最近 30 天预测下一天的滚动预测方式，利用已建立的 LSTM 预测模型对预测区间内的高频、低频和趋势项 3 个子序列进行预测。表 2 给出了各子序列预测模型的多指标评价结果：高频子序列预测模型的 R^2 只有 0.408 1，表明该模型解释能力不足、预测误差较大，结合图 8 所示，高频子序列预测值相对真实值的右偏特征，表明高频子序列预测模型存在明显的滞后问题，需进一步改进；而低频和趋势子序列预测模型对应的 R^2 都已超过 0.997，表明两者的预测值和实际值均已非常接近，预测效果出色。

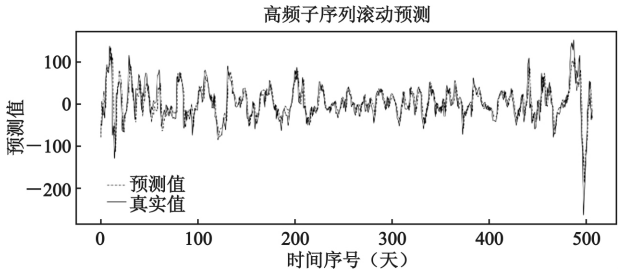


图 8　高频子序列的 LSTM 滚动预测结果

表 1　不同神经元个数条件下高频分量预测模型的评估结果

隐藏层神经元个数	模型评价指标			
	R^2	EVS	RMSE	MAE
(6,6)	0.397	0.401	28.166	19.640
(8,8)	0.406	0.437	28.982	20.348
(10,10)	0.462	0.463	28.006	19.518
(16,16)	0.420	0.427	28.149	19.998
(20,20)	0.455	0.455	27.751	19.291
(26,26)	0.432	0.433	27.634	19.467
(32,32)	0.414	0.414	27.972	19.455

表 2　各子序列预测模型的多指标评价结果

序列	R^2	EVS	RMSE	MAE
高频子序列	0.408 1	0.414 8	27.738 4	19.678 6
低频子序列	0.999 5	0.999 5	8.523 9	7.122 1
趋势子序列	0.997 3	0.999 8	4.091 0	3.959 4

2. 高频子序列重组中 IMF 组合方式的优化。重组构成高频子序列的 IMF1～IMF4 在包含股市指数波动特征信息的同时携带大量的噪声。因此，以 IMF1～IMF4 的不同子集重组产生的高频子序列，也会同时包含指数的波动特征信息以及不同比例的噪声。显然，在 IMF1～IMF4 中，IMF3 和 IMF4 的频率相对更低，包含指数的波动信息相对更多，而 IMF1、IMF2 则含有更多的噪声。因此，优化后重组形成高频子序列的 IMF 集合，至少应包含 IMF3 和 IMF4。进一步采用本文所确定的子序列 LSTM 预测模型参数，并根据预测效果评估结果确定最优的高频 IMF 组合方式。如表 3 所示，在 4 种合理的 IMF 组合方式中，剔除 IMF1 后利用 IMF2～IMF4 重组产生高频子序列，进而构建的高频预测模型的预测效果最优。相对于第一种组合方式，最优组合方式的 R^2 值、EVS、RMSE、MAE 分别提升了 116.3%、113.1%、47.6%、45.4%。图 9 为最优组合方式下高频子序列的滚动预测结果，其滞后性比图 8 中高频子序列的预测结果有明显改善。

表 3　不同 IMF 组合方式下高频预测模型的评估结果

IMF 组合的序号	R^2	EVS	RMSE	MAE
1+2+3+4	0.408 1	0.414 8	27.738 4	19.678 6
2+3+4	0.883 9	0.883 9	14.544 0	10.754 1
1+3+4	0.220 2	0.227 5	30.766 7	22.108 7
3+4	0.536 8	0.539 9	25.202 1	18.213 5

