目录

[1 绪论 3](#_Toc17309834)

[1.1 研究背景、意义及价值 3](#_Toc17309835)

[1.2 文献综述 4](#_Toc17309836)

[1.2.1 股票指数时间序列的预测研究文献综述 4](#_Toc17309837)

[1.2.2 经验模态分解时间序列的应用研究文献综述 7](#_Toc17309838)

[1.3 本文主要内容与创新 8](#_Toc17309839)

[2 EMD算法概述 8](#_Toc17309840)

[2.1 EMD的概念和基本算法 8](#_Toc17309841)

[2.1.1 EMD算法的基本概述 8](#_Toc17309842)

[2.1.2 EMD算法的停止准则 10](#_Toc17309843)

[2.1.3 EMD算法的端点问题 11](#_Toc17309844)

[2.2 EMD算法的改进研究 13](#_Toc17309845)

[2.2.1 集成经验模态分解 13](#_Toc17309846)

[2.2.2 完整集成经验模态分解 14](#_Toc17309847)

[2.2.3 具有自适应噪声的完整集成经验模态分解 15](#_Toc17309848)

[2.3 本章小结 16](#_Toc17309849)

[3 基于经验模态分解和神经网络的混合预测模型 16](#_Toc17309850)

[3.1 相关算法介绍 17](#_Toc17309851)

[3.1.1 ANN算法介绍 17](#_Toc17309852)

[3.2 EMD与ANN混合预测模型及改进 18](#_Toc17309853)

[3.2.1 基于EMD与ANN的预测模型 18](#_Toc17309854)

[3.2.2 基于EMD与ANN的单一模型预测结构 20](#_Toc17309855)

[3.3 实证分析研究 21](#_Toc17309856)

[3.3.1 数据介绍 21](#_Toc17309857)

[3.3.2 基于EMD的数据分解 22](#_Toc17309858)

[3.3.3 模型性能评价指标 25](#_Toc17309859)

[3.3.4 实证结果分析 28](#_Toc17309860)

[3.3.5 同类研究对比 33](#_Toc17309861)

[3.4 本章小结 34](#_Toc17309862)

[4 基于经验模态分解与神经网络的自适应预测模型 34](#_Toc17309863)

[4.1 EMD算法中的前视偏差分析 34](#_Toc17309864)

[4.2 EMD与ANN混合的自适应预测模型 36](#_Toc17309865)

[4.2.1 基于EMD与ANN的自适应预测模型 36](#_Toc17309866)

[4.2.2 基于EMD与ANN自适应预测模型的改进 37](#_Toc17309867)

[4.3 实证分析 40](#_Toc17309868)

[4.4 本章小结 45](#_Toc17309869)

[5 结论与展望 45](#_Toc17309870)

[5.1 本文结论 45](#_Toc17309871)

[5.2 研究展望 46](#_Toc17309872)

1. 绪论

1.1 研究背景、意义及价值

时间序列是按照时间为顺序记录的一系列数据点，是一种非常常见的数据形式。理想的时间序列数据都存在着一定的内在规律，表现出一定的特性。而实际的时间序列数据常常都会受到外部因素的影响，使得在分析和寻找其内在规律和特性时存在一定的困难。在现实中有很多问题的分析都会归结为时间序列的分析处理问题，如金融数据，能源数据，气象数据，GDP数据，网络流量数据，电力数据，一些医疗数据等。这些变量随着时间的推移而变化包含着大量与观测系统相关的信息，比如频率特征和时间特征等。通过对时间序列的分析，我们能够更好的认识时间序列所代表的意义以及解释其内在的变动规律，从中找寻有价值的信息从而指导我们做出正确的决策。

近年来，国内外大量的学者致力于研究时间序列分析模型，并且不断地把新的思想和方法融入到新的模型结构中。同时各个领域的学者也在不断地把新方法运用到各个领域里，因此时间序列分析及应用得以在各个领域迅速发展。如生物信息学，遗传学，多媒体，社会学，经济学，金融学等。近十年来，在数据挖掘领域里，积累了大量的时间序列数据，以及各个平台开源数据和技术，也为时间序列数据的研究和发展打下了坚实的基础。目前时间序列的研究主要集中于序列匹配，模式识别，异常检验，主题发现，索引，聚类，分类，可视化，分割，趋势分析，相似检验，自动文摘和长短预测[57]。作为时间序列分析的主要用途之一，时间序列预测是统计学、经济学和管理学等研究中的热点和难点[58]。通过对被预测事物历史值的分析和研究，发现其内在规律并建立时间序列预测模型对未来时刻的值进行预测。本文的主要研究内容就是有关金融时间序列中较常见的股票市场指数预测的问题。

金融市场在现代社会经济中扮演着极其重要的角色，经济活动影响着世界各国的经济发展。预测金融资产的未来价格一直是金融市场上的一个经久不衰的话题，因为准确的预测金融市场的未来的表现情况无论对个人投资者、机构投资者还是政府金融机构都具有显著的意义。他们可以通过提前预估未来市场的走势而提前指定相关策略来及时调整资产配置降低决策风险。股票市场指数的预测是投资者和研究人员最重要，最具有挑战性的金融时间序列预测问题之一。尤其是近来量化投资的火热又把金融时间序列的处理和预测分析推向了一个新的高潮，再加上机器学习的快速发展，使得越来越多的学者投身于机器学习在金融领域的研究之中。

早期的金融时间序列主要通过统计学手段分析其随时间的变化。早期预测模型主要有随机游走模型，指数平滑（ESM）、自回归（AR）、移动平均法（MA）、加权移动平均法、自回归移动平滑模型（ARMA）、自回归差分移动平滑模型（ARIMA）、平滑过渡自回归模型[54]（STAR）、向量自回归（VAR）模型、自回归条件异方差模型（ARCH），广义自回归条件异方差模型（GARCH）。大量的研究证明，GARCH模型在实际应用中可以更加精准的对价格以及价格波动率进行有效的分析和预测。目前为止大量的股价预测、股指预测以及其它计量经济学方面的预测和研究仍然是以GARCH模型为基础的[59]。

机器学习技术的发展和计算机技术的快速发展，使得机器学习和深度学习模型在金融时间序列的分析和应用中变得越来越广泛。如树（TREE）模型，集成树（Ensemble Tree）模型，朴素贝叶斯（Naive Bayes）模型，支持向量机（SVM）模型，人工神经网络（ANN）以及基于上述模型的一些发展和改进模型（RNN、LSTM、GAN等）。大量学者的研究表明，机器学习和深度学习算法在金融时间序列的分析和预测中具有良好的表现性能。

本文主要基于EMD（经验模态分解）方法把原始时间序列分解成不同的时间序列，这些序列代表着原序列不同时域和频域的特征，然后对分解得到的序列进行建模，充分挖掘序列隐藏的信息。对于一个趋势性较强的时间序列往往在不同的时间内表现出不同的趋势，很难捕捉其内在的变化规律，利用经验模态分解方法可以把序列分解成不同的特征序列，来达到简化序列的目的。本文不同于其他学者的把所有分解出来的子序列都进行建模，然后再把模型集成。本文使用方法的特殊之处在于使用分解出来序列的特征构建一个模型，大大简化了模型，提高了效率。而且本文为了避免在分解序列时使用到未来的信息，近而建立基于EMD分解的自适应模型，来研究在不涉及未来信息的情况下经验模态分解是否对预测有帮助。

本文研究一个基于EMD的新的预测框架，可以明显提高预测效率。使用EMD进行序列分解把复杂的序列分解成若干简单的、具有不同特征的序列，然后依据不同序列的特征进行建模，简化传统的基于EMD的预测模型框架，提高模型的预测效率及模型预测精度，为基于EMD的时间序列处理问题提供新的思路。而且本文也着重研究了检验模态分解中前视性偏差对模型结果的影响并提出了一种消除前视性偏差的预测模型，使得模型的预测结果更加具有现实意义。

1.2 文献综述

1.2.1 股票指数时间序列的预测研究文献综述

股票市场指数的预测是投资者和研究人员最重要也是最具有挑战性的金融时间序列预测问题之一[19]。关于股票市场的是否可预测性的问题由来已久，早在1900年Bachelier第一次用一个随机游走方式来表征股票价格的变动。后来也有一部分学者凭借自己的经验去测试价格变化的随机游走特征[42-45]。1970年Malkiel和Fama[46]基于有效市场假说对股票市场进行了一些研究。有效市场假说表明所有新信息都会立刻反应到资产价格中，因此未来资产价格变动于过去和现在的信息无关。然而很多研究者通过实验反驳了市场有效假说，经验证据表明股票市场在某种程度上是可以预测的。如文献[47-53]中等研究者和学者的实验验证都表明股票市场在一定程度上是可预测的，并且不断涌现的基本面分析方法、技术分析方法以及大量研究股票市场走势、回报率以及波动率的文章[1-41]，也间接表明了股票市场在一定程度上是可以预测的。

传统的时间序列预测方法主要是一些计量经济学模型如指数平滑（ESM）、自回归（AR）、移动平均法（MA）、加权移动平均法、自回归移动平滑模型（ARMA）、自回归差分移动平滑模型（ARIMA）、平滑过渡自回归模型（STAR[54]）、向量自回归（VAR）、自回归条件异方差模型（ARCH），广义自回归条件异方差模型（GARCH）这些模型基本思想是一致的，都是基于一定的假设，或对数据进行变换，然后建立线性模型。例如ARIMA就是对数据差分后的ARMA模型，GARCH是对收益或者收益残差和波动率进行平方然后对取平方得到的数据进行回归预测分析。

Ariyo, Adewumi and Ayo(2014)[55,13]用ARIMA预测纽约证券交易所指数（NYSE）和尼日利亚证券交易所指数（NSE）,他们认为ARIMA模型在股票指数的短期预测中具有特殊的潜力，并且他们又对比研究了ARIMA和人工神经网络（ANN）来预测NYSE指数。E. Chong, C. Han and F. C. Park(2017)[7]等比较了AR模型和以及应用主成分分析（PCA）、受限的玻尔兹曼机（RBM）和自编码（AE）的降维方法的ANN以及DNN对一些股票价格进行预测分析，他们相信没有理由认为机器学习得到的结果比自回归得到的结果更加具有优越性。M. Kumar and M. Thenmozhi (2009)[35] 收益率变换对数据处理，然后使用ARMA模型、ANN、以及ANN与ARMA的混合模型来对NIFTY指数进行预测。

Ju-Jie Wang等(2012)[27]利用指数平滑模型（ESM）、ARIMA、ANN以及用遗传算法（GA）优化的混合模型来对道琼斯指数（DJIAI）和深圳综合指数（SZII）进行预测研究。Ha Young Kim等(2018)[6]研究GARCH模型、LSTM模型以及他们的混合模型在股票价格波动率的预测中的应用，他们认为将LSTM模型的与多个GARCH相结合可以提高预测效果。传统的统计模型ARMA和ARCH一直在金融时间序列中起着及其重要的作用，绝大多是计量经济学模型都与他们有关。而且ARCH模型作为获得2003年诺贝尔经济学奖的计量经济学成果之一，足以表明其在时间序列处理中的地位。而且目前许多有关计量经济学的分析和预测研究仍然是以GARCH为基础。

然而，金融时间序列本质上是复杂的、嘈杂的、动态的、非平稳的、非线性的、非参数的和混沌的[1,16,19,,56]。大量的基于非线性处理模型的出现也为金融时间序列的分析和预测带来了新的处理方法。特别是一些传统机器学习模型和深度学习模型在金融时间序列中得到了非常广泛的应用。

Y. Baek and H. Y. Kim(2018)[1]利用数据增强提出了有两个LSTM模型组成的组合模型对标普500（S&P500）指数和KOSPI200指数进行预测研究，结果表明他们提出的有两个LSTM的组合模型由于一般的DNN、RNN和单独的LSTM模型。D. M. Q. Nelson, A. C. M. Pereira(2017)[10]利用LSTM模型预测巴西圣保罗IBovespa 指数的走势。并且与随机森林（RF）、多层感知机（MLP）和随机模型相比，LSTM的准确率更高。Y. Chen and Y. Hao(2017)[19]利用加权的支持向量机（SVM）和加权的最近邻（KNN）模型预测了短期、中期和长期上证综合指数和深圳综合指数以及指数走势。

还有些学者利用机器学习模型来分析国际市场间的相互影响对预测的作用如：L. S. Malagrino, N. T. Roman 和 A. M. Monteiro(2018)[20]利用朴素贝叶斯模型结合全球市场其它地区主要的股票市场指数的走势预测了巴西圣保罗IBovespa 指数的走势，虽然结果一般但是朴素贝叶斯模型利用解释和分析市场间的联系。同样M. Thenmozhi 和 G. Sarath Chand(2016)[4]利用支持向量机（SVM）回归分析研究了全球六个市场美国（道琼斯，标普500）英国（FTSE-100）印度（NSE）新加坡（SGX）香港（恒生）和中国（上证）股票市场指数之间的价格信息传递，实证结果表明加入了全球市场因子的预测结果高于仅仅只有滞后价格信息的预测结果。

还有一些学者在使用历史数据的前提下增加了技术指标或者是市场新闻情绪、投资者情绪等指标。如X. Zhang(2018)[2]加入在雪球网提取的情感指数利用SVM来预测上证指数的走势。B. Weng(2018)[15]使用在线数据源预测短期股票价格，结合历史股票价格、几项有名的技术指标、特定股票已发布新闻的数量和情绪分数、谷歌搜索给定股票的搜索趋势、维基百科特定页面访问量。使用神经网络（ANN）回归集成，支持向量机（SVM）回归集成，提升树（Boosting Tree）和随机深林（Random Forest）回归来预测股票的价格，实证表明提升树对于股票价格预测效果更好。

