Statistics and Application 统计学与应用, 2024, 13(4), 1219-1228
Published Online August 2024 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2024.134124

Hans汉斯

改进的隐马尔可夫模型股票价格预测分析

何瑞博，李俊刚

北方工业大学理学院统计学系，北京

收稿日期：2024年7月15日；录用日期：2024年8月6日；发布日期：2024年8月19日

摘　要

本文基于隐马尔可夫模型(HMM)，选取上证指数近10年的历史数据(开盘价、最高价、最低价和收盘价)进行实证分析，得出HMM模型在股票预测方面具有一定的可行性。同时，通过对传统HMM模型的输入和预测方法进行改进，对股票价格变化作出了更加准确的预测。主要步骤为：1) 数据处理。对股票价格序列进行检验并做处理，以股价波动率作为HMM模型的输入。2) 根据池化信息准则(AIC)和贝叶斯信息准则(BIC)固定最佳隐状态数目，并通过训练模型确定参数。3) 预测。相较于传统HMM模型根据股票价格序列直接得到预测数据，改进后的HMM模型则通过股价波动率计算后得出的预测得到了进一步提升。

关键词

隐马尔可夫模型，波动率，股价预测

Improved Hidden Markov Model Stock Price Prediction Analysis

Ruibo He, Jungang Li

Department of Statistics, College of Science, North China University of Technology, Beijing

Received: Jul. 15th, 2024; accepted: Aug. 6th, 2024; published: Aug. 19th, 2024

Abstract

Based on the Hidden Markov Model (HMM), this paper selects the historical data of the Shanghai Composite Index in the past 10 years (opening price, high price, low price and closing price) for empirical analysis, and concludes that the HMM model has certain feasibility in stock prediction. At the same time, through the improvement of the input and prediction methods of the traditional HMM model, more accurate predictions are made for stock price changes. The main steps are: 1) Data processing. The stock price series is tested and processed, and the stock price volatility is used as the input to the HMM model. 2) The number of optimal hidden states is fixed according to the Pooling

文章引用: 何瑞博, 李俊刚. 改进的隐马尔可夫模型股票价格预测分析[J]. 统计学与应用, 2024, 13(4): 1219-1228.
DOI: 10.12677/sa.2024.134124

Information Criterion (AIC) and Bayesian Information Criterion (BIC), and the parameters are determined by training the model. 3) Forecasting. Compared with the traditional HMM model, which directly obtains the forecast data based on the stock price series, the improved HMM model further improves the prediction obtained by calculating the stock price volatility.

Keywords

Hidden Markov Model, Volatility, Stock Price Prediction

1. 引言

股票市场作为社会经济生活中不可忽视的一部分，股票市场的发展不仅反映出国家经济的发展状况，同时也是经济运行的“晴雨表”，对股票市场运行规律的研究一直是各国政府、学术界、企业投资者努力追求的目标。

隐马尔可夫模型(Hidden Markov Model，简称 HMM)是由 Baum 和 Petrie (1966)提出的一种用于信号检测的模型。是一种含有隐状态的马尔可夫过程，隐状态代表股票市场在一段时间内所处的状态。HMM 模型是关于时序数据的概率模型，描述含有隐状态的马尔可夫链随机生成一个不可观测的状态序列，再由每个状态生成一个观测序列。股票市场可以看作是一个隐马尔可夫过程，投资者只能看到股票价格走势，而股票市场在当前时刻所处的状态是未知的。由于数据形式的不同，隐马尔可夫模型分为离散型和连续型，其中连续型的 HMM 又称为高斯隐马尔可夫模型(GHMM)。在本文中，由于股票数据是时间序列数据，因此采用的是高斯隐马尔可夫模型。