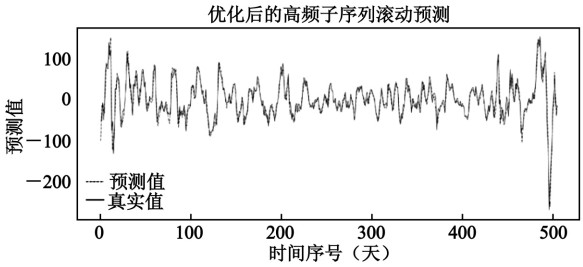


图 9　最优组合方式下高频子序列的预测结果

（五）指数整体预测值的加和集成及效果评价

将组合方式优化后产生的高频、低频与趋势项 3 个子序列的预测值加和，获得优化后的指数集成预测结果，如图 10 所示。可观察到未经优化的预测值相对于真实值存在较明显的整体右偏缺陷，表明未经优化的预测值具有一定的滞后性，而优化后的预测值明显更加贴近真实值。表 4 给出了优化前后指数集成预测效果的多指标评估结果：R^2 值提升了 0.5%，EVS 提升了 0.5%，RMSE 提升了 43%，MAE 提升了 40.3%。这表明，IMF 组合方式优化处理显著降低了模型预测的滞后性，并提升了预测精确度。

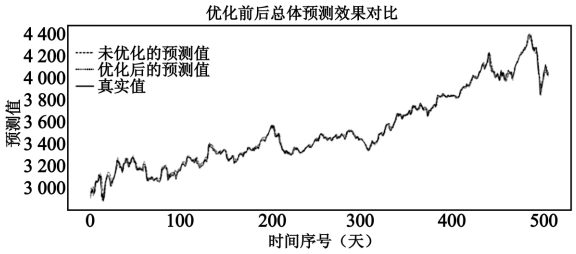


图 10　优化前后指数的 CEEMDAN-LSTM 集成预测结果

表 4　优化前后指数的集成预测效果的评估结果

评价指标	R^2	EVS	RMSE	MAE
优化前	0.992 4	0.992 5	29.272 2	21.104 7
优化后	0.997 6	0.997 6	16.678 7	12.593 0
提升比率	0.5%	0.5%	43%	40.3%

五、CEEMDAN-LSTM 模型预测效果的对比验证分析

本文选取包括沪深 300、上证综指等 5 个最具代表性的国内股市指数为实验数据，并以已经研究证实预测效果较突出的指数预测建模方法[5,21]，包括多层感知器 MLP、支持向量回归 SVR 和 LSTM，作为实验的对比方法，对 CEEMDAN-LSTM 模型的预测有效性、适应性进行评估。

（一）基于沪深 300 指数的预测效果对比分析

以本文所述训练期内的沪深 300 指数时序数据为输入，分别运用 MLP、SVR 和 LSTM 三种指数预

测建模方法，直接针对指数序列数据构建相应的指数预测模型，然后以可视化方式呈现各种预测模型的滚动预测效果，并对各模型的预测效果进行采用多指标评估与对比分析，以客观地评估 CEEMDAN-LSTM 指数预测建模方法的有效性。

多层感知器（MLP）是一种前向结构的人工神经网络（ANN），其中包含多个节点层，每个节点代表一个带有非线性激活函数的神经元。MLP 是一个有向图，每一层都全连接到下一层，能将一组输入向量映射到输出向量，通常采用反向传播 BP 算法训练网络权值。本文通过交叉验证与网格搜索方法设定 MLP 参数的最优值：正则化惩罚项系数 α 为 0.1，隐层层数为 3，相应的节点数为（13，23，9），激活函数为“tanh”，优化算法为“lbfgs”。图 11 给出了指数 MLP 预测模型的滚动预测结果，存在一定的预测滞后性。