在机器学习快速发展的今天，大量的机器学习和深度学习算法已经渗入到金融时间序列的预测中如人工神经网络ANN（E. Guresen(2011)[3], E. Chong, C. Han and F. C. Park(2017)[7], M. Qiu, Y. Song(2018)[11], B. Weng(2018)[15], X. Zhong(2017)[16]）、支持向量机SVM（X. Zhang(2018)[2], M. Thenmozhi(2016)[4], B. Weng(2018)[15], Q. Xu(2019)[18], Y. Chen(2017)[19]），深度神经网络DNN（循环神经RNN，LSTM）（Y. Baek and H. Y. Kim(2018)[1], T. Fischer and C. Krauss(2018)[5], H. Y. Kim and C. H. Won(2018)[6], H. M(2018)[9], D. M. Q. Nelson[10]）等。

1.2.2 经验模态分解时间序列的应用研究文献综述

经验模态分解（Empirical Mode Decomposition , EMD）是1998年由美籍华人专家Huang在瞬时频率概念的基础上提出的一种新型自适应信号时频分析技术,是从时序的角度进行时间序列分析的工具，主要用于信号分解和去噪。在建立了EMD基本框架的情形下，Huang等人定义了本征模函数（Intrinsic Mode Function , IMF），提出了构建包络线和用连续均值筛选的过程，并进一步论证了EMD的正交性和完备性。经过十几年的发展，基于EMD的时频分析方法逐步形成了一套完备的理论体系，受到国内外诸多学者的广泛关注和研究[58]。EMD本质上可以被看作是一种滤波，将不同尺度的波动逐级进行分解。经验模态分解因为其前提假设少，适用性广，提出之后很快就被用于许多理工科领域[59]。

现有文献按照对子序列（IMF）是否使用相同的模型来说可以主要分为两类。

第一类是对不同的IMF使用不同的模型。Y. Xiang(2018)[61] 等利用EMD对降雨时间序列进行分解，然后对第一个IMF使用SVR进行预测，其他分量使用ANN进行预测，把所有的预测结果加总得到最后结果，他们认为分解后的时间序列能够得到更好的预测效果。田大中（2015[63] 利用EMD对网络流量序列分解，然后结合ARIMA和SVM分别对各分量的IMF和余项进行预测，然后将预测值利用ANN非线性叠加作为最终的预测结果。他认为经过组合的模型优于单个ARIMA和SVM模型。张文风（2018）[59]对使用经验模态分解后的子序列分别使用不同的模型进行预测，然后把预测值利用RBF（径向基）神经网络再次进行建模得到最终的预测值。

第二类也是最流行的一种方式，对每个IMF（包括余项）使用相同的模型进行预测。S. Huang(2014)[67]等对基于EMD分解的个子序列分别利用SVM进行建模预测，然后把预测的值直接进行相加。他们认为经过改进的EMD—SVM模型的预测效果优于单独的ANN的SVM模型。Z. Qu(2019)[68]等对风速时间序列进行EMD分解，并对高频的IMF进行二次分解，然后对每一个序列使用Flower-pollination算法优化的BPNN进行预测然后把预测结果相加。陈渝（2019）[60]等利用EMD分解门诊量序列，然后对每个分量IMF序列建立LSTM模型去预测门诊量，最后将结果直接相加。他们的结果表明用EMD进行分解数据时得到的组合模型的预测精度比不进行序列分解时的单一LSTM和单一SVM的预测精度更好。

H. Zhou(2019)[62]等利用EMD对股票指数走势进行预测研究，他们把分解的IMF利用因子分解机（Factorization machine）结合神经网络来预测上证综合指数、标普500以及纳斯达克综合指数的走势。E. Meng(2019)[64]等利用EMD分解水流量序列并结合SVM对水流量进行预测，他认为各分量的组合模型所得到的结果优于单个序列的ANN模型和SVM模型。N. Sun(2018)[65]提出了一种自适应的两层分解序列方法-结合EMD和VMD(变分模态分解, Variational Mode Decomposition, VMD)两种序列分解技术来分解序列，然后利用PCA降维技术分别选取各分量结合超限学习机（extreme learning machine, ELM）对风速进行预测，他们认为两阶段分解后的组合超限学习机模型优于序列单次分解的模型和序列不分解所构建的模型。Q. Tan(2108)[66]等提出了自适应EEMD与ANN的混合方法对水流量进行预测，他们认为在雨季的时候新提出的模型可以更准确的预测未来的水流量，但是在旱季的时候新提出的模型预测效果不如季节自回归（SAR）的预测效果。

1.3 本文主要内容与创新

2 EMD算法概述

经验模态分解(Empirical Mode Decomposition, EMD)是由Huang(1998) [73]等于1998年提出的一种处理非线性、非平稳性信号的方法, 其本质是对时间序列进行平稳化处理。与小波分解不同的是EMD是由数据驱动的一种序列分解方式，其更具原序列的极值特征直接对原序列进行分解不依赖于基函数。EMD本质上可以被看作是一种滤波，将不同尺度的波动逐级进行分解。经验模态分解因为其前提假设少，适用性广，提出之后很快就被用于许多理工科领域[59]。

2.1 EMD的概念和基本算法

2.1.1 EMD算法的基本概述

首先Huang在其文章中定义本征模态函数(Intrinsic mode functions)。当函数满足以下两个条件时，我们视其为一个本征模态函数：

（1）在整个数据集中，极值点的个数和过零点的个数必须相等或最多相差1个；

（2）在任意时间点，由局部极大值定义的包络线和由局部极小值定义的包络线的均值为零。

EMD分解是一种直观的、后验的、自适应的分解，完全由数据驱动。该方法的实质是对数据中固有振荡模态的特征时间尺度进行经验识别，然后对数据进行相应的分解。EMD分解基于以下假设：

（1）信号序列至少有两个极值，一个极大值一个极小值；

（2）时间尺度的特征由极值之间的时间间隔来定义；

（3）如果数据没有极值只有拐点，那么可以通过对数据差分来获取极值点。

从本质上讲经验模态分解过程是一个筛选过程，从该过程中依次筛选出频率由高到低的本征模态函数，并最终剩余一个不能再分解的单调残差序列，也称为趋势项。也就是说EMD可以自适应地把一个时间序列分解成有限个IMF和一个残差项。其算法描述如下：

Step1：输入原始序列信号；

Step2：获取原始序列的所有极大值点和极小值点，并分别对极大值点和极小值点用2条3样条插值曲线拟合，得到一条上包络线（upper envelope）和下包络线（lower envelope）。

Step3：计算包络线的均值；

 公式（2-1）

Step4：计算和的差别并令；

 公式（2-2）

Step5：检查是否满足上述的两个IMF的条件，如果是一个IMF或满足停止准则，则转到Step6；否则令，并重复Step1-4直到是一个IMF或满足停止准则。

Step6：计算残差项（residue），如果是一个单调函数或者只有一个极值点则分解完成；否则令，重复Step1-5。

至此，我们将原始信号分解为若干个本征模态函数和一个趋势项，用公式（2-3）表示：

 公式（2-3）

在上述的分解过程中，分解得到的本征模函数很容易满足第一个条件，即极值点的个数与过零点的个数相等或最多相差1个，然而第二个条件即局部上包络线和下包络线的均值为0却不一定能通过变换满足。而且当第二个条件发挥到极致时，可能会消除物理意义上的振幅波动，因此在筛选过程中要格外小心，因为如果将筛选进行到极致，可能使得到的IMF成为恒定的纯调频信号。为了确保IMF各组成部分保持足够的振幅和频率调制的物理意义，我们必须确定一个停止筛选过程的标准[73]。

2.1.2 EMD算法的停止准则

根据2.1.1中IMF的定义，严格IMF应满足局部极大值定义的包络函数和局部极小值定义的包络函数的均值恒为0，而在实际分解过程中很难做到分解出的上下包络之和严格等于0，只能接近于0。针对这种情况，一些学者通过实验给出了一些相对较为理想的停止准则也被称为局部包络条件。

EMD的提出者Huang(1998)[73]在其论文中给出了标准偏差(SD)准则，通过限制SD的值来判断是否停止筛选。SD是由两个连续的筛选结果计算得到的，计算过程用公式（2-4）表示[73]：

 公式（2-4）

其中为第个IMF第次迭代的值，计算公式见（2-2）。当SD小于研究者给定的某个值时，就认为该为第个本征模态函数。Huang在其文章中通过大量实验给出取值在0.2和03之间时分解效果较好。后来Huang等人又给出了另一个相似的停止准则用公式（2-5）表示：

 公式（2-5）

其中为第个IMF第次迭代的上下包络线的均值见（2-1）。然而，由于以下原因SD标准很难实现:首先，SD小到什么程度就足够小，这其实并没什么标准。其次，这个准则不依赖于IMF的定义，因为平方差很小但不能保证函数会有相同数量的过零点和极值。为了弥补这些缺点，Huang(1999,2003)等人提出了第二种类型的标准，称为S停止准则。使用这种类型的停止准则，筛选过程只有在零点个数和极值个数相等或最多相差1且连续S次保持相同时才停止。Huang(2003)等人的大量试验表明，S的最佳范围应该在3到8之间，但更倾向于较低的数字。显然，任何选择都是有特定的使用范围，需要一个严格的理由[74]。

法国学者Gabriel Rilling(2003)[75]等基于Huang等的研究提出了一个较为常用评价函数来控制筛分过程的停止

 公式（2-6）

其中为上下包络线的均值，。并设置三个阈值、、，当满足以下两个条件：

（1）当里面小于的比率达到；

（2）不存在大于的值。

则认为分解结束，中止筛选，一般取，和[75]。这也是本文中所采用的停止准则。

2.1.3 EMD算法的端点问题

在EMD相关算法的应用过程中，端点效应的处理是数据预处理过程中一个重要的步骤。所谓端点效应，是指在运用EMD对信号进行分解时，需要去“筛选”出本征模函数，而“筛选”过程需要对极值点使用三样条插值构造上下包络线，如果数据在端点处不是极值，那么在对极值点插值拟合时，将会出现误差从而产生端点效应。在端点附近时，三样条插值将出现发散现象，而随着本征模函数的一步一步提取，端点效应向内扩散最终导致污染数据[59]。Huang在其文章中指出在信号的两端，根据端点信号的振幅和频率，分别加两个特征波，可以有效抑制端点效应，但没有给出具体的做法。

处理端点问题最简单的方法是在迭代过程中不断地剔除非极值点的部分，但是对于预测问题来讲抛弃右端点的值也就意味着抛弃了近期的数据，然而近期的数据包往往是含着信息最大并且最重要的数据。所以如果是对时间序列值做简单的分析分解的话可以使用此方法，但是对于预测问题，往往不适合此方法。本文主要研究的是对股价指数的预测问题，不适合使用此方法。

另外一种解决端点问题的思想就是人为的在端点处添加一些数据，也称为延拓法。随着EMD算法的广泛应用，一些学者陆续的提出了一些较为经典的延拓方法，这些延拓方法大致可以概括为两种直接延拓法和预测延拓法[59]。

直接延拓法就是通过一定的技术手段将原有数据向外延拓。目前主要的延拓方法有，镜像延拓[76]、偶延拓、周期延拓[80]、极值延拓、平行线延拓、加窗口延拓[81]、、Coughlin法（Coughlin and Tung 2004）[79]、斜率法(Datig and Schlurmaim 2004)[77]和Rato法(Rato 2008)[78]等。

预测延拓法主要是指根据各种预测模型或方法对未来一些点进行外推预测的方法。常见的预测延拓方法有时间序列预测延拓，神经网络预测延拓，以及一些利用机器学习方法进行端点预测的方法。与直接延拓相比预测延拓往往需要更大的计算量，而且要去根据具体情况选择合适的预测方法，这样给延拓带来一些不确定性。

本文主要研究的是模型的简化和优化问题，并不关注这些延拓方法的区别。所以本文取最为常用并且简单的延拓方法—镜像延拓法来处理端点问题。镜像延拓法通过对原始数据序列邻近边缘的极值点添加镜像,以对称的方式增加n个极值点的方式来抑制分解过程中的端点效应。对于图2-1镜像延拓法过程如下：

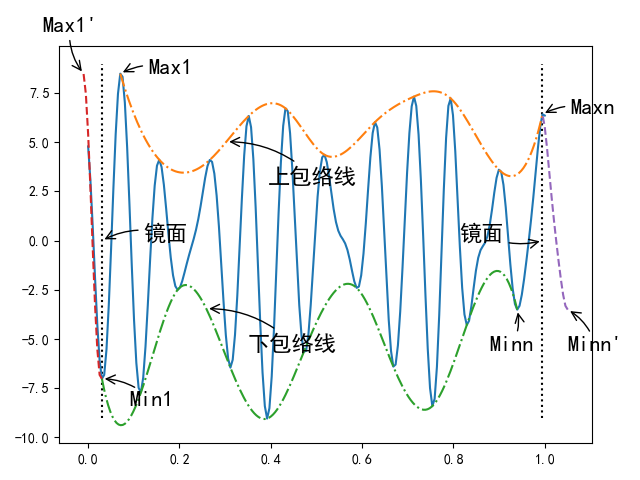


图2-1镜像延拓法过程

Step1：找到原始序列最左端的极值点Min1（第一个极小值点），以及最右端的极值点Maxn（最后一个极大值），如图2-1中所示；

Step2：分别以Min1和Maxn为镜面得到Max1的镜面对称点Max1’和Minn的镜面对称点Minn’,如图2.1所示；

Step3：最后根据数据原有的极大值点和极小值点加上有镜面对称得到的点，分别对运用三样条插值构建上包络线和下包络线。这样可以在一定程度上抑制端点的发散（图2.1给的上、下包络线为未进行延拓下的包络线）。

2.2 EMD算法的改进研究

因为停止条件的限制，在经验模态分解过程中可能会出现模态混叠，使得不同特征的序列分解成一个子序列。针对此类问题一些学者提出了一系列的改进算法。其中在处理信号序列中最常用的是一下介绍的EEMD算法和CEEMD算法。在其它序列数据处理中EMD算法和EEMD算法较为常用。本文研究的是有序列分解和简单神经网络混合模型的预测框架的优化问题，并不注重于不同的分解算法对结果的影响。所以本文选择最简单而且效率最高的EMD算法。