近年来，国内外有很多基于 HMM 模型对股票价格预测的研究。瞿惠和肖斌卿[1]提出，马尔可夫模型的状态具有实际意义，并与股票市场所处的走势有一一对应的关系，即股票价格处于上涨时，对应的股市状态为“牛市”。应用状态转移矩阵，成功刻画出股票收益在不同隐状态下的分布情况。余文利等人[2]利用 BIC 算法确定了 HMM 模型的隐状态数目，并基于 HMM 模型对苹果、DELL、IBM 三家上市公司股票进行了单步预测分析。Hassan 和 Nath [3]固定了 HMM 模型的隐状态，从历史数据中找到与当前股票数据模式类似的股票数据来预测一些航空公司的股票价格变化。Hassan 等人[4]根据 HMM 模型对股票预测研究的效果进一步加入了人工神经网络(ANN)和遗传算法(GA)，建立了混合模型预测股市变化。Nguyen [5]在 Hassan 和 Nath 的基础上，使用池化信息准测(AIC)和贝叶斯信息准测(BIC)来确定隐状态数目，并测试不同隐状态数目下 HMM 模型的预测效果。Gupta 和 Dhingra [6]基于隐马尔可夫模型的历史追溯法与模糊逻辑理论得到 HMM-Fuzzy 模型和人工神经网络作对比。Nguyen [7]通过把 HMM 与蒙特卡洛模拟的估计方法结合，对股票价格指数的单变量以及多变量并固定的窗口大小进行了预测，表明多变量的预测效果较好。

我们选取上证指数(000001.SS) 2014 年 12 月 12 日至 2023 年 12 月 11 日(近十年)的股票数据作为观测序列。通过 python 爬取上证指数每日的股价(开盘价、收盘价、最高价、最低价)作为分析数据。

2. 隐马尔可夫模型的理论及算法

2.1. 隐马尔可夫模型

HMM 是一种生成式的概率模型。由初始状态概率向量 π、状态转移概率矩阵 A 和观测概率矩阵 B

确定。因此，隐马尔可夫模型可以用三元符号表示，即 $\lambda=(A,B,\pi)$。HMM 模型主要解决三个问题：1) 概率计算问题。给定模型 $\lambda=(A,B,\pi)$ 和观测序列 $O=(o_1,o_2,\cdots,o_T)$，计算在模型 λ 下的观测序列 O 出现的概率 $P(O|\lambda)$。2) 学习问题。已知观测序列 $O=(o_1,o_2,\cdots,o_T)$，估计模型 $\lambda=(A,B,\pi)$ 参数，使得在该模型下观测序列概率 $P(O|\lambda)$ 最大，即用极大似然估计的方法估计参数。3) 预测问题，也称为解码(decoding)问题。已知模型 $\lambda=(A,B,\pi)$ 和观测序列 $O=(o_1,o_2,\cdots,o_T)$，求对给定观测序列条件概率 $P(I|O)$ 最大的状态序列 $I=(i_1,i_2,\cdots,i_n)$。这三个问题的解决分别对应着 HMM 模型的三个算法。概率问题采用前向算法；学习问题(即参数估计)运用 Baum-Welch 算法；预测问题运用 Viterbi 算法。

2.2. HMM 的基本算法

2.2.1. 前向算法

定义前向概率：给定隐马尔可夫模型 λ，定义到时刻 t 部分观测序列为 $o_1,o_2,\cdots,o_T$ 且状态为 q_i 的概率为前向概率，记作：

$$\alpha_t(i)=P(o_1,o_2,\cdots,o_t,i_t=q_i|\lambda) \tag{1}$$

可以递推求得前向概率 $\alpha_t(i)$ 及观测序列概率 $P(O|\lambda)$。

输入：隐马尔可夫模型 λ，观测序列 O；

输出：观测序列概率 $P(O|\lambda)$。

1) 初值：

$$\alpha_1(i)=\pi_i b_i(o_1),i=1,2,\cdots,N \tag{2}$$

2) 递推：对 $t=1,2,\cdots,T-1$，

$$\alpha_{t+1}(i)=\left[\sum_{j=1}^{N}\alpha_t(j)a_{ji}\right]b_i(o_{t+1}),i=1,2,\cdots,N \tag{3}$$