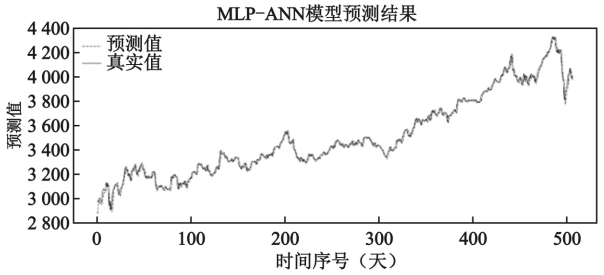


图 11　MLP 模型的滚动预测结果

按照本文所述 LSTM 预测建模方法，针对沪深 300 指数序列的前 2 450 个数据，创建训练集并直接建立基于 LSTM 的指数序列预测模型，其参数设定参照前文所述的子序列 LSTM 预测模型参数。直接通过 LSTM 建模的指数预测结果如图 12 所示，其中预测值相对真实值有一定右偏，表明该预测方法存在预测滞后性问题。

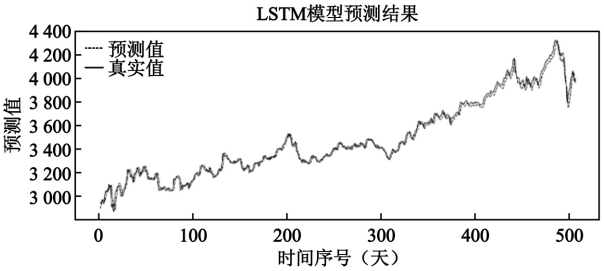


图 12　LSTM 模型的指数滚动预测结果

参照贺毅岳等关于股市指数 SVR 预测建模的研究结果，以指数时序数据为基础创建训练集和测试集[21]。在对指数的 SVR 建模过程中，为避免模型超参数较多导致参数搜索计算代价过高的问题，本文限定系数 γ 和惩罚系数 C 的搜索区间为[0.01，20]，待选核函数为：多项式核、线性核函、高斯核，进一步采用随机参数优化方法进行参数寻优实验，搜索到 SVR 预测模型中最优的核函数为“linear”、惩罚系数 α 为 20，其余参数设定为 Sklearn 库中 SVR 函数提供的默认值。图 13 给出了基于 SVR 的指数滚动预测结果，仍存在一定的预测滞后性。

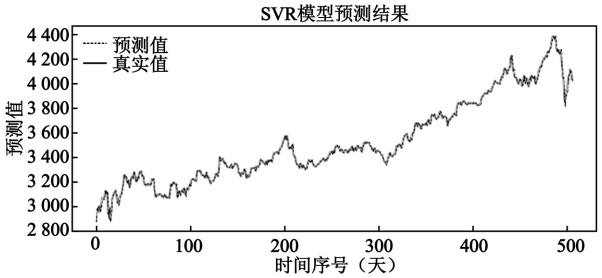


图 13　SVR 模型的指数滚动预测结果

在 CEEMDAN 分解的基础上进一步构建预测模型 CEEMDAN-MLP 和 CEEMDAN-SVR，在模型参数寻优过程中均采用了随机参数优化方法[3]。图 14 给出了本文方法及上述 5 种对比建模方法在预测区间的前 100 个指数值上的滚动预测对比结果。表 5 进一步给出了各预测方法的多指标评估结果，其中 CEEMDAN-LSTM 的 R^2 和 EVS 最大，RMSE 和 MAE 最小，其在所有评估指标上一致优于其他 5 种对比方法。表 5 中模型 CEEMDAN-MLP 和 CEEMDAN-SVR 显著优于 SVR 和 MLP 直接建模的预测效果，也证实了对指数进行 CEEMDAN 分解与重组处理能显著提升建模的精确度。这表明通过 CEEMDAN 分解与重组产生子序列，再建模预测并集成最终预测值的思路是合理有效的。