2.2.1 集成经验模态分解

针对经验模态分解存在的模态混叠现象，Wu and Huang(2009)[82] 等人提出集成经验模态分解(Ensemble Empirical Mode Decomposition, EEMD)算法，显著地提高了经验模态分解算法的应用价值。

集成经验模态分解（EEMD）算法的主要思想是：向原始的序列数据添加一个白噪声，然后对白噪声和原始序列的混合序列进行分解，重复N次此过程，将所得到的N组IMF和N个余项分别取平均值作为最后的分解的结果。集成经验模态分解过程中认为经过N次集成平均，序列中添加的白噪声可以相互抵消，同时克服了经验模态分解中出现的模态混叠现象，能够更好的实现信号序列的分解，是经验模态分解算法发展中的一个极其重要的改进。在2.1.1章节EMD算法的基础上，EEMD算法步骤如下：

Step1：在原始信号中添加一个白噪声信号，得到下列混合信号：

 公式（2-7）

其中，表示第次信号混合，，为信号混合的总次数，是一个人为设定参数，通常取[82]。

Step2：对混合后的信号序列进行经验模态分解，得到第次分解的子序列和其中为分解的个IMF（为了确保分解信号个数相同一般要求添加的白噪声为低信噪比的白噪声序列）。重复该步骤，直至。

Step3：对每个IMF和余项取平均，作为最终分解结果，得到第项IMF和余项：

 公式（2-8）

 公式（2-9）

Step4：得到最终的分解结果：

 公式（2-10）

其中，为分解的到的本征模函数的个数。

2.2.2 完整集成经验模态分解

根据2.2.1中的介绍，我们知道集成经验模态分解算法是通过添加噪声项来解决模态混叠问题。在引入噪声项后我们假设经过多次添加噪声项，噪声之间可以相互抵消。但实际上在信号重构时，引入的噪声项并不能完全抵消，一般都是通过增加集成的次数来减弱残留噪声项的相对大小，，但是这样做往往会大大的增加计算量。为解决经验模态分解过程中的模态混叠问题，同时消除集成分解过程中添加的白噪声对信号重构的影响, Yeh J R and Shieh J S(2010)[83]提出了完整集成经验模态分解(Complete Ensemble Empirical Mode Decomposition, CEEMD)。该算法的主要思想是成对的添加噪声，并且使得所添加的白噪声和为零。基于EMD和EEMD算法，CEEMD算法分解过程如下。

Step1：将白噪声成对的添加到原始信号中，得到一组待分解信号，并且该组信号的白噪声序列和为零：

 公式（2-11）

其中，表示原始信号序列，表示加入的白噪声，表示与正的白噪声叠加后的信号序列，表示与负的白噪声叠加后的信号序列。

Step2：分别对、进行EMD分解，重复N次。得到分解后的IMF子序列和余项：

 公式（2-12）

 公式（2-13）

其中，、和、分别表示第次混合信号、分解得到的子序列和余项。

Step3：最终，由CEEMD算法将原始信号分解为

 公式（2-14）

其中，为分解的到的本征模函数的个数。

2.2.3 具有自适应噪声的完整集成经验模态分解

EEMD和CEEMD都是通过对原序列添加噪声来处理模态混叠问题。有些学者考虑在分解过程中添加噪声来处理模态混叠问题，进而提出了具有自适应噪声的完整集成经验模态分解(Complete Ensemble Empirical Mode Decomposition with Adaptive Noise, CEEMDAN)算法，该算法的主要思想是：在序列分解过程中的每个阶段都添加自适应白噪声，通过计算唯一的余量信号获得IMF子序列，在较少的集成次数后就能够使重构误差几乎为零[84]。CEEMDAN算法过程如下（Humeau-Heurtier(2105)[85]）：

Step1：在原始信号序列中添加一些自适应白噪声：

 公式（2-15）

其中，表示第次加入白噪声后的信号，表示噪声系数（噪声标准差），表示第次添加的白噪声，表示集成次数。

Step2: 对每个信号，分别采用EMD算法分解获得第一个IMF分量，然后对这个分量取均值：

 公式（2-16）

其中，表示分解得到的第一个IMF信号。则余项信号为：

 公式（2-17）

Step3：定义为使用EMD算法分离出第个本征模函数的过程，分解，，直到分解出一个IMF。(在此步)在每一步允许选择的信噪比（SNR）。然后，得到第二个IMF：

 公式（2-18）

Step4: 计算第个余项：

 公式（2-19）

使用EMD算法分解，直到分解出一个IMF，定义第IMF：

 公式（2-20）

Step5：回到Step4，并令，直到余项不能再被分解。得到最终的余项：

 公式（2-21）

其中为分解得到的IMF个数。最终原信号序列被写作：

 公式（2-22）

2.3 本章小结

本章对经验模态分解算法进行了详细的介绍。介绍了经验模态分解算法的思想、基本原理以及算法实现过程。而且本章也对经验模态分解存在的问题以及其改进算法做了一些简单的介绍。本文并没有深入探讨不同的停止准则、不同的端点处理手段、不同的经验模态分解改进算法对结果的影响。本文主要的研究内容是经验模态分解与简单神经网络混合预测模型结构的改进。所以在此说明本文所采用的停止准则为2.1.2中提到的法国学者Gabriel Rilling(2003)[75]提出的停止准则，端点问题的处理依据2.1.3中详细介绍的而且较为常用的镜像延拓法，分解算法采用最简单并且效率最高的EMD算法分解序列。

3 基于经验模态分解和神经网络的混合预测模型

3.1 相关算法介绍

3.1.1 ANN算法介绍

人工神经网络(ANN)是一种模拟生物神经网络结构和功能的数学模型。因为其在变量间非线性关系强大的识别能力，在过去的几十年里得到了广泛的应用。在众多不同类型的人工神经网络中，反向传播人工神经网络(BP-ANN)由于其结构简单、易于实现，已被证明是解决许多问题的有力工具[66]。本文中使用三层神经网络预测股票指数，网络结构大致如图3-1所示。假设输入层有个节点，隐藏层有 个节点，输出层有一个节点。



图3-1 简单神经网络节点图

假设输入层有个节点，隐藏层有个节点，输出层有一个节点。则b1是维向量，b2是1维向量。对于隐藏层的节点来说

 公式（3-1）

其中，为个隐藏层的第个节点，为输入变量的权重值，为第个隐藏层节点的偏置，为激活函数。

 公式（3-2）

其中(图3.1只有一个输出，可以有多个输出)，为个输出层的第个节点，为隐藏变量的权重值，为第个输出节点的偏置，为激活函数。

3.2 EMD与ANN混合预测模型及改进

自1998年经验模态分解法提出以来，经验模态分解算法已经取得了较大的发展，并且理论知识日趋完善。经验模态分解因为其前提假设少，适用性广，提出之后很快就被用于许多理工科领域[59]。正如第一章节文献综述所介绍，使用EMD分解算法，然后对其子序列分别进行预测，最后在把预测结果相加这一预测模型框架在许多领域的研究都表明能取得较好的表现性能。然而这一预测模型框架，往往存在模型过多、计算量过大、参数难以调优等一系列问题。特别是使用EMD算法与机器学习算法结合时，例如EMD-ANN[59,61,62,63,66,68,71,72,84,86]、EMD-SVM[58,61,63,64,67,]、EMD-LATM[60]等都会出现模型过多所带来的计算量大、超参数过多难以对模型调优等问题。本文针对这些问题提出了一种改进的预测模型结构，并结合实证实验研究模型的表现性能。人工神经网络模型由于具有非线性、非局限性的优良性质，在进行组合预测模型中是最常用的基本模型之一，本章主要针对EMD与ANN模型的混合提出一种基于EMD的单模型ANN预测模型。

3.2.1 基于EMD与ANN的预测模型

现实中许多时间序列都具有高度复杂的特征，其高度非线性非平稳性性等一系列性质，使得直接对序列分析变得比较困难，也难以用具体模型去捕捉时间序列的特性，进而难以建立精确的预测模型。近年来对时间序列的研究越来越广泛，一系列对时间序列的处理方法相继被提出，如单尺度分析、多尺度分析、DFA趋势分析、MF-DFA趋势分析[57]、先分解后组合[84]分析等一些方法，本文多研究的就是基于“先分解后组合”思想的一种方法。先分解后组合的模型框架即首先将复杂的时间序列使用一些分解算法分解成若干子序列，然后再对其子序列进行建模分析预测。已有文献所给出的预测模型结构大多和本文使用的组合预测模型EMD-ANN相似如图3-2所示，只是基础预测模型和分解算法略有不同。



图3-2 EMD与ANN混合预测模型流程图

EMD-ANN的预测框架结构如图3-2所示，其主要算法步骤如下：

Step1：对时间序列进行使用EMD算法进行分解，等到一系列的子序列IMF1，IMF2,……IMFm和一个余项RES。

Step2：对所有的子序列建立ANN模型，使用前个序列值预测第项的值，具体变量预测方式如图3-3所示。

Step3：对所有子序列的预测结果进行叠加等到最终预测值。

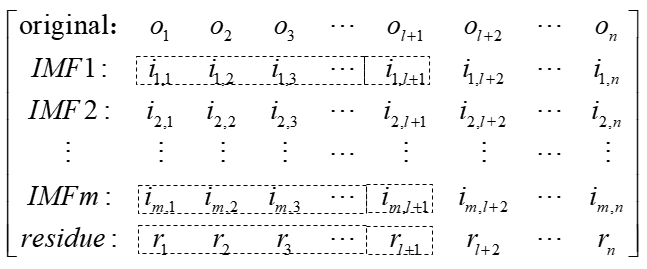


图3-3 EMD-ANN模型变量预测示意图

如图3-3所示绝大多数学者都是对每一个IMF和余项都进行建模，最后把每个模型得到的结果叠加作为最后的预测值，如图3-3虚线标识所示，对每一个子序列（IMF和residue）分别用前个序列值预测第项，并使用滑动窗口法预测余下各项。

3.2.2 基于EMD与ANN的单一模型预测结构

传统的基于EMD的预测模型框架都是对于每一个分解出来的IMF进行建模预测，然而对于每个IMF进行建模代价往往要面临极大的计算量这一问题。特别是在使用机器学习模型对每个IMF进行预测时往往需要很大的计算量，特别是现在使用支持向量机模型、人工神经网络等算法在构建模型的时候往往有一定量的超参数要进行试验调优，所以所组合的模型个数越多也就会有越多的超参数需要去确定。而且在对每个IMF结果进行叠加的时候往往会造成大量的累积误差，如果有个模型叠加作为最后结果就会有项的模型误差累积。再者有学者研究（H. Zhou(2019)[62]、张文风（2017）[59]）表明经过EMD分解的各项IMF往往具有一定的短期相关性，各IMF的特征之间具有一定的“交互作用[62]”，所以如果只是对各个子序列进行建模往往会忽略各子序列之间的相互影响。考虑到这些问题本文提出了一种基于EMD算法的单一模型预测框架并基于此框架建立S(single)-EMD-ANN模型。模型框架如图3-4所示。



图3-4 EMD与ANN混合的单一预测模型流程图

S-EMD-ANN的预测框架结构如图3-4所示，其主要算法步骤如下：

Step1：对时间序列进行使用EMD算法进行分解，等到一系列的子序列IMF1，IMF2,……IMFm和一个余项RES。

Step2：对所有的子序列建立一个ANN模型，用每一个子序列的前个序列值直接预测原始序列数据第项的值，具体变量预测方式如图3-5所示。

Step3：直接预测最终的结果值。

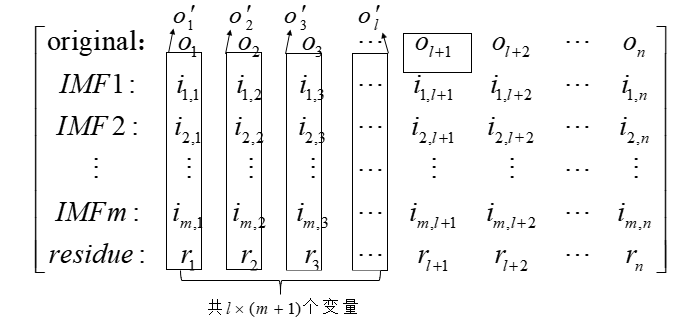


图3-5 S-EMD-ANN变量预测示意图

如图3-5所示S-EMD-ANN算法设计思想是把经过EMD分解出来的子序列看做是不同时域和频域的特征子序列，然后在每个子序列上选择一些滞后变量作为子序列的特征直接预测最后的结果。如图===所示把每个时间点的所有分解值作为的特征向量，最后用这些特征向量组成变量集，直接去预测最终值。用每个子序列前个序列值，共个的变量去预测第项，并使用滑动窗口法预测余下各项。

3.3 实证分析研究

3.3.1 数据介绍

本文实证研究主要选择上海证券综合指数简称“上证指数”和标准普尔500指数。上证指数选择从1990年12月19日（基准日）-2019年4月2日的数据共6916个时间点序列数据，为了数据的一般性对于标准普尔500指数我们也选择相同日期的数据即1990年12月19日-2019年4月2日的数据共7123个时间点序列数据。考虑到要与线性模型结果做对比，线性模型不适合做长期的预测；并且对于股票指数的预测研究在实际中都是经常根据新的数据来更新参数所以大部分研究都做的是短时间内的预测如：Zhou(2019)[62]对几个不同地区的股指预测选择250个左右的作为测试集，Carnelossi Furlaneto(2017)[70]选择250（大约一年的交易日数）各值最为测试集。对于本文中每一个数据集我们都选择留下最后300个数据点即一年多的股票指数作为测试数据。数据情况描述见表3-1。

表3-1 数据集描述

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 数据名称 | 训练数据 | 测试数据 | 均值 | 方差 | 数据日期范围 |
| 上证指数 | 6616 | 300 | 1924.18 | 1071.92 | 1990.12.19-2019.04.02 |
| 标普500 | 6823 | 300 | 1253.42 | 601.77 | 1990.12.19-2019.04.02 |