3) 终止：

$$P(O|\lambda)=\sum_{i=1}^{N}\alpha_T(i) \tag{4}$$

2.2.2. Baum-Welch 算法

假设给定训练数据集只包含 S 个长度为 T 的观测序列 $\{O_1,O_2,\cdots,O_s\}$ 而没有对应的状态序列，并由观测序列数据优化 HMM 模型 $\lambda=(A,B,\pi)$ 的参数。

输入：观测数据 $O=(o_1,o_2,\cdots,o_T)$；

输出：隐马尔可夫模型参数。

1) 初始化：对 $n=0$，选取 $a_{ij}^{(0)},b_j(k)^{(0)},\pi_i^{(0)}$，得到模型：

$$\lambda^{(0)}=\left(A^{(0)},B^{(0)},\pi^{(0)}\right) \tag{5}$$

2) 递推：对 $n=1,2,\cdots,N$，

$a_{ij}^{(n+1)}=\dfrac{\sum_{t=1}^{T-1}\xi_t(i,j)}{\sum_{t=1}^{T-1}\gamma_t(i)}$、$b_j(k)^{(n+1)}=\dfrac{\sum_{t=1,o_t=v_k}^{1}\gamma_t(j)}{\sum_{t=1}^{T}\gamma_t(j)}$、$\pi_i^{(n+1)}=\gamma_1(i)$ 右端各值按观测 $O=(o_1,o_2,\cdots,o_T)$ 和模型 $\lambda^{(n)}=\left(A^{(n)},B^{(n)},\pi^{(n)}\right)$ 计算。式中，$\gamma_t(i),\xi_t(i,j)$ 由前向算法中的期望计算得到，文中未详细推导。

3) 终止：得到模型参数：

$$\lambda^{(n+1)}=\left(A^{(n+1)},B^{(n+1)},\pi^{(n+1)}\right) \tag{6}$$

2.2.3. Viterbi 算法

维特比算法实际是用动态规划求解隐马尔可夫模型预测问题，即用动态规划求概率最大路径。此时，一条路径对应着一个状态序列。

输入：模型 $\lambda=(A,B,\pi)$ 和观测数据 $O=(o_1,o_2,\cdots,o_T)$；

输出：$I^*=(i_1^*,i_2^*,\cdots,i_T^*)$。

1) 初始化：

$$\delta_1(i)=\pi_i b_i(o_1),i=1,2,\cdots,N \tag{7}$$

$$\Psi_1(i)=0,i=1,2,\cdots,N \tag{8}$$

2) 递推：对 $t=1,2,\cdots,T$，

$$\delta_t(i)=\max_{1\le j\le N}\left[\delta_{t-1}(j)a_{ji}\right]b_i(o_t),i=1,2,\cdots,N \tag{9}$$

$$\Psi_t(i)=\arg\max_{1\le j\le N}\left[\delta_{t-1}(j)a_{ji}\right],i=1,2,\cdots,N \tag{10}$$

3) 终止：

$$P^*=\max_{1\le i\le N}\delta_T(i) \tag{11}$$

$$i_T^*=\arg\max_{1\le i\le N}\left[\delta_T(i)\right] \tag{12}$$

4) 最优路径回溯：对 $t=T-1,T-2,\cdots,1$，

$$i_t^*=\Psi_{t+1}\left(i_{t+1}^*\right) \tag{13}$$

求得最优路径 $I^*=(i_1^*,i_2^*,\cdots,i_T^*)$。

2.3. 数据处理与模型选择

本文选取上证指数 2014 年 12 月 12 日至 2023 年 12 月 11 日的股票开盘价、最高价、最低价和收盘价作为分析特征。将最近 100 天的观测数据用于测试模型准确性，其余数据作为训练数据估计模型参数。

2.3.1. 数据处理

我们首先通过相关矩阵来分析观测序列的相关性，从表 1 的相关矩阵中清楚地看出股票的开盘价、最低价、最高价和收盘价之间高度相关。

Table 1. Stock price correlation coefficient matrix table
表 1. 股票价格相关系数矩阵表

	开盘价	最高价	最低价	收盘价
开盘价	1.00	0.99	0.99	0.98
最高价	0.99	1.00	0.99	0.99
最低价	0.99	0.99	1.00	0.99
收盘价	0.98	0.99	0.99	1.000