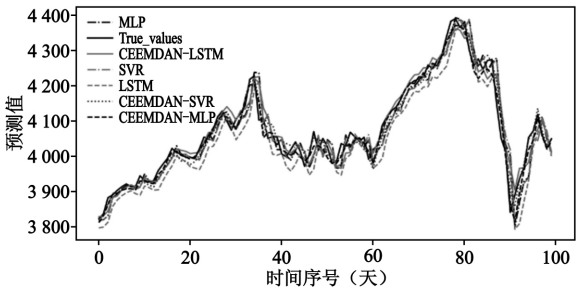


图 14　六种建模方法的滚动预测结果对比

表 5　不同预测建模方法在沪深 300 上的评估结果

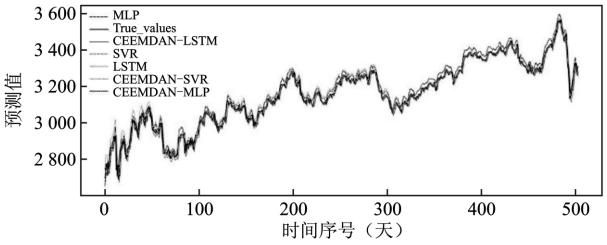
指数建模方法	R^2	EVS	RMSE	MAE
CEEMDAN-LSTM	0.997 6	0.997 6	16.678 7	12.593 0
LSTM	0.987 8	0.988 3	36.597 2	26.923 7
SVR	0.990 6	0.990 8	33.366 9	23.420 5
MLP	0.990 6	0.990 8	33.351 8	23.621 5
CEEMDAN-SVR	0.996 8	0.997 4	19.508 0	14.643 9
CEEMDAN-MLP	0.995 9	0.995 9	20.987 8	16.870 4

(二)基于上证综指等四个典型指数的预测效果对比分析

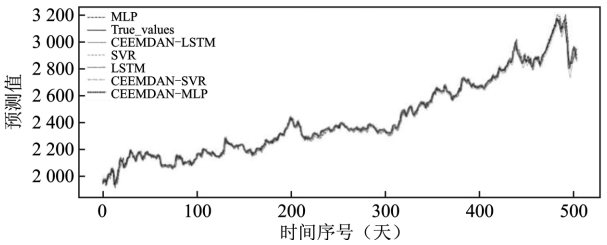
利用 Python 从聚宽量化平台在线提取 2006 年 1 月 1 日至 2018 年 2 月 1 日之间上证综指、上证 50、深圳成指 3 个典型股市指数的收盘价,提取 2008 年 1 月 1 日至 2018 年 2 月 1 日之间中证 500 指数的收盘价,剔除节假日等因素的影响,前 3 个指数均含有 2 955 个数据,中证 500 含有 2 472 个数据,作为指数预测建模的输入数据。对上述 4 个指数的统计性质分析与检验表明:与沪深 300 指数类似,上述 4 个指数包含大量的噪声,具有显著的非正态、非平稳特征,对应的收益率序列波动聚集性显著,直接应用传统的计量方法难以获得高精度的预测效果,因而选用 CEEMDAN-LSTM 对各指数进行预测建模。

分别以上述 4 种指数的 2016 年 2 月 1 日之前共 2 450 个数据(中证 500 前 1 967 个数据)作为建模输入数据,参照前文所述沪深 300 指数序列 CEEMDAN-LSTM 建模过程,采用滚动预测建模方式,以最近 30 天的指数值为输入变量来预测下一天的指数值,依次通过指数序列的 CEEMDAN 分解和重组、子序列 LSTM 预测建模及高频子序列重组中 IMF 组合方式优化、加和集成指数整体预测值等一系列建模步骤,构建出与每一种指数对应的 CEEMDAN-LSTM 预测模型。同时,参照本文基于沪深 300 指数的预测效果对比分析部分所述,针对上述每一种指数,采用滚动预测建模方式,分别运用 MLP、SVR、LSTM、CEEMDAN-SVR 和 CEEMDAN-MLP 建模方法,构建 5 个对应的指数预测对比模型。然后,以每一个指数 2016 年 2 月 2 日至 2018 年 2 月 1 日共 505 个数据构建滚动预测的测试集,分别利用上述 6 种预测模型进行周期为 30 天的按日滚动预测,以对比分析各模型的预测效果。