3.3.2 基于EMD的数据分解

本小节主要是依据数据差分变换规则，对数据进行预处理然后再进行分解。EMD分解算法最初提出是解决复杂的信号问题，信号序列一般都是高频、趋势性不强。然而如图3-6和图3-7虚线所示，本文处理的股票指数时间序列具有较强的趋势性并且频率较低。所以本文对数据进行一步差分预处理。文献[87,88]认为差分运算可以放大高频信息，显示高频信息引起的极值点，从而为提高频率分辨率提供了可能[88]。而且差分数据可以对数据进行缩放，在学习类算法中对数据缩放是加快学习速度的一个重要方法。

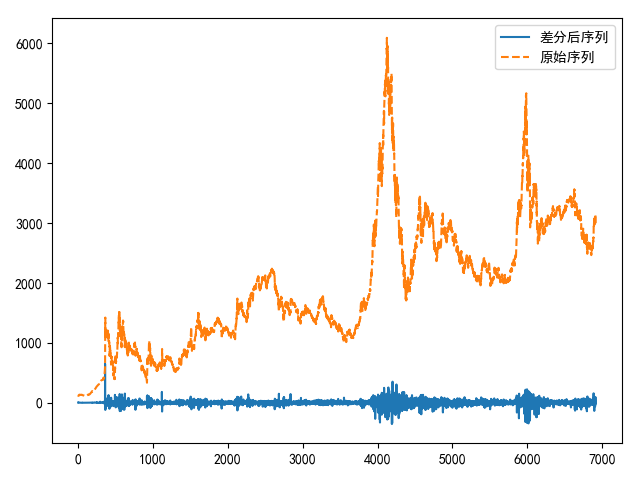


图3-6 上证指数原时间序列数据和差分后的时间序列数据

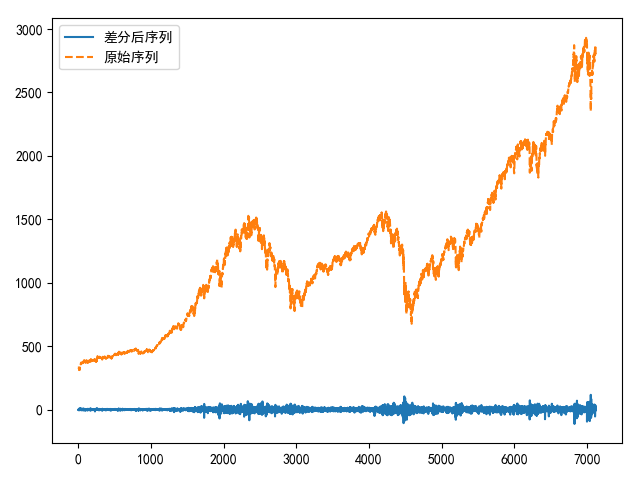


图3-7 标普500指数原时间序列和差分后的时间序列数据

此外，根据第二章EMD分解原理内容我们知道检验模态分解是有数据驱动的序列分解方式，再进一步讲经验模态分解是由序列的极值点驱动的序列分解方式，因为它主要是依据由极值点拟合的上下包络线分解的一种分解算法。所以我们可以知道当一个序列的极值点越多的话则上下包络线越接近真实的序列。所以表3-2中我们比较了股票指数的原序列和差分后序列的极值点个数，认为本文中的两个股票指数序列差分后会有更多的极值点可以使得上下包络线更加接近原始序列，也使得更多的点在包络线上。从图3-6和图3-7也可以看出差分后的序列的极值点比原序列极值点要多。增加极值点个数减弱端点效应的影响这一点对于第四章有重要的意义。

表3-2 数据集的极值点情况

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 数据名称 |  | 极大值点个数 | 极小值点个数 | 极值点个数 |
| 上证指数 | 原序列 | 1666 | 1666 | 3332 |
|  | 差分后序列 | 2265 | 2264 | 4529 |
| 标普500指数 | 原序列 | 1827 | 1828 | 3655 |
|  | 差分后序列 | 2380 | 2379 | 4759 |

对上证指数和标普500指数差分后的时间序列进行检验模态分解得如图3-8和图3-9的分解结果。上证指数分解得到12个IMF和一个余项，标普500指数分解得到11个IMF和一个余项。

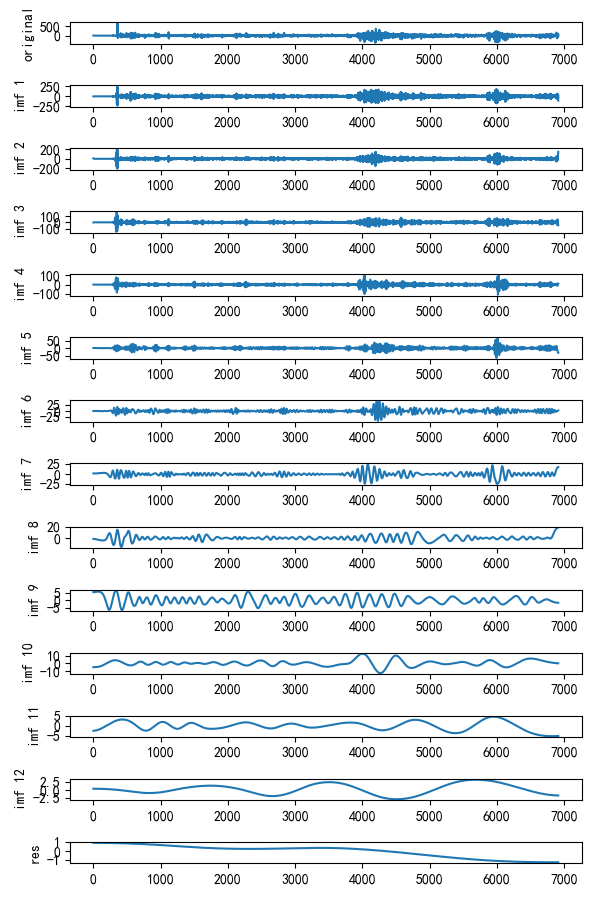


图3-8 上证指数差分后的序列分解结果

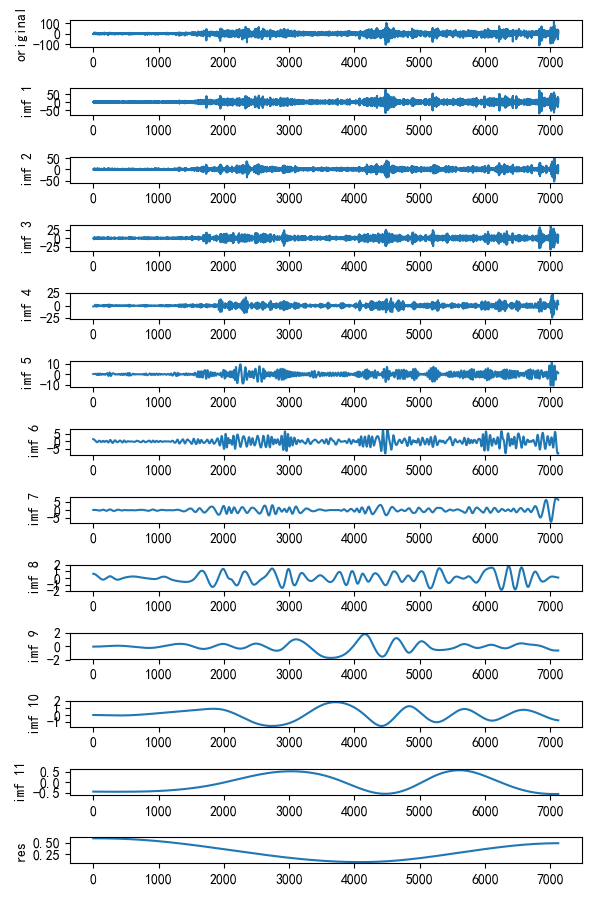


图3-9 标普500指数差分后的序列分解结果

3.3.3 模型性能评价指标

本章节实证结果分别从模型的总体预测精度以及模型预测结果之间差异性显著检验准则对预测效果进行综合评价。

3.3.3.1 整体预测准确度评估指标

本研究采用平均绝对误差(MAE)、平均绝对百分误差(MAPE)和均方根误差(RMSE)三个常用的总体误差评价指标。他们的计算公式如下：

 公式（3-3）

 公式（3-4）

 公式（3-5）

其中为观测值，为预测值，为样本数量。

为了定量评价不同模型的性能，定义指标RMSE、MAE和MAPE的改进百分比。

 公式（3-6）

其中为RMSE、MAE或MAPE，和分别是模型1和模型2的评价指标值。

3.3.3.2 显著性检验指标

此外，本文还采用Diebold Mariano (DM)检验[90,91]和Wilcoxon Signed rank (WS)检验来验证了两种模型预测精度的等价性，即表明两种模型的测试数据的预测值之间是否存在统计学上显著的差异。

首先介绍一下DM检验，DM检验是一种非参数检验方法。假设一对预测值的误差，。为测试样本数。模型预测结果的质量依据一些指定的损失函数来判断，常用的损失函数，，等。假如两个模型预测效果相等则两个模型的的差的期望为0。即：

 公式（3-7）

并令

 公式（3-8）

很自然的根据观察到的样本均值进行检验：

 公式（3-9）

其中为样本的观测值，为第个模型的第项样本的预测值。

假设向前步（本文为向前一步预测也就是提前一天预测）预测。

 公式（3-10）

 公式（3-11）

则得到DM（Diebold-Mariano）统计量：

 公式（3-12）

其中有公式（3-10）和公式（3-11）得到。在零假设下DM统计量服从渐进标准正态分布（证明见[91]）。即：

 公式（3-13）

学者Harvey, Leybourne 和 Newbold[92]给出了更严谨的修正统计量，即DM检验的修正也称为HLN检验。

 公式（3-14）

并得出HLN服从自由度为的分布（证明见[92]）。即

 公式（3-15）

现在常用的DM检验一般指DM检验的修正，即HLM检验。本文DM统计量的检验也是使用HLM检验值。

其次介绍一下WS检验。Wilcoxon signed-rank test（**威尔科克森符号秩检验**）也是一种非参数假设检验方法，它成对的检查2个数据集中的数据（即 paired difference test），判断 2个数据集是不是来自相同分布的总体。如果可以得到精确的有限样本检验，则可以根据零中值损失差的零假设检验检验观察到的损失差；即有的中位数为零。如果采样的损失差异具有对称的分布，则零假设与上述DM测试一致，表示相同的准确度[1]。WS统计检验值定义如下：

 公式（3-16）

 公式（3-17）

 公式（3-18）

其中如公式3-8所示，为样本数量

3.3.4 实证结果分析

实证研究数据依据3.3.1,3.3.2小节给出的数据，测试数据预测评价指标结果如表3-3和表3-4所示。对于模型自变量的选择，文献[62]分别选择了窗口大小为3,4,5的滞后变量进行建模分析。本文为了研究模型之间的差异分别选择了窗口大小为3-9的滞后变量（表格中Lag代表滞后变量的个数）。模型结构是一个输入层、一个隐藏层和一个输出层。输出层节点为1即向前一步预测，在ANN和EMD-ANN模型中网络隐藏层节点数我们选择经验值128，在S-EMD-ANN中由于其输入变量远多于EMD-ANN所以我们选择经验节点数256，并且对于所有的模型都在输入层和隐藏层加入Dropout层，并根据实验法得出当值为0.4模型表现更好（对比了0.3、0.4和0.5），输入层节点对于ANN模型和EMD-ANN模型来说为滞后变量（Lag）的个数，对于S-EMD-ANN模型来说输入节点个数为滞后变量与子序列的乘积（Lag\*m）。

表3-3 不同模型下上证指数的预测结果评价

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model |  | Lag | | | | | | | |
|  | 3 | 4 | 5 | 6 | 7 | 8 | 9 | Mean |
| ARIMA | MAPE | 3.120 | | | | | | | |
|  | MAE | 89.90 | | | | | | | |
|  | RMSE | 117.2 | | | | | | | |
| ANN | MAPE (%) | 1.381 | 1.377 | 1.378 | 1.377 | 1.378 | 1.381 | 1.378 | 1.379 |
|  | MAE | 39.58 | 39.45 | 39.49 | 39.44 | 39.49 | 39.57 | 39.47 | 39.50 |
|  | RMSE | 52.82 | 52.65 | 52.70 | 52.62 | 52.68 | 52.81 | 52.65 | 52.70 |
|  | （1,T） | 0.071 | 0.072 | 0.074 | 0.076 | 0.071 | 0.074 | 0.074 | 0.073 |
| E-A | MAPE (%) | 0.687 | 0.673 | 0.690 | 0.688 | 0.695 | 0.696 | 0.696 | 0.689 |
|  | MAE | 19.77 | 19.39 | 19.88 | 19.83 | 20.06 | 20.09 | 20.07 | 19.87 |
|  | RMSE | 26.04 | 25.95 | 26.45 | 26.30 | 26.44 | 26.53 | 26.42 | 26.30 |
|  | （13,T） | 0.943 | 0.932 | 0.963 | 1.002 | 0.968 | 0.956 | 0.954 | 0.96 |
| S-E-A | MAPE (%) | 0.673 | 0.684 | 0.683 | 0.678 | 0.676 | 0.671 | 0.673 | 0.677 |
|  | MAE | 19.39 | 19.72 | 19.68 | 19.53 | 19.50 | 19.40 | 19.46 | 19.53 |
|  | RMSE | 25.29 | 25.59 | 25.52 | 25.91 | 26.06 | 26.01 | 26.27 | 25.81 |
|  | （1,T） | 0.082 | 0.085 | 0.089 | 0.094 | 0.095 | 0.096 | 0.097 | 0.091 |