2.3.2. 模型选择

HMM 模型隐状态数目的确定一定程度上影响预测的准确性。我们通常是使用池化信息准测 AIC 和贝叶斯信息准测 BIC，首先设置多组不同的隐状态数目，然后把不同的隐状态数目放入模型训练，根据

AIC 和 BIC 的值来检验不同的隐状态数目的 HMM 模型的性能。这两种检验的方法都适用于 HMM，因为在 Baum-Welch 算法中，EM 方法是最大化模型的对数似然。由于隐状态的数目均对应着实际含义，因此对应到股市现实状态，我们将隐状态数目设定在 2~4 之间，以保证模型的简洁性和股票预测的可能性。通常隐状态数为 2 代表“牛市”和“熊市”；数目 3 代表股市处于“上涨”、“震荡”、“下跌”。AIC 和 BIC 的计算公式为：

$$\mathrm{AIC} = -2\ln(L) + 2k \tag{14}$$

$$\mathrm{BIC} = -2\ln(L) + k\ln(M) \tag{15}$$

其中，L 表示的是模型的对数似然函数，k 是模型中估计参数的个数，M 是观测点的个数。在本文中，每个隐藏状态对应的分布为高斯分布。因此，参数的个数 $k = N^2 + 2N - 1$，其中 N 为隐状态的个数。

Table 2. Likelihood function values, AIC, and BIC values in different implicit states
表 2. 不同隐状态下的似然函数值、AIC 和 BIC 值

隐状数	似然函数值	AIC	BIC	说明
2	−43860.07	87732.14	87766.28	4 状态时，AIC 与 BI 最小
3	−42832.71	85689.42	85757.72	
4	−42152.09	84344.19	84458.01	

根据表 2 中 AIC 和 BIC 数值的大小，确定隐状态的数目为 4。即可以假定股票市场大致可分为上涨、小幅上涨、下跌、小幅下跌 4 种状态。

3. 股票价格预测

本节中，我们将使用 HMM 模型来预测股票价格。根据上节确定的隐状态数目，下面的分析采用固定状态数为 4 的固定模型。预测过程主要分为三个步骤。首先通过 Baum-Welch 算法估计出模型的参数，计算出股票价格序列每日的似然函数值。其次，找到历史数据中的似然函数值与当前一天的似然函数值最接近的股票价格序列。最后，采用历史似然法，将找出的历史数据中似然函数值最接近的股票价格与其后一天之差作为股价变化值，再将变化值加上当前的股票价格，即作为第二天股票价格的预测值。用式子表示为：$O_{t+1} = O_t + \left(O_{t-j+1} - O_{t-j}\right)$，其中 O_t 为当前股票价格，O_{t-j} 为历史数据中与当前股价似然函数值最接近的股价，O_{t-j+1} 为 O_{t-j} 后一天的股价，O_{t+1} 为预测的股价。

我们将运用此方法对未做处理的股票数据和处理后股票数据分别进行预测，通过预测误差和预测值与真实值的拟合效果来对比两种情况下预测的准确性。

3.1. 传统的 HMM 预测

我们将 python 爬取到的股票数据(开盘价、最高价、最低价、收盘价)最近 100 天的数据作为测试数据，其余作为训练数据输入 HMM 模型中进行训练，估计出模型的参数并进行预测。HMM 的隐状态数目为 4，得到预测之后的股票价格拟合图与预测误差。

图 1~4 是固定隐状态数为 4，将股票价格序列输入 HMM 模型后得到 100 天的预测值与真实值之间的拟合效果。从预测效果来看，HMM 模型能够很好地对股票价格走势做出预测。本文的预测误差采用均方误差(MSE)、绝对平均误差(RMSE)和平均绝对百分比误差(MAPE)来评价模型的预测效果。计算公式如下：