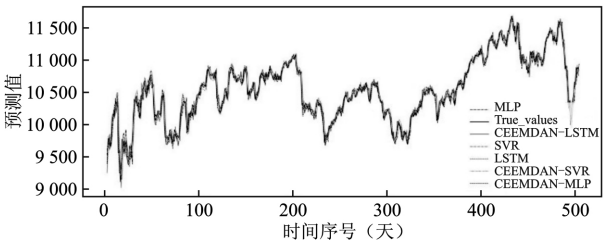
图 15 依次给出了上述 6 种建模方法在上证综指、上证 50、深圳成指和中证 500 四个指数预测区间上的滚动预测对比结果。其中,相比其他对比建模方法的预测曲线,CEEMDAN-LSTM 预测曲线与原指数曲线贴合最紧密,时间滞后性最弱。进一步地,表 6～9 依次给出了各模型在四个指数上滚动预测效果的多指标评估结果,其数据证实,在对上述每一个指数的预测表现中,相对于其他 5 种对比建模方法,CEEMDAN-LSTM 的 R^2 和 EVS 值均最大,而 RMSE 和 MAE 值均最小,即其在所有评估指标上一致优于包括 LSTM 在内的其他 5 种对比方法;同时,在对每一个指数的预测表现中,CEEMDAN-SVR 和 CEEMDAN-MLP 又分别优于 SVR 和 MLP 直接建模,证实了对指数进行 CEEMDAN 分解与重组处理能显著提升进一步预测建模的有效性。因此,将 CEEMDAN 的自适应分解功能与 LSTM 的长期依赖关系提取能力结合运用,进而构建股市指数的高精度混合预测模型的思路是合理、有效的。



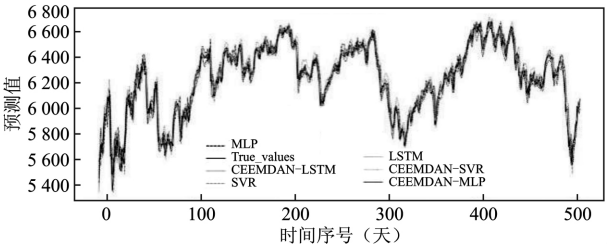
(a)上证综指



(b)上证 50



(c)深圳成指



(d)中证 500

图 15　6 种建模方法在 4 个典型股市指数上的滚动预测效果对比

表 6　不同预测建模方法在上证综指上的评估结果

指数建模方法	R^2	EVS	RMSE	MAE
CEEMDAN-LSTM	0.992 8	0.994 0	15.293 0	11.050 1
LSTM	0.970 1	0.970 8	30.029 6	21.269 3
SVR	0.960 3	0.992 6	34.278 1	31.063 4
MLP	0.973 9	0.974 0	28.360 2	19.472 4
CEEMDAN-SVR	0.972 4	0.973 6	29.285 7	20.619 0
CEEMDAN-MLP	0.977 3	0.992 8	27.034 8	24.079 4

表 7　不同预测建模方法在上证 50 上的评估结果

指数建模方法	R^2	EVS	RMSE	MAE
CEEMDAN-LSTM	0.997 6	0.997 9	14.314 0	9.809 0
LSTM	0.991 0	0.992 2	27.353 9	19.694 6
SVR	0.984 1	0.984 1	22.609 8	15.784 1
MLP	0.981 7	0.982 4	27.310 9	19.884 6
CEEMDAN-SVR	0.996 5	0.997 4	17.186 7	12.855 1
CEEMDAN-MLP	0.987 6	0.987 6	14.481 8	9.911 3

表 8　不同预测建模方法在深圳成指上的评估结果

指数建模方法	R^2	EVS	RMSE	MAE
CEEMDAN-LSTM	0.982 9	0.983 6	63.221 0	48.041 0
LSTM	0.923 2	0.923 3	133.030 6	92.263 2
SVR	0.929 6	0.929 6	128.041 5	90.237 2
MLP	0.928 3	0.928 3	130.646 6	92.700 6
CEEMDAN-SVR	0.977 6	0.979 2	70.987 3	55.611 0
CEEMDAN-MLP	0.978 9	0.981 1	71.090 7	53.867 1