表3-4 不同模型下标普500指数的预测表现结果

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model |  | Lag | | | | | | | |
|  | 3 | 4 | 5 | 6 | 7 | 8 | 9 | Mean |
| ARIMA | MAPE (%) | 1.334 | | | | | | | |
|  | MAE | 35.98 | | | | | | | |
|  | RMSE | 48.96 | | | | | | | |
| ANN | MAPE (%) | 1.039 | 1.036 | 1.039 | 1.038 | 1.035 | 1.038 | 1.037 | 1.038 |
|  | MAE | 28.06 | 27.98 | 28.05 | 28.02 | 27.96 | 28.04 | 28.02 | 28.02 |
|  | RMSE | 39.86 | 39.75 | 39.81 | 39.79 | 39.68 | 39.78 | 39.76 | 39.78 |
|  | （1,T） | 0.072 | 0.073 | 0.072 | 0.072 | 0.072 | 0.072 | 0.072 | 0.072 |
| E-A | MAPE (%) | 0.519 | 0. 528 | 0.529 | 0.525 | 0.527 | 0.528 | 0.530 | 0.527 |
|  | MAE | 14.02 | 14.27 | 14.28 | 14.19 | 14.24 | 14.27 | 14.32 | 14.23 |
|  | RMSE | 19.38 | 19.88 | 19.81 | 19.79 | 19.81 | 19.88 | 19.89 | 19.78 |
|  | （12,T） | 0.868 | 0.869 | 0.873 | 0.868 | 0.869 | 0.870 | 0.876 | 0.879 |
| S-E-A | MAPE (%) | 0.511 | 0.514 | 0.516 | 0.521 | 0.523 | 0.518 | 0.524 | 0.518 |
|  | MAE | 13.84 | 13.93 | 13.96 | 14.11 | 14.16 | 14.04 | 14.19 | 14.03 |
|  | RMSE | 19.21 | 19.10 | 19.12 | 19.28 | 19.45 | 19.40 | 19.64 | 19.31 |
|  | （1,T） | 0.083 | 0.086 | 0.087 | 0.089 | 0.091 | 0.094 | 0.095 | 0.089 |

E-A表示EMD-ANN，S-E-A表示S-EMD-ANN，（N，T）表示模型个数和训练模型所需时间（分钟）。

表3-3和表3-4中的数据说明：ARIMA模型的平稳性和白噪声检验结果以及对应的自回归项数和移动平均项数见附录1，其中、值的确定依据AIC信息准则选取。ANN、EMD-ANN和S-EMD-ANN中每个窗口下（自变量Lag的值）对应的MAPE、MAE、RMSE、T评估指标值为20次实验的平均值。Mean列表示的是所有选取的窗口下的模型表现评估指标的平均值即140（20\*7）次实验的预测结果评估指标的均值。

从表3-3和表3-4我们可以得出以下结论：

结论1：对于上证指数和标普500指数时间序列数据集来说，ANN、EMD-ANN、S-EMD-ANN模型的预测结果要好于ARIMA模型，表现评估指标以MAPE为例，在上证指数数据集上分别改进55.8%（(3.12-1.379)/3.12）、77.9%、78.3%，在标普500指数数据集上分别改进22.2%、60.5%，61.2%。所以可以知道非线性模型ANN、EMD-ANN、S-EMD-ANN模型优于线性模型ARIMA模型。

结论2：正如大多数文献[=====]研究的结论一样，在本文的数据集上EMD-ANN模型预测结果优于单独的ANN模型，表现评估指标以MAPE为例，在上证指数数据集上改进50.0%（(1.379-0.689)/1.379），在标普500指数数据集上改进49.2%。本文所提出的S-EMD-ANN也同样优于单独的ANN模型，表现评估指标以MAPE为例，在上证指数数据集上改进50.9%（(1.379-0.689)/1.379），在标普500指数数据集上改进50.1%，同时我们也发现单从结果来看本文所提出的EMD-ANN的改进模型S-EMD-ANN的表现结果略优于EMD-ANN的表现结果。

结论3：本章节主要的研究内容，从图3-10和图3-11我们可以看出除了在上证指数数据上窗口（Lag=4）等于4时S-EMD-ANN模型的MAPE值大于EMD-ANN模型的MAPE值，在其他选取的窗口下S-EMD-ANN模型的MAPE值都要小于EMD-ANN模型的MAPE值，从附录2和附录3中的MAE和RMSE指标值也可以得出类似的结果。从整体预测结果的MAPE评价指标上看S-EMD-ANN比EMD-ANN在上证指数和标普500指数分别改进了1.74%（(0.689-0.677)/0.689），1.71%，在MAE评价指标上分别改进了1.71%，1.41%，在RMSE评价指标数上分别改进了1.86%、2.38%。虽然本文提出的改进模型在精度上提升比较小，但是在效率上却有很大的提高，从表3-3和表3-4中我们看出在上证指数数据上本文所提出的S-EMD-ANN只需要1个模型而一般的EMD-ANN模型需要13（与EMD分解得到的子序列数相等）模型,而且模型训练时间减少了90.5%，在标普500指数数据上S-EMD-ANN只需要1个模型而一般的EMD-ANN模型需要12个模型，模型训练时间减少了89.9%，单个模型还有一点优势在于像ANN等一些学习类模型往往需要大量的超参数需要要去搜索确定，模型数量的减少也大大的提高了超参数择优的效率。为了更进一步说明两个模型结果的差异，我们对两个模型的预测结果进行假设检验，利用DM检验和WX检验检验两个模型的结果差异性是否显著。假设检验结果表3-5和表3-6所示，从表中数据可以看出无论对于绝对百分比误差(APE)、绝对误差(AE)、还是方差(SE)，S-EMD-ANN和EMD-ANN两种模型在任何窗口下预测结果都没有显著的差异，假设检验的值都远大于0.05，所以我们可以认为S-EMD-ANN模型的预测结果和EMD-ANN模型的预测结果没有显著性的差异。从前面的表述也可以看出S-EMD-ANN模型的预测结果相较于EMD-ANN模型的预测结果在三种评估指标上都仅仅改进了1.5%左右。但从模型的训练时间和模型数量上看S-EMD-ANN模型远优于EMD-ANN模型。

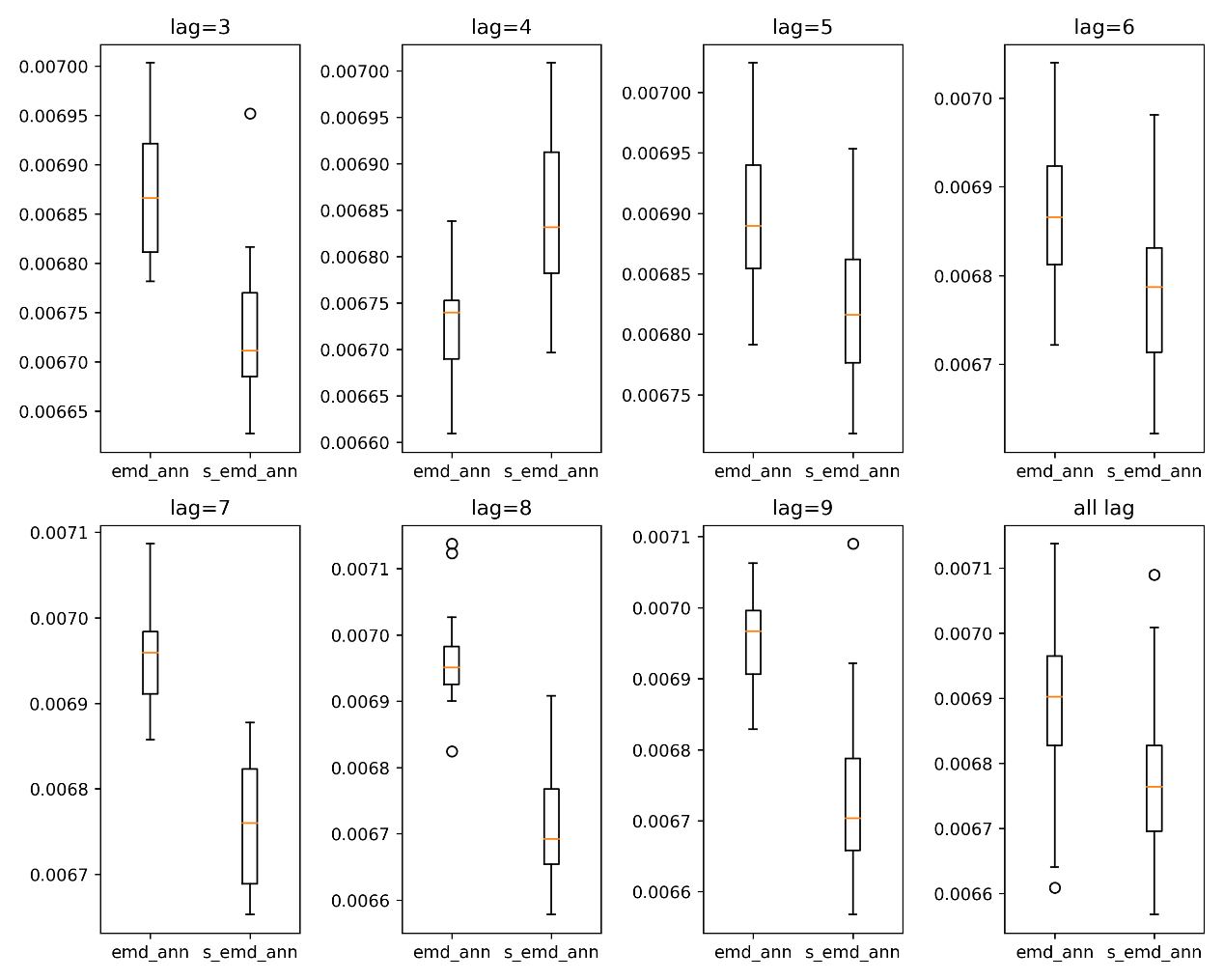


图3-10 上证指数数据中EMD-ANN与S-EMD-ANN模型的MAPE指标值

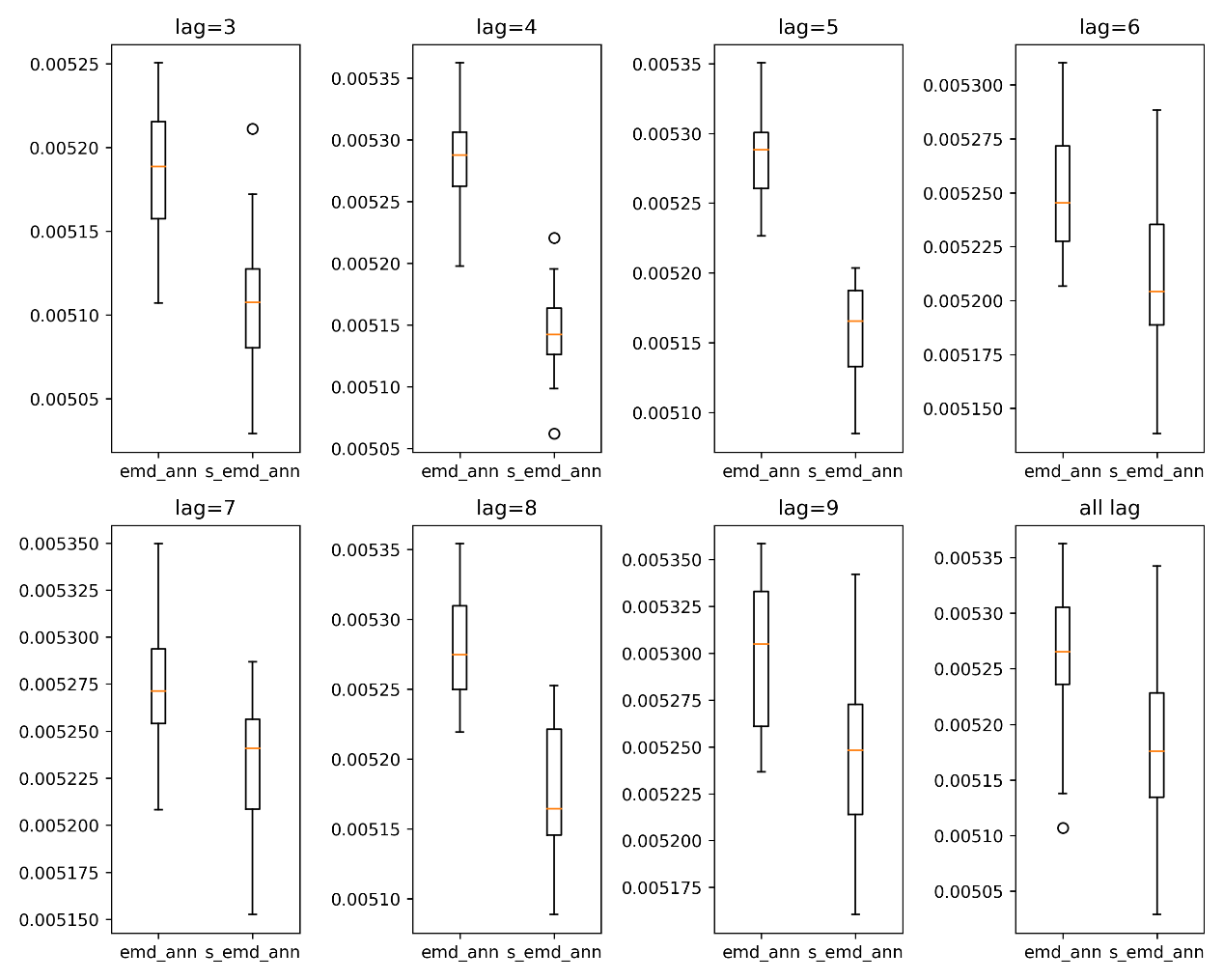


图3-11 标普500指数数据中EMD-ANN与S-EMD-ANN模型的MAPE指标值

表3-5 上证指数模型差异性检验DM与WX检验p值

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| S-E-A vs E-A | | Lag | | | | | | | |
| 3 | 4 | 5 | 6 | 7 | 8 | 9 | Mean |
| APE | DM | 0.377 | 0.713 | 0.633 | 0.480 | 0.186 | 0.099 | 0.103 | 0.282 |
|  | WX | 0.554 | 0.390 | 0.708 | 0.824 | 0.344 | 0.166 | 0.349 | 0.511 |
| AE | DM | 0.416 | 0.705 | 0.626 | 0.469 | 0.191 | 0.111 | 0.118 | 0.296 |
|  | WX | 0.623 | 0.394 | 0.707 | 0.846 | 0.338 | 0.180 | 0.358 | 0.552 |
| SE | DM | 0.225 | 0.560 | 0.189 | 0.507 | 0.426 | 0.328 | 0.520 | 0.299 |
|  | WX | 0.421 | 0.520 | 0.885 | 0.964 | 0.526 | 0.305 | 0.761 | 0.403 |