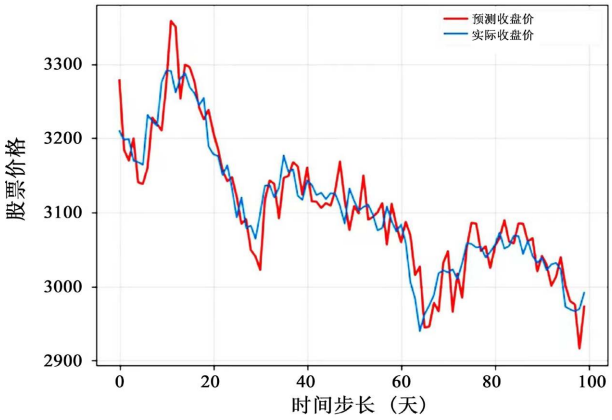

Figure 1. Comparison chart of the true and forecast values of the 100-day closing price of the Shanghai Composite Index
图 1. 上证指数 100 天的收盘价真实值与预测值对比图

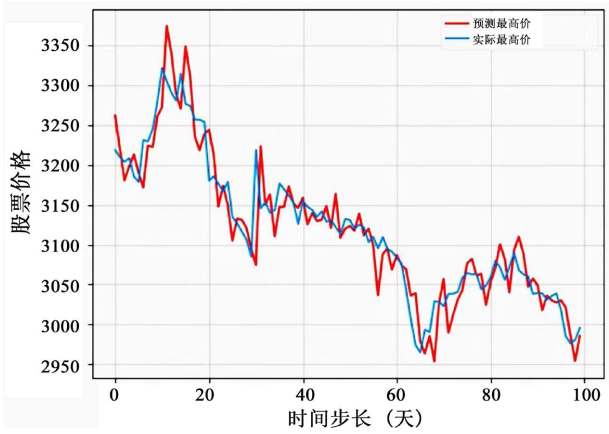

Figure 2. Comparison chart of the true and forecast values of the 100-day maximum price of the Shanghai Composite Index
图 2. 上证指数 100 天的最高价真实值与预测值对比图

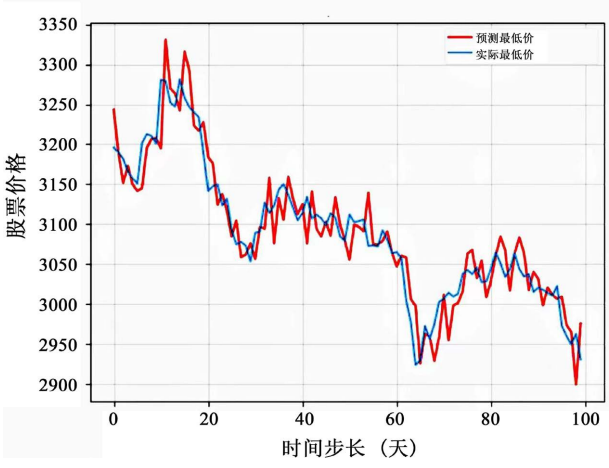

Figure 3. Comparison chart of the true and forecast values of the 100-day minimum price of the Shanghai Composite Index
图 3. 上证指数 100 天的最低价真实值与预测值对比图

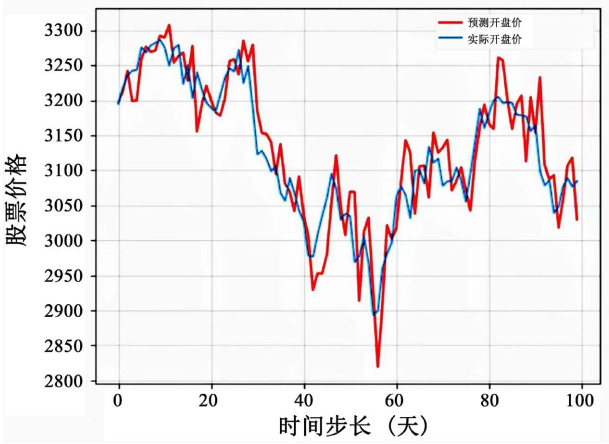

Figure 4. Comparison chart of the true and forecast values of the 100-day opening price of the Shanghai Composite Index
图 4. 上证指数 100 天的开盘价真实值与预测值对比图

$$\mathrm{MSE}=\frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{Predicted}(i)-\mathrm{True}(i)\right)^2 \tag{16}$$