表 9　不同预测建模方法在中证 500 上的评估结果

指数建模方法	R^2	EVS	RMSE	MAE
CEEMDAN-LSTM	0.987 5	0.988 8	30.185 7	22.095 2
LSTM	0.880 6	0.910 2	93.495 9	67.323 7
SVR	0.913 3	0.914 0	82.256 5	58.415 7
MLP	0.905 7	0.909 7	85.326 7	58.817 0
CEEMDAN-SVR	0.979 8	0.978 0	39.371 0	29.746 0
CEEMDAN-MLP	0.969 5	0.985 7	46.844 5	39.997 6

六、结论及展望

本文针对股票市场指数预测建模这一金融投资领域的核心问题，运用 CEEMDAN 分解与重组产生波动特征更简单的高频、低频及趋势子序列，为进一步构建子序列预测模型充分提取子序列的复杂波动模式创造了有利条件，显著降低了指数序列高精度预测建模的难度，使得本文通过 CEEMDAN 进行指数分解与重组后分别构建预测模型，进而加和集成获得指数整体预测值的思路合理、可行。本文在对指数 CEEMDAN 分解与重组的基础上，充分利用 LSTM 对复杂序列中长期依赖关系高效提取的优势，提出并详细阐述了一种 CEEMDAN 和 LSTM 结合的股市指数集成预测建模方法 CEEMDAN-LSTM。最后，选取了包括沪深 300、上证综指等 5 个最具代表性的国内股市指数为实验数据，并以经研究证实预测效果较突出的主流机器学习指数建模方法，包括 MLP、SVR 与 LSTM，作为实验的对比方法，对 CEEMDAN-LSTM 模型的预测有效性、适应性进行多维度量化评估。实验结果证实，CEEMDAN-LSTM 的预测表现一致性地优于现有建模方法，其预测结果误差小、精度高，且相对真实指数值具有更低的时间滞后性。然而，本文在 CEEMDAN-LSTM 的建模过程中，对网络隐藏层个数等部分参数的选取仍具有一定的主观性；同时，双层 LSTM 单元未必能充分挖掘出非线性复杂指数序列中蕴含的深层次变化模式信息，故还需进一步研究模型参数的最优化处理。

参考文献：

[1] Chen Y W, Chou R K, Lin C B. Investor Sentiment, SEO Market Timing, and Stock Price Performance[J]. Journal of Empirical Finance, 2019, 51(C): 28-43.

[2] 谢合亮，游涛. 基于深度学习算法的欧式股指期权定价研究——来自 50ETF 期权市场的证据[J]. 统计与信息论坛，2018，33(6)：99-106.

[3] 苏治，卢曼，李德轩. 深度学习的金融实证应用：动态、贡献与展望[J]. 金融研究，2017，443(5)：111-126.

[4] 陈卫华，徐国祥. 基于深度学习和股票论坛数据的股市波动率预测精度研究[J]. 管理世界，2018，34(1)：180-181.

[5] 杨青，王晨蔚. 基于深度学习 LSTM 神经网络的全球股票指数预测研究[J]. 统计研究，2019，36(3)：65-77.

[6] LeCun Y, Bengio Y, Hinton G. Deep Learning[J]. Nature, 2015, 521(7553): 436-444.

[7] Mabu S, Hirasawa K, Obayashi M, et al. Enhanced Decision Making Mechanism of Rule-based Genetic Network Programming for Creating Stock Trading Signals[J]. Expert Systems with Applications, 2013, 40(16): 6311-6320.

[8] Wang F, Yu L H, Cheung D W. Combining Technical Trading Rules Using Particle Swarm Optimization[J]. Expert Systems with Applications, 2014, 41(6): 3016-3026.

[9] 梁淇俊，郑贵俊，徐守萍. 基于生存分析的择时策略择优体系研究——以技术指标交易信号为例[J]. 金融经济学研究，2015，30(1)：96-106.

[10] Hassan M R. A Combination of Hidden Markov Model and Fuzzy Model for Stock Market Forecasting[J]. Neurocomputing, 2009, 72(16): 3439-3446.

[11] 张超. 基于误差校正的 ARMA-GARCH 股票价格预测[J]. 南京航空航天大学学报(社会科学版)，2014，16(3)：43-48.