表3-6 标普500指数模型差异性检验DM与WX检验p值

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| S-E-A vs E-A | | Lag | | | | | | | |
| 3 | 4 | 5 | 6 | 7 | 8 | 9 | Mean |
| APE | DM | 0.435 | 0.247 | 0.244 | 0.630 | 0.527 | 0.307 | 0.435 | 0.281 |
|  | WX | 0.943 | 0.804 | 0.819 | 0.706 | 0.893 | 0.688 | 0.727 | 0.835 |
| AE | DM | 0.476 | 0.281 | 0.267 | 0.674 | 0.580 | 0.339 | 0.477 | 0.313 |
|  | WX | 0.940 | 0.790 | 0.819 | 0.696 | 0.885 | 0.699 | 0.732 | 0.837 |
| SE | DM | 0.626 | 0.135 | 0.161 | 0.274 | 0.343 | 0.275 | 0.482 | 0.220 |
|  | WX | 0.791 | 0.970 | 0.993 | 0.538 | 0.704 | 0.976 | 0.995 | 0.986 |

3.3.5 同类研究对比

为了进一步研究本文所提出模型的预测能力，本小节选取了一些文献并利用文献中的数据使用本文的预测模型，并与文献结果进行对比研究。本章模型参数同3.3.4小节相同，模型结果为20次实验的均值。

表3-7 上证指数同类模型研究结果

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 上证指数 | Zhou (2019) [62]EMD-FNN | | | 本文模型结果（S-EMD-ANN） | | |
| Lag | MAPE | MAE | RMSE | MAPE | MAE | RMSE |
| 3 | 0.0123 | 37.1778 | 61.5138 | 0.0074 | 21.83 | 31.45 |
| 4 | 0.0143 | 43.5513 | 66.7576 | 0.0076 | 22.42 | 32.11 |
| 5 | 0.0152 | 54.8499 | 89.9065 | 0.0075 | 22.35 | 32.74 |

表3-8 标普500指数同类模型研究结果

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 标普500 | Zhou (2019) [62]EMD-FNN | | | 本文模型结果（S-EMD-ANN） | | |
| Lag | MAPE | MAE | RMSE | MAPE | MAE | RMSE |
| 3 | 0.0105 | 13.0396 | 17.6591 | 0.00812 | 10.09 | 13.67 |
| 4 | 0.0122 | 15.1386 | 20.3978 | 0.00831 | 10.33 | 13.91 |
| 5 | 0.0130 | 16.1700 | 22.1277 | 0.00843 | 10.47 | 14.09 |

表3-7和表3-8中虚线部分，为Zhou(2019)[62]的研究结果其中表3-7为上证指数研究结果，数据为2012年1月4日到2016年12月30日，共1214个数据点，表中数据为最后242个数据点的预测结果评价指标值。从表中的结果评价指标来看本文所提出的模型的预测效果要远远优于文献中的预测结果，从窗口为3（文献中选择的最优窗口）时的结果来看MAPE指标改进了39.84%，MAE指标改进了41.28%，RMSE指标改进了48.87%。表3-8为标普500指数研究结果，数据为2007年1月3日到2011年12月30日，共1260个数据点，表3-7和表3-8中数据为最后252个数据点的预测结果评价指标值。从表中的结果评价指标来看本文所提出的模型的预测效果要远远优于文献中的预测结果，从窗口为3（文献[62]中选择的最优窗口）时的结果来看MAPE指标改进了22.67%，MAE指标改进了22.62%，RMSE指标改进了22.59%。同时从本文模型结果中我们也可以进一步证明文献[62]所选的窗口大小时，确实是当窗口等于3时模型的表现结果较好。

表3-9 沪深300股指期货指数同类研究结果

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 预测模型 | 史(2015)[86] | | | | 本文模型 |
| GARCH | ARIMA | RBF(ANN) | EMD-RBF(ANN) | S-EMD-ANN |
| MAE | 22.8945 | 20.1335 | 18.6743 | 15.8276 | 11.09 |
| RMSE | 30.2314 | 25.4589 | 24.7865 | 21.3992 | 13.75 |

表3-9中虚线部分为史(2015)[86]的研究结果，其研究的是对沪深300股指期货的预测，其研究结果如表虚线所示，数据为2012年5月22日到2014年9月9日，表中数据为后61个数据点的预测结果评价指标。文献[86]选择的窗口大小也为3但是没给出理由，为了进行对比研究表中本文的模型也选择窗口大小为3时的结果。从表中的结果评价指标来看本文所提出的模型的预测效果要优于文献中的预测结果，其中MAE指标改进了29.93%，RMSE指标改进了35.74%。

3.4 本章小结

本章主要是提出了一种改进的EMD-ANN预测模型S-EMD-ANN预测模型，通过大量的实证研究并结合对比其它文献研究的研究结果我们可以得到以下结论。1、本文所提出的S-EMD-ANN模型在模型的数量以及复杂程度是上都远小于EMD-ANN模型，并且S-EMD-ANN模型在训练是需要的时间远小于EMD-ANN训练需要的时间，所以S-EMD-ANN的效率更高。2、从两种模型的结果差异性检验上我们可以看出两种预测模型的结果没有显著性差异。但是在各个评价指标上我们可以看出S-EMD-ANN预测模型的结果要优于EMD-ANN预测模型的结果。所以我们可以得出结论S-EMD-ANN预测模型的预测结果要比EMD-ANN预测结果要好，但结果改进不是很大，从3.3.4章节我们知道在各个指标上的改进大约在1.5%左右。3、在3.3.5章节中在与其它学者的研究成果进行对比时，发现S-EMD-ANN模型的预测结果比其它学者的研究结果改进了20%-50%之间，说明本文所提出的模型在其他数据上也能有良好的表现结果。

4 基于经验模态分解与神经网络的自适应预测模型

4.1 EMD算法中的前视偏差分析

前视误差（Look-ahead bias）也有称为前瞻性偏差可以定义为对直到稍后的日期才可用信息的无意使用，换句话说，就是无意中用未来的数据预测未来。Mahfoud,Mani(1996)[93]指出，商业和公开的金融数据可能从一开始就含有前瞻性偏见。例如，与政府经济指标相关的数据在通过审查过程，可能因为未来的某项指标而修改过去的数据[70]。

除了人为的向数据本身添加的前视偏差之外，使用一些技术手段对数据进行预处理时也有可能引入前视偏差。例如，数据标准化技术是非常流行的数据预处理方式也是处理时间序列的常用方式。Min-max和z-score标准化，可以说是两种最流行的标准化技术，利用一些数据的统计信息对数据进行变换处理，然而对于一个数据集来说最大值，最小值以及标准差都是整个数据的统计信息，特别是大部分研究在计算这些统计信息时是用了个数据集来计算的，那么对于我们要用来测试的结果来说，我们在预测之前就已经知道了它的最大值、最小值或者标准差的情况，这显然是引入了前视偏差，该偏差可能对其它具有一定统计特征的数据影响不大，但在对一些复杂的事件序列来说可能影响就会偏大，特别是一些具有一定周期性和波动性的数据。在经验模态分解中也是存在这样的前视性误差，而且这种前视误差影响比较大。不幸的是，在文献的研究工作中，这种性质的偏差经常被忽视，而且由于EMD算法的使用包含有前视偏差，在文献中包含前视偏差测试结果似乎很普遍[70]。这也使得诸多学者的研究结果都看起来很“漂亮”。

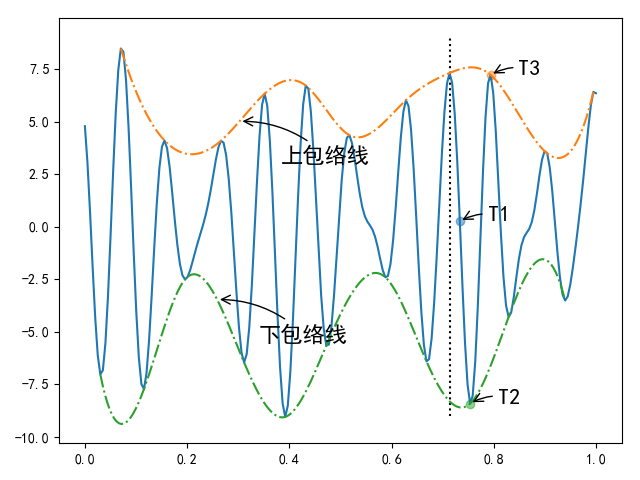


图4-1 EMD分解示意图

正如N. Sun (2018)[65] 、Q. Tan(2018)[66] 和X. Zhang(2015)[69]提出的问题，现有的大部分文献[57-64]对于数据集划分都存在一定的瑕疵，现有的文献在处理数据是都会忽略掉“前视性[70]”误差。如图4-1所示，现在有大部分文献在划分数据是都是将整个时间序列进行经验模态分解，然后将分解后的序列划分为训练集和测试集。然而这样的序列分解方式考虑了未发生时间点的信息，所以不是真实的预测。如图4-1所示对T1点的值分解使用了极值点T2和T3的值，而T2和T3对于T1来说属于未来的信息。所以如果我们要预测T2时刻的值那么对序列分解时只能用T2时刻之前的历史值。在文献[66]中学者Q. Tan将涉及未来信息的实证实验称为后视实验（Hindcast experiment），把不涉及未来信息的预测称为预测实验（Forecast experiment）。在基于EMD的一系列混合模型的研究中后视实验是一种及其理想化的实验，是一种不考虑未来极值点对结果影响下的实验，然而绝大多数的研究者都停留在理想化的研究中，显然本文的第3章的实验也是理想化的实验也包含前视性误差。

4.2 EMD与ANN混合的自适应预测模型

4.2.1 基于EMD与ANN的自适应预测模型

为了解决基于EMD的混合模型的前视偏差问题，N. Sun (2018)[65] 、Q. Tan(2018)[66]、Dennis(2017)[70]和X. Zhang(2015)[69]学者都使用了一种自适应方法来构建基于EMD的一系类预测模型。其基本预测流程如图4-2所示。并用AEMD-ANN表示该模型。



图4-2 AEMD-ANN预测模型流程图

图4-2所示AEMD-ANN算法的核心思想是：对于测试数据来说当预测T+1时刻的值时，在分解时只使用T+1时刻之前的数据，避免使用到未来的极值信息；然后再把T+1时刻的值添加到训练数据中再重新分解数据，重新训练模型然后再预测T+2时刻的值，依次类推直到所有的测试数据被预测出来。自适应的意思就是每增加一个值，都要重新训练模型的参数，使得模型能够学习到最新数据的特征。其算法流程如下：

Step1：把原始数据序列分解为训练集和测试集 ，并初始化k=1；

Step2：使用EMD算法把原序列分解成n个IMF和一个余项；

Step3：对每一子序列建立ANN模型，用子序列的前项的值，预测第项的值，具体变量预测方式如图3-3所示。

Step4：每个模型向前预测一步，预测结果求和得到测试数据集的第k项的值的预测值；

Step5：如果k=m，则输出所有测试数据的预测值，否则，将加入得到新的训练数据，并令k=k+1重复Step2-Step5。

4.2.2 基于EMD与ANN自适应预测模型的改进

根据3.2章节的分析我们知道传统的EMD与ANN得混合模型，存在模型结构复杂，模型数量大，模型训练效率低等问题，这些问题在自适应模型中同样存在并且更加严重。因为对于测试数据的每个数据点都要重新分解数据和训练模型，所以可想而知传统的自适应混合模型的计算量是非常庞大的，在后面的实证结果中这个问题也充分的显现了出来。基于这些问题我们同样给出EMD与ANN混合的单一模型预测框架，如图4-3所示。并用S-AEMD-ANN表示该模型。



图4-3 S-AEMD-ANN预测模型流程图

S-AEMD-ANN的算法步骤和AEMD-ANN只有下面两步的差异：

Step3：对所有的子序列建立一个ANN模型，用每一个子序列的前项序列值直接预测原始序列数据的第项的值，具体变量预测方式如图3-5所示。

Step4:向前预测一步得到测试数据集的第项的值的预测值。

根据4.3的结论3我们知道，如果只是简单地使用自适应EMD分解去改善ANN的预测结果，然而得到的结果非常不理想。根据检验模态分解算法的原理我们可以知道以下因素使得预测结果不理想的原因之一：对于训练数据来说每一个时刻的数据点，用来预测时刻数据的自变量（时刻的数据）都是真实的，也就是说这些自变量都是根据真实的数据分解出来的数据，是存在前视性误差的。而对于测试数据来说当我们在测试数据集中预测t时刻的值时，我们用来分解时间序列的时间点只能使用到t-1时刻，那么可想而知我们不知道t-1时刻的值是不是极值点也不知道下一个极值点在哪一时刻，所以只能依据延拓等手段来预估未来的极值点已完成数据的分解，那么在这种情况下就会得到一些虚假的分解值，而且时间点越靠近t时刻的数据失真越严重，当极值点的个数越少，失真数据点就越多（这也是3.3.2小结对数据差分来增加极值点的一个重要原因）。因此我们使用真实分解数据训练得到的参数在应用于测试数据的失真分解数据时有可能会给测试的预测结果带来巨大的误差。由于测试数据存在虚假的分解数据，因此本文对AEMD-ANN和S-AEMD-ANN模型进行简单的改进，算法改进如下：