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{Predicted}(i)-\mathrm{True}(i)\right)^2} \tag{17}$$

$$\mathrm{MAPE}=\frac{1}{n}\sum_{i=1}^{m}\frac{\left|\mathrm{Predicted}(i)-\mathrm{True}(i)\right|}{\mathrm{True}(i)} \tag{18}$$

其中，n 为预测点的个数。预测误差分析结果如表 3 所示。MSE、RMSE 和 MAPE 的数值越小，说明预测效果越好、越准确。

Table 3. Error analysis table of the model
表 3. 模型的误差分析表

指标	MAPE	MSE	RMSE
开盘价	0.0108363	2398.3590	48.9730
最高价	0.0082474	1296.0316	36.0004
最低价	0.0083075	1268.4298	35.6150
收盘价	0.0095636	1406.6549	37.5053

由图 1~4 的真实值与预测值之间的拟合效果可以看出，预测模型能够较为有效地捕捉到股票价格的趋势和波动，同时为了能够进一步量化分析预测效果，选取了均方误差(MSE)、绝对平均误差(MSE)和平均绝对百分比误差(MAPE)作为评价指标，并从表 3 中清楚地看到 MAPE 数值均在 0.009 左右，说明预测值与真实值之间的误差较小。

综上所述，HMM 模型对上证指数股票价格的预测表现出很好的性能，具有较高的现实意义。考虑到 HMM 对模型的输入数据有一定的要求，即要求观测数据之间相互独立，在上述的拟合过程中，我们并未对数据做处理，也得到不错的预测效果，但是下面我们对数据做简单的变形处理，让模型输入的观测序列之间满足独立性，以期得到更为准确的预测。

3.2. 改进的 HMM 模型预测

从表 1 的相关矩阵中清楚地看出股票的开盘价、最低价、最高价和收盘价之间高度相关，因此对数据做了简单处理。处理方式为将股票的价格序列数据转换为波动率来输入 HMM 模型训练。计算方式为：

$$fc = \frac{cp - op}{op},\ fh = \frac{hp - op}{op},\ fl = \frac{op - lp}{op} \tag{19}$$

其中，op 表示开盘价，cp、hp、lp 分别为收盘价、最高价和最低价，计算得到的 fc 为收盘价波动率，fh 为最高价波动率，fl 为最低价波动率。

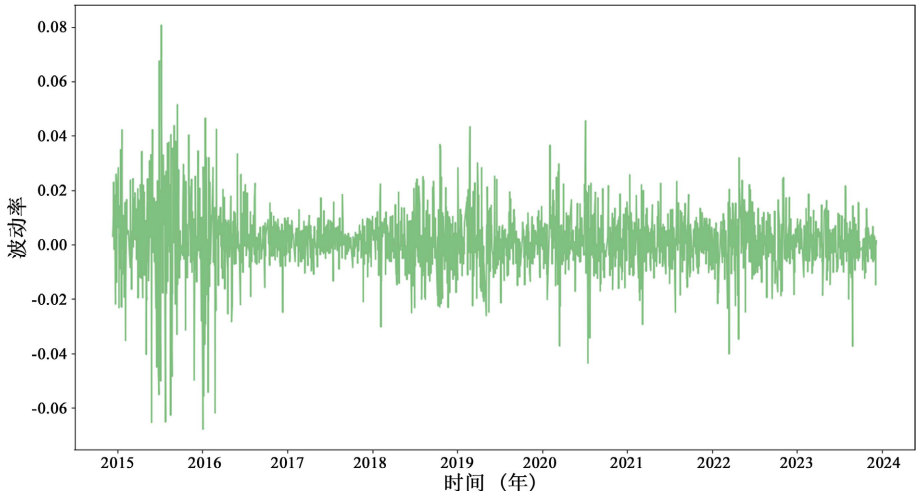

Figure 5. Stock closing price volatility
图 5. 股票收盘价波动率

从图 5 股票收盘价的波动率可以看出，尽管股票价格数据表现出的规律性不明显，且股价波动的幅度较大，但将股票价格数据转换为波动率后可以看出其波动率呈现出高斯分布的特征，更加符合 HMM 模型。