[12] 张蓓. 高斯隐马尔科夫模型在金融预测中的应用[D]. 武汉：华中科技大学，2018.

[13] Tay F H, Cao L J. Application of Support Vector Machines in Financial Time Series Forecasting[J]. Omega, 2001, 29(4):

309-317.

[14] Chen Y, Hao Y. A Feature Weighted Support Vector Machine and K-nearest Neighbor Algorithm for Stock Market Indices Prediction[J]. Expert Systems with Applications, 2017, 80(Sep): 340-355.

[15] Bao W, Yue J, Rao Y. A Deep Learning Framework for Financial Time Series Using Stacked Autoencoders and Long-short Term Memory [J/OL]. PLOS ONE, 2017, 12 (7): e0180944. [2019-08-20]. https://doi. org/10. 1371/journal. pone. 0180944.

[16] Thomas F, Christopher K. Deep Learning with Long Short-term Memory Networks for Financial Market Predictions[J]. European Journal of Operational Research, 2018, 270(2): 654-669.

[17] Huang N E, Shen Z, Long S R, et al. The Empirical Mode Decomposition and the Hilbert Spectrum for Nonlinear and Non-stationary Time Series Analysis[J]. Proceedings of the Royal Society of London, 1998, 454(1971): 903-995.

[18] Wu Z H, Huang N E. A Study of Characteristics of White Noise Using the Empirical Model Decomposition Method[J]. Proceedings of the Royal Society of London, 2004, 460(2046): 1597-1611.

[19] Torres M E, Colominas M A, Schlotthuer G, et al. A Complete Ensemble Empirical Mode Decomposition with Adaptive Noise[J]. Brain Research Bulletin, 2011, 125(3): 4144-4147.

[20] Yang H L, Lin H C. Applying the Hybrid Model of EMD, PSR, and ELM to Exchange Rates Forecasting [J]. Computational Economics, 2017, 49(1): 1-18.

[21] 贺毅岳,高妮,王峰虎,等. EMD 分解下基于 SVR 的股票价格集成预测[J]. 西北大学学报(自然科学版), 2019, 49(3): 329-336.

[22] 李合龙,冯春蛾. 基于 EEMD 的投资者情绪与股指波动的关系研究[J]. 系统工程理论与实践, 2014, 34(10): 2495-2503.

[23] Zhang X, Zhang Q, Zhang G, et al. A Novel Hybrid Data-Driven Model for Daily Land Surface Temperature Forecasting Using Long Short-Term Memory Neural Network Based on Ensemble Empirical Mode Decomposition[J]. International Journal of Environmental Research & Public Health, 2018, 15(5): 1-23.

Research on Predictive Modeling on Stock Market Index Based on CEEMDAN-LSTM

HE Yi-yue[1], LI Ping[2], HAN Jin-bo[1]

(1. School of Economic & Management, Northwest University, Xi'an 710127, China;
2. Xi'an University of Finance and Economics, Xi'an 710100, China)

Abstract: In order to meet the demand of high-precision forecast of stock market index in the field of active quantitative investment, complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) is introduced into predictive modeling on stock market index. Combing CEEMDAN with Long Short-Term Memory (LSTM), which can model long-term dependencies of complex time series efficiently, an integrated prediction method CEEMDAN-LSTM for stock market index is proposed following "decomposition-reassembly-prediction-integration" process. In the method, CEEMDAN is applied to decompose the index and reconstruct its high frequency component, low frequency component and trend items. Then, the LSTM prediction model is established for each component respectively, and the predicted values of each component are furtherly integrated as the overall predicted index value. Finally, comparative experiments are conducted and the prevalent machine learning modeling methods of financial time series, with five representative stock market indexes including CSI 300 as test data. The experimental results prove that CEEMDAN-LSTM has lower prediction error and lag, and its prediction performance is significantly better than that of the existing modeling methods.

Key words: quantitative investment; stock market index; predictive modeling; CEEMDAN; LSTM

(责任编辑:崔国平)