AEMD-ANN算法的改进，并将改进后的模型记为AEMD-ANNa。

Step3：对每一子序列建立ANN模型，用(删除项距离预测值较近的滞后变量)项序列值，预测第项的值，具体变量预测方式如图4-4所示。

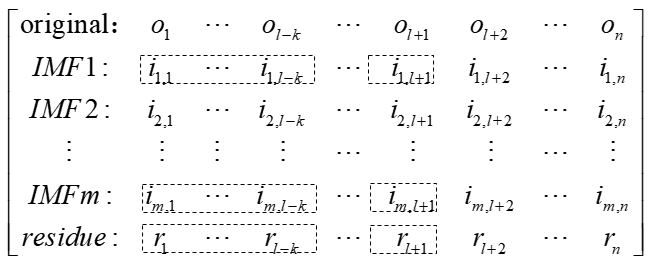


图4-4 AEMD-ANNa模型变量预测示意图

S-AEMD-ANN算法的改进，并将改进后的模型记为S-AEMD-ANNa。

Step3：对所有的子序列建立一个ANN模型，用(删除项距离预测值较近的滞后变量)项序列值直接预测第项的值，具体变量预测方式如图4-5所示。

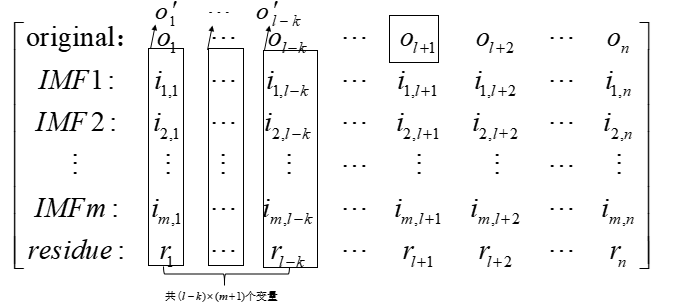


图4-5 S-AEMD-ANNa模型变量预测示意图

改进的AEMD-ANNa和S-AEMD-ANNa模型主要依据以下两种假设：

第一，对于一些失真的数据点我们选择对其进行直接删除处理，尽量选择失真较小的数据点作为自变量。从而减少失真数据点对预测结果的影响。

第二，对于所要预测的数据点来说，我们假设它与多个滞后数据点相关，在一定程度上可以舍弃一些可能严重失真的滞后变量，来减少失真数据对预测结果的影响。我们从表4-1可以看出这种假设是完全合理的。从表4-1中我们看出对于上证指数原数据和标普500指数原数据来说T时刻的值与滞后9个变量的值相关系数相差不大，都与T时刻的值具有很大的相关性。对于本文使用的差分后的序列来说相关性也不仅仅是依据距离T时刻远近来决定的，例如在上证指数中T时刻与T-1、T-2时刻的相关系数分别为0.036、0.037，而与T-3、T-4时刻数据的相关系数分别为0.045、0.072。所以在一定程度上牺牲一些失真的数据点来换取预测结果的准确性是可取的，而且从4.3的结论4我们也可以看出，这种做法对于优化模型来说是可取的。

表4-1 数据自相关系数

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 数据 | T | T-1 | T-2 | T-3 | T-4 | T-5 | T-6 | T-7 | T-8 | T-9 |
| SZ-A | 1 | 0.999 | 0.999 | 0.998 | 0.997 | 0.996 | 0.995 | 0.994 | 0.993 | 0.993 |
| SZ-B | 1 | 0.036 | 0.037 | 0.045 | 0.072 | 0.004 | 0.052 | 0.021 | 0.007 | 0.009 |
| SP-A | 1 | 0.999 | 0.999 | 0.999 | 0.999 | 0.999 | 0.999 | 0.998 | 0.998 | 0.998 |
| SP-B | 1 | 0.048 | 0.037 | 0.007 | 0.017 | 0.038 | 0.004 | 0.008 | 0.017 | 0.014 |

注：SZ-A表示上证指数序列的原序列，SZ-A表示上证指数序列差分后的序列。SP-A表示标普500指数序列的原序列，SP-B表示标普500指数差分后的序列。

4.3 实证分析

实证研究数据依据3.3.1,3.3.2小节给出的数据，测试数据预测表现结果如表4-2和表4-3所示。对于模型自变量的选择，本文为了研究模型之间的差异分别选择了窗口大小为5-11的滞后变量（表中Lag=5,…,11）。模型结构与第三章相同。

表4-2 不同自适应模型下上证指数的预测表现结果

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model |  | Lag | | | | | | | |
|  | 5 | 6 | 7 | 8 | 9 | 10 | 11 | Mean |
| ARIMA | MAPE | 3.323 | | | | | | | |
|  | MAE | 96.99 | | | | | | | |
|  | RMSE | 130.81 | | | | | | | |
| ANN | MAPE (%) | 1.529 | 1.540 | 1.535 | 1.532 | 1.534 | 1.525 | 1.537 | 1.533 |
|  | MAE | 44.48 | 44.79 | 44.64 | 44.57 | 44.62 | 44.35 | 44.70 | 44.59 |
|  | RMSE | 61.97 | 63.36 | 62.13 | 62.07 | 62.15 | 61.81 | 62.24 | 62.10 |
|  | （1,T） | 0.072 | 0.072 | 0.072 | 0.072 | 0.072 | 0.078 | 0.075 | 0.073 |
| AE-A | MAPE (%) | 1.759 | 1.740 | 1.737 | 1.728 | 1.727 | 1.713 | 1.724 | 1.733 |
|  | MAE | 51.37 | 50.82 | 50.77 | 50.49 | 50.45 | 50.04 | 50.37 | 50.62 |
|  | RMSE | 70.98 | 70.11 | 70.39 | 69.80 | 69.42 | 69.33 | 69.83 | 69.98 |
|  | （13\*,T） | 51.45 | 52.10 | 51.56 | 51.68 | 52.00 | 51.71 | 52.16 | 51.81 |
| S-AE-A | MAPE (%) | 1.664 | 1.662 | 1.664 | 1.666 | 1.663 | 1.669 | 1.670 | 1.665 |
|  | MAE | 48.55 | 48.48 | 48.56 | 48.61 | 48.48 | 48.69 | 48.70 | 48.58 |
|  | RMSE | 65.17 | 65.22 | 65.30 | 65.61 | 65.20 | 65.54 | 65.05 | 65.30 |
|  | （1,T） | 5.601 | 5.797 | 5.617 | 5.824 | 5.619 | 5.599 | 5.787 | 5.692 |
| AE-Aa | MAPE (%) | 1.262 | 1.254 | 1.305 | 1.316 | 1.323 | 1.323 | 1.337 | 1.303 |
|  | MAE | 36.67 | 36.41 | 37.87 | 38.21 | 38.47 | 38.44 | 38.87 | 37.85 |
|  | RMSE | 49.96 | 52.06 | 53.67 | 54.23 | 54.19 | 54.67 | 55.43 | 53.46 |
|  | （13\*,T） | 49.81 | 47.05 | 48.61 | 48.24 | 48.24 | 48.48 | 47.65 | 48.33 |
| S-AE-Aa | MAPE (%) | 1.181 | 1.172 | 1.173 | 1.180 | 1.174 | 1.181 | 1.181 | 1.177 |
|  | MAE | 34.26 | 33.97 | 34.02 | 34.23 | 34.06 | 34.26 | 34.26 | 34.15 |
|  | RMSE | 49.03 | 48.72 | 48.96 | 48.87 | 48.74 | 49.21 | 49.10 | 48.94 |
|  | （1,T） | 5.344 | 4.983 | 5.057 | 5.136 | 5.097 | 5.104 | 5.106 | 5.118 |

表4-3 不同自适应模型下标普500指数的预测表现结果

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model |  | Lag | | | | | | | |
|  | 5 | 6 | 7 | 8 | 9 | 10 | 11 | Mean |
| ARIMA | MAPE | 0.936 | | | | | | | |
|  | MAE | 25.91 | | | | | | | |
|  | RMSE | 32.31 | | | | | | | |
| ANN | MAPE (%) | 0.822 | 0.819 | 0.823 | 0.820 | 0.820 | 0.819 | 0.819 | 0.820 |
|  | MAE | 22.64 | 22.57 | 22.66 | 22.60 | 22.61 | 22.57 | 22.58 | 22.60 |
|  | RMSE | 28.47 | 28.39 | 28.50 | 28.41 | 28.42 | 28.37 | 28.40 | 28.42 |
|  | （1,T） | 0.071 | 0.067 | 0.067 | 0.069 | 0.069 | 0.068 | 0.069 | 0.069 |
| AE-A | MAPE (%) | 1.056 | 1.064 | 1.069 | 1.075 | 1.078 | 1.081 | 1.077 | 1.071 |
|  | MAE | 28.91 | 29.12 | 29.26 | 29.39 | 29.47 | 29.55 | 29.44 | 29.31 |
|  | RMSE | 37.02 | 37.34 | 37.53 | 37.78 | 38.19 | 38.05 | 38.11 | 37.72 |
|  | （12\*,T） | 41.81 | 41.85 | 42.04 | 42.54 | 42.17 | 42.28 | 42.83 | 42.22 |
| S-AE-A | MAPE (%) | 0.960 | 0.952 | 0.954 | 0.941 | 0.959 | 0.953 | 0.954 | 0.953 |
|  | MAE | 26.36 | 26.13 | 26.19 | 25.84 | 26.33 | 26.16 | 26.21 | 26.17 |
|  | RMSE | 33.51 | 33.13 | 33.16 | 32.91 | 33.33 | 33.12 | 33.24 | 33.20 |
|  | （1,T） | 4.332 | 4.308 | 4.319 | 4.445 | 4.316 | 4.323 | 4.356 | 4.343 |
| AE-Aa | MAPE (%) | 0.874 | 0.859 | 0.898 | 0.913 | 0.924 | 0.932 | 0.938 | 0.905 |
|  | MAE | 23.97 | 23.60 | 24.63 | 25.03 | 25.34 | 25.55 | 25.73 | 24.84 |
|  | RMSE | 32.07 | 31.60 | 33.63 | 34.28 | 34.74 | 35.02 | 35.13 | 33.78 |
|  | （12\*,T） | 42.21 | 42.50 | 42.38 | 42.65 | 42.53 | 42.72 | 44.06 | 42.72 |
| S-AE-Aa | MAPE (%) | 0.717 | 0.720 | 0.713 | 0.716 | 0.720 | 0.719 | 0.717 | 0.717 |
|  | MAE | 19.77 | 19.84 | 19.65 | 19.74 | 19.83 | 19.82 | 19.76 | 19.77 |
|  | RMSE | 24.69 | 24.75 | 24.54 | 24.63 | 24.71 | 24.77 | 24.67 | 24.68 |
|  | （1,T） | 4.278 | 4.164 | 4.159 | 4.178 | 4.164 | 4.169 | 4.183 | 4.185 |

表4-2和表4-3中的数据说明：ARIMA模型的平稳性和白噪声检验结果以及对应的自回归项数和移动平均项数见附录==，其中、值的确定依据AIC信息准则选取。ANN、AEMD-ANN和S-AEMD-ANN中每个窗口下（自变量Lag的值）对应的MAPE、MAE、RMSE、T评估指标值为20次实验的平均值。Mean列表示的是所有选取的窗口下的模型表现评估指标的平均值即140（20\*7）次实验的预测结果评估指标的均值。由于计算量的限制，本章节的测试数据选择了大约两个月的交易数据共50个数据点作为测试数据，也即是留下最后50个数据点为测试数据集。