然后我们将处理后得到的收盘价波动率、最高价波动率、最低价波动率作为观测值输入高斯 HMM 模型进行股票价格预测。如此通过输入下一日的波动率对下一日的股票价格进行预测。具体的预测方法是将波动区间划分为 50 个小区间，根据积分的方式计算出波动率在每个小区间的概率，概率值最大的区间两端点的均值作为波动率的预测值，最后将得到的波动率值分别与第 n 日的收盘价进行计算，即得到第 $n+1$ 日的收盘价、最高价和最低价。

将 2014 年 12 月 12 日至 2023 年 7 月 16 日的数据作为样本数据训练模型参数，并将 HMM 模型的隐状态数目固定为 4，预测得到 2023 年 7 月 17 日到 2023 年 12 月 11 日(100 天)的股票收盘价、最低价和最高价。预测得到的上证指数股票价格的预测值与真实值的走势对比图如图 6~8 所示。

从图 6~8 可以看出，对数据做处理后得到改进的 HMM 模型，相较于传统的 HMM 模型预测性能有显著的提升，能够更好地拟合上证指数股票价格的走势。我们也通过预测误差来更好地量化预测效果的提升，如表 4 所示。

表 4 中 MAPE 的数值大小可以清楚地看到，改进后的 HMM 模型 MAPE 的数值最小为 0.004859，相较于之前的模型对股票数据的预测精度提高了一倍左右。因此，我们有理由相信改进后的 HMM 模型能够更好地对上证指数股票数据做出预测。

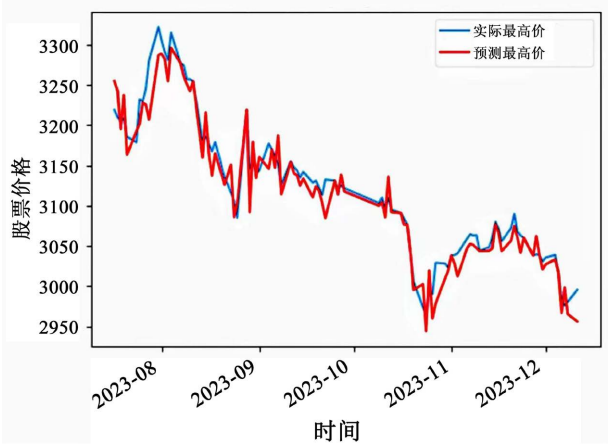

Figure 6. Trend chart of the true and forecast values of the highest price of stocks of Shanghai Composite Index
图 6. 上证指数股票最高价真实值与预测值走势图

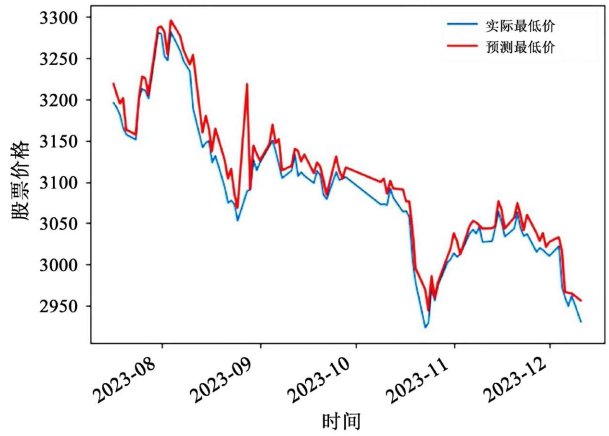

Figure 7. Trend chart of the real and forecast values of the lowest prices of stocks of Shanghai Composite Index
图 7. 上证指数股票最低价真实值与预测值走势图