为了更直观的表现数据结果，以箱线图来呈现研究结果：

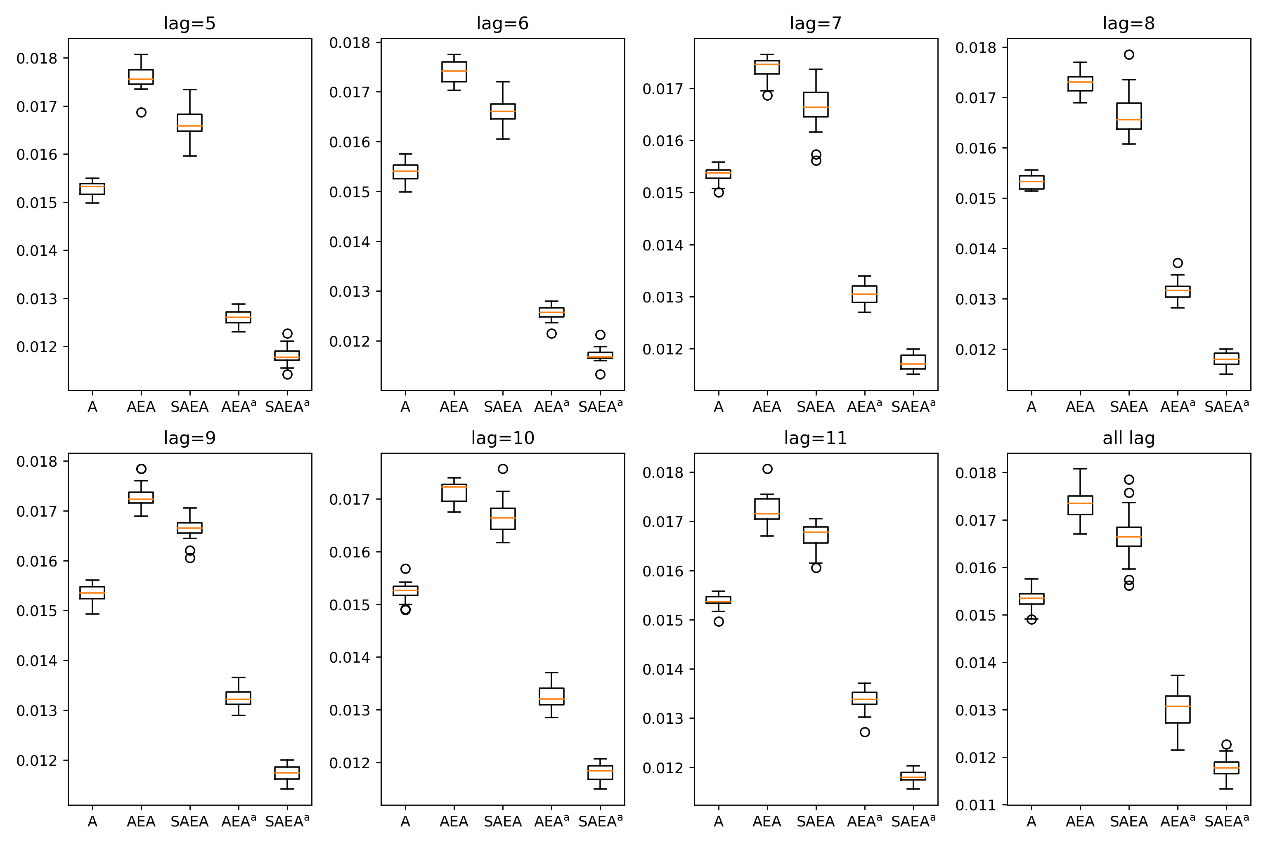


图4-6 上证指数数据中剔除后视误差后模型的MAPE指标值

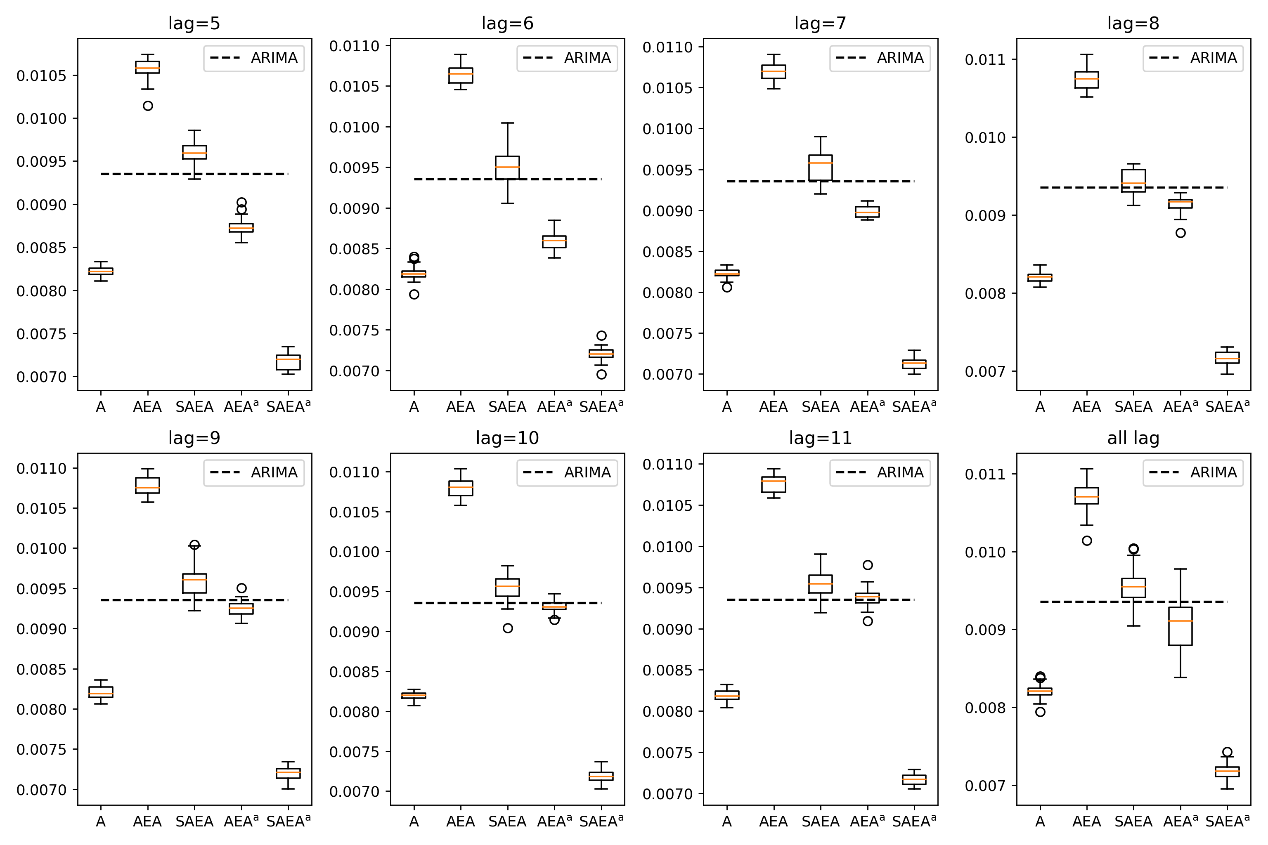


图4-7 标普500指数数据中剔除后视误差后模型的MAPE指标值

从表4-2和表4-3我们可以得出以下结论：

结论1 从表4-2和图图4-6我们可以看出在上证指数数据集上ANN，AEMD-ANN，S-AEMD-ANN，AEMD-ANNa和S-AEMD-ANNa模型的结果都优于ARIMA模型的结果，然而从表4-3和图4-7中我们可以看出在标普500数据集上只有ANN模型和S-AEMD-ANNa的结果是优于ARIMA模型的结果。所以我们认为只有ANN和S-AEMD-ANNa模型是优于线性模型ARIMA模型的，而其他几种模型在剔除前视性偏差之后模型表现性能并不一定比线性模型ARIMA模型好。

结论2表4-2，表4-3和图4-6，图4-7的AEMD-ANN和S-AEMD-ANN以及AEMD-ANNa和S-AEMD-ANNa模型的对比结果再次验证了3.3.4章节的结论3的正确性，即本文所提出的单个混合预测模型要优于传统的混合预测模型。以MAPE评价指标为例，S-AEMD-ANN模型比AEMD-ANN在上证指数数据集上平均改进了3.92%（(1.733-1.665)/1.733），在标普500数据集上平均改进了11.02%，S-AEMD-ANNa模型比AEMD-ANNa在上证指数数据集上平均改进了9.67%（(1.303-1.177)/1.303），在标普500数据集上平均改进了20.77%。在模型数量上和模型训练效率上也有显著提高，在训练时间上，S-AEMD-ANN模型比AEMD-ANN在上证指数数据集上平均减少了89.01%（(51.81-5.692)/ 51.81），在标普500数据集上平均减少了89.71%，S-AEMD-ANNa模型比AEMD-ANNa在上证指数数据集上平均减少了89.41%，在标普500数据集上平均减少了90.20%。

结论3 表4-2，表4-3和图4-6，图4-7的ANN模型结果与AEMD-ANN、S-AEMD-ANN模型的结果对比我们可以知道在本文所使用的两个数据集下，如果剔除前视误差对结果的影响那么EMD分解算法对ANN预测模型不仅没有起到改善结果的作用反而使得预测结果比单独的ANN模型还要差。而且结合附录中图===和===我们可以知道不管从MAPE、MAE、RMSE中任何一个评价指标来衡量模型结果的好坏，AEMD-ANN、S-AEMD-ANN这两种模型预测结果都要差于单独的ANN模型预测结果。所以在一般情况下如果剔除前视性误差对预测结果的影响那么我们没有理由认为EMD分解算法对ANN预测模型有提升作用。Dennis(2017)[70]在利用EMD和SVM以及EMD和RF(Random Forest)预测走势的时候也认为如果剔除前视性误差EMD-SVM模型和EMD-RF模型的测结果不一定优于单独的SVM和RF模型。

结论4 从表4-2、表4-3和图4-6、图4-7中的AEMD-ANN、S-AEMD-ANN和AEMD-ANNa、S-AEMD-ANNa模型的表现结果看，本文在4.2.2章节提出的模型改进方式有利于提高文献中[65、66、69、70]自适应模型的预测精度。以MAPE结果评价指标来看，而且我们可以得出AEMD-ANNa模型预测结果精度比AEMD-ANN模型预测精度在上证指数数据上和标普500指数数据上结果分别改进了24.81%（(1.733-1.303)/1.733）和15.50%。S-AEMD-ANNa模型预测结果精度比S-AEMD-ANN模型预测精度在上证指数数据上和标普500指数数据上结果分别改进了29.31%和24.76%。而且从图4-6和图4-7中我们可以看出本文所提出的改进模型S-AEMD-ANNa在两种数据集上预测表现性能都是最好的，并且都优于单独ANN模型，也就是说明EMD分解在一定程度上即使剔除前视性误差也可以改进ANN预测模型的表现性能。

4.4 本章小结

本章在4.1小结中分析前视性误差的来源及成因，并且说明了在有关EMD分解算法中存在较大的前视性误差。并在4.2小结中结合相关文献提出了一种改进的自适应模型。在4.2小节的介绍中可以知道自适应模型可以剔除前视性误差，但研究表明[70]剔除前视性误差后基于EMD的混合预测模型表现似乎不一定会优于单独的预测模型。在4.3小节的结论3中我们也证实了这一点，基于此本文提出了一种改进的混合模型S-AEMD-ANNa，结合实证研究分析可知本文提出的混合模型不论在效率上还是在预测精度上都优于文献中所提出的模型。而且本文所提出的混合预测模型S-AEMD-ANNa优于单独的ANN预测模型。另外本章的研究内容也再次证明了第三章本文所提出的基于EMD分解算法的单个模型预测结构的优越性，基于EMD的单个模型预测结构（见图）比文献中的组合模型预测机构（见图）在效率上有很大的改善，并且模型也比较简单，预测精度也有一定程度的改善。

5 结论与展望

5.1 本文结论

本文主要的研究内容是基于经验模态分解的金融时间序列预测方法的改进与优化。主要针对目前基于检验模态分解算法的组合预测模型存在的问题，如预测模型复杂、组合模型规模过大、计算量大以及现在研究存在前视性偏差等问题提出了两种改进和优化方法。而且本文通过一些真实的金融时间序列进行实证研究证明了模型改进的有效性。本文主要内容可以分为以下两点：

1 针对基于EMD分解的组合预测模型中模型数量较多、规模复杂和计算量大等问题，本文提出了一种基于EMD分解算法的单一模型预测结构。本文提出的预测模型不同于以往的预测模型，以往的预测模型是对经EMD算法分解后的每一个子序列都进行建模预测然后把预测结果相加得到最终预测结果，本文的预测模型是对经EMD算法分解后的序列建立一个预测模型，避免叠加模型过多而使得预测结构复杂、计算量大、参数过多等问题，同时模型数量的减少也在一定程度避免了模型叠加时所造成的累积误差等问题。本文第三章的“后视实验”研究以及第四章的“预测实验”研究都证明了本文提出的单一预测模型S-EMD-ANN、S-AEMD-ANN和S-AEMD-ANNa都优于文献中传统意义上的组合模型EMD-ANN[66,68,71,72]，AEMD-ANN[65、66]和AEMD-ANNa。基于EMD的单一预测模型有效缩减了网络结构，从而有效减少了模型个数和参数调节，使得模型的训练时间得到大幅度改善。并且从预测结果的各个评价指标来看单一预测模型比传统的组合预测模型在预测精度上也有一定的提升。

2 针对基于EMD分解过程中存在的前视性误差问题，本文结合文献Q. Tan(2108)[66]、X. Zhang(2015)[69]、Dennis(2017)[70]中的自适应混合模型的概念提出了一样改进的自适应预测模型。本文提出的模型主要从两个方面对文献中的自适应混合预测模型进行改善，第一个方面主要基于第一部分的研究从模型结构上进行改善将文献中组合自适应预测模型改为单一自适应预测模型，从4.3小结的实证实验研究中我们知道此改进对模型的运算效率有很大的提升，极大缩减了模型结构和训练时间同时也提升了模型的预测精度；第二方面主要基于检验模态分解数据中端点失真问题的改进，由于自适应预测模型会使得端点效应在测试数据中极大的暴露出来所以靠近端点的数据分解会严重失真，而在训练数据中却不会存在这样的问题，这样就会使得使用训练数据训练出来的模型可能在测试数据上表现很差，甚至并不一定优于单独的预测模型，X. Zhang(2015)[69]、Dennis(2017)[70]和本文4.3小结的结论3都说明如果剔除前视性偏差基于EMD分解的SVM、RF、ANN等一些混合预测模型都不一定优于单独的SVM，RF和ANN等预测模型，经过4.2.2的分析为此本文提出了一种改进的自适应预测模型，在模型训练和测试时删除一些可能存在严重失真的变量尽量用真实的数据去建模预测。4.3节大量实证实验结果表明经过变量处理的AEMD-ANNa预测模型优于不经过变量处理的AEMD-ANN预测模型，经过变量处理的S-AEMD-ANNa预测模型优于不经过变量处理的S-AEMD-ANN预测模型。并且与文献中结论不同的是本文提出的改进模型S-AEMD-ANNa在文中的各种实证结果上都要优于单独的ANN模型，这也说明了本文提出改进方案的有效性。

5.2 研究展望

本文主要是从基于经验模态分解的混合预测模型结构和混合模型中存在的前视性误差两个方面进行研究，并没有特别注重模型参数的调优后续如果进行参数调优的话可能有更好的预测结果。

针对检验模态分解中高频子序列分解不稳定，而且无序的问题，有学者提出了二次分解技术就是高频无序的第一个IMF分量进行二次分解，再对二次分解出来的序列进行建模预测。如Qu (2019)[68]在对风速预测中对最无序的第一个本征模函数（IMF）进行二次经验小波分解（EWT empirical wavelet transform），然后对各分解的子序列进行预测。Sun (2018)在对短期风速预测时对无序的第一个本征模函数进行二次自适应变分模态分解（AVMD adaptive variational mode decomposition），然后对各子序列预测。在他们的研究中二次分解可以提高预测模型的精度。在未来的研究中可以把二次分解技术运用到本文所提出的单一模型预测结构，应该也能提高预测的精度。

本文只要以最常用的ANN为基础预测模型结合EMD分解构建单一预测模型。因为文章中有大量的重复试验，ANN在各种神经网络中计算量相对较小适合大量的重复试验。在未来可能会研究更加适合处理时间序列的循环神经网络（RNN）模型如LSTM等模型结合本文的方法进行研究，在少量的实证研究表明以LSTM为基础模型集合EMD分解的预测结果会更加精确，因为未进行大量的重复试验该部分结果在附录====中给出。