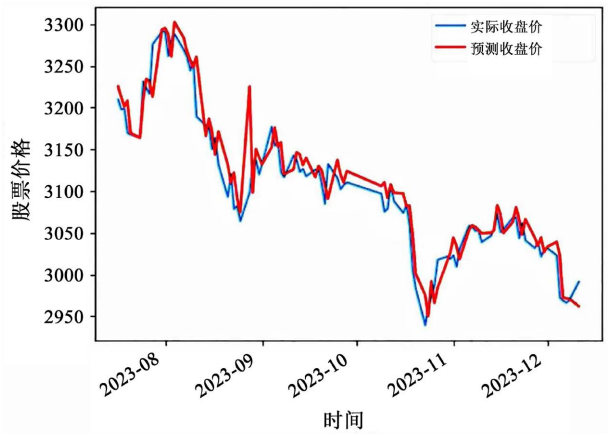

Figure 8. Trend chart of the true and forecast values of the closing prices of stocks of Shanghai Composite Index
图 8. 上证指数股票收盘价真实值与预测值走势图

Table 4. Error analysis table of the improved HMM model
表 4. 改进后 HMM 模型的误差分析表

指标	MAPE	MSE	RMSE
最高价	0.004859	411.690044	20.290146
最低价	0.005554	544.742399	23.339717
收盘价	0.005739	626.800099	25.035976

4. 结论

本文对上证指数历史股票数据进行分析，选取了开盘价、最高价、最低价和收盘价作为观测数据，基于 HMM 模型对股票价格进行预测。实证的过程主要包括数据选取及检验、隐状态数目确定、参数估计、预测和对比分析等步骤。得到如下两个结论：

1) 隐马尔可夫模型进行股票预测分析有一定的可行性，但股票观测序列(开盘价、最低价、最高价、收盘价)之间具有较强的相关性，传统的 HMM 模型并不能很好地刻画股票价格的走势。

2) 改进后的 HMM 模型在预测性能上有显著的提升，预测的精度有明显提高。同时，也说明对 HMM 模型进行合理的改进能够更好地刻画股市的波动。相信随着研究的不断深入，应用 HMM 模型对股市做预测将得到更好的效果。

基金项目

本论文工作由北京市属高校基本科研业务费(No. 110052971921/103)资助。

参考文献

[1] 瞿慧, 肖斌卿. 基于马尔科夫状态转移模型的股指收益率研究[J]. 管理科学, 2011, 24(5): 111-119.

[2] 余文利, 廖建平, 马文龙. 一种新的基于隐马尔可夫模型的股票价格时间序列预测方法[J]. 计算机应用与软件, 2010, 27(6): 186-190.

[3] Hassan, M.R. and Nath, B. (2005). Stock Market Forecasting Using Hidden Markov Model: A New Approach. 5*th International Conference on Intelligent Systems Design and Applications* (*ISDA*'05), Warsaw, 8-10 September 2005, 192-196. https://doi.org/10.1109/isda.2005.85

[4] Hassan, M.R., Nath, B. and Kirley, M. (2007) A Fusion Model of HMM, ANN and GA for Stock Market Forecasting. *Expert Systems with Applications*, **33**, 171-180. https://doi.org/10.1016/j.eswa.2006.04.007

[5] Nguyen, N. (2016) Stock Price Prediction Using Hidden Markov Model. Youngstown State University.

[6] Gupta, A. and Dhingra, B. (2012) Stock Market Prediction Using Hidden Markov Models. 2012 *Students Conference on Engineering and Systems*, Allahabad, 16-18 March 2012, 1-4. https://doi.org/10.1109/sces.2012.6199099

[7] Nguyen, N.T. (2014) Probabilistic Methods in Estimation and Prediction of Financial Models. Ph.D. Thesis, The Florida State University.