基于 HMM-XGBoost 的股价预测

李方圆* 张涛
（广西科技大学 理学院，广西 柳州 545006）

摘 要 股价预测一直是国内外很热门的研究内容，但由于股票市场会同时受到多方面的影响，这给股价预测带来了一定的挑战。论文提出了一个全新的 HMM-XGBoost 模型，并将提出的模型应用到国内股票市场，在与其他模型的预测结果对比后该模型显现出一定的优势。最后通过该模型提出了相应的投资建议，能够一定程度上指导投资者获取最大的收益。

关键词 隐马尔科夫模型；XGBoost 模型；HMM —XGBoost 模型；股价预测
中图分类号：F832 51 **文献标志码**：A **文章编号**：2095-4859(2021)04-0484-05

近年来，股价预测有众多国内外学者进行研究。ZUO Y 等[1]利用了贝叶斯网络结构进行股价市场的预测分析，结果表明利用贝叶斯网络的模型预测精度要优于传统的时间序列模型；ADEBIYYI A A 等[2]使用 ARIMA 模型和神经网络来预测股票价格；NGUYEN T T[3]提出了一种新的框架，即具有相关股票信息的深度转移(DTRSI)，利用了深度神经网络和转移学习，使用长短时记忆网络(LSTM)的基础模型从许多不同库存获得的大量数据进行预训练，以优化初始训练参数，再对模型进行微调，从而提高模型的性能。贺毅岳等[4]在股市指数建模过程中引入自适应噪声完备集合经验模态分解(CEEMDAN)结合长短时记忆网络对复杂序列中长期依赖关系高效的建模能力，提出一种指数预测建模方法 C —LSTM；杨青等[5]基于深度神经网络优化技术，构造了一个深层 LSTM 神经网络并将其应用于全球 30 个股票指数的不同期限预测中；商晔等[6]考虑了时域相关性的隐马尔科夫模型，并且利用该模型对股票市场的价格进行了预测，同时也将遗传算法应用于隐马尔科夫模型的参数训练中；李玉明等[7]在股票交易决策和股价预测中提出了深度强化学习理论，通过实验数据证明了模型的可靠性和可用性，并将该模型与传统模型进行了比较，证明了其优越性；徐颖等[8]提出了一种结合 k 均值聚类和集成学习的混合预测模型，对一些国内股票的实验结果表明，混合预测模型具有较高的股票价格预测精度；李辉等[9]为了解决股票模型预测中由于数据复杂而导致的预测精度低、训练复杂等问题，提出了一种基于特征选择(FS)和长短期记忆(LSTM)算法的股票收盘价预测方法。

随着计算机技术与现代金融理论的发展，非线性预测方法逐渐成为金融序列预测分析的研究热点，本文为了在环境复杂多变的股票市场中研究出一个相对稳定的股票投资策略，取代主观性强的人工交易方式，能够较大限度地保障投资者的收益。利用了 BAUM L E[10]提出的 HMM 模型的自身特点对当前国内股票市场的状态进行识别，接着采用陈天奇[11]近些年提出的 XGBoost 模型进行预测分析。根据查询文献显示，以往研究者均采用的是单一的 HMM 模型或者 XGBoost 模型对股价的预测，并没有将两者进行结合使用，因此，本文提出了一个新的 HMM —XGBoost 模型来对股价进行预测。本次研究是从 HMM 的隐状态出发结合分类模型 XGBoost，即从各个不同的隐状态出发研究不同情形下应该在哪个时刻进行合理的投资，来获取较大收益。最后将提出的模型应用到国内股票市场进行收益预测，结果显示了该模型的优越性。

1 数据说明

选取 2010 年 1 月 —2021 年 3 月贵州茅台的统计数据作为预测的源数据，数据来源于网易财经。其中各变量信息如下表 1 所示：

* 作者简介：李方圆，男，陕西咸阳人。硕士研究生。研究方向：金融统计。

表 1　变量说明表

原始变量		指标变量	
变量名称	变量说明	变量名称	变量说明
Date	交易日期	MA	移动平均线
Clspr	收盘价	MACD	平滑异同平均线
Openprc	开盘价	LogDel	最高价与最低价的对数差
Hiprc	当日最高价	KDJ	随机指标
Loprc	当日最低价	BIAS	乖离率指标
Dnshrtrd	成交量	OBV	能量潮指标

2　基于 HMM-XGBoost 模型的股价预测

2.1　HMM 模型

HMM 模型(隐马尔科夫模型)是关于时序的模型概率,描述由一个隐藏的马尔科夫链随时生成的不可观测的状态随机序列,再由各个状态生成一个观测从而产生随机序列的过程[12]。隐马尔科夫模型具有以下两点基本假设:

1)隐藏的马尔科夫链在任意时刻的状态只依赖前一时刻的状态,与其他的时刻状态无关。

2)任意时刻的观测只依赖该时刻的马尔科夫链的状态,与其他观测状态无关。

隐马尔科夫模型是由不可观测到的隐式状态 Hidden state 和由每个隐式状态生成的观测序列组成。由于隐马尔科夫模型中的路径状态预测问题采用的是维特比算法,根据该预测问题可以确定不同的隐状态在不同时期所处的最大概率。

针对本次预测问题,首先需要确定当前股票市场的隐状态,采用 HMM 模型对市场状态进行了识别预测[13]。查阅资料显示我国国内股票市场的隐藏状态个数一般设定为 5,分别为:平稳上升、震荡上升、缓慢下降、震荡下降、基本持平时期。将预处理后的数据带入 HMM 模型中,可以得到图 1 中的 5 种不同隐状态及其走势情况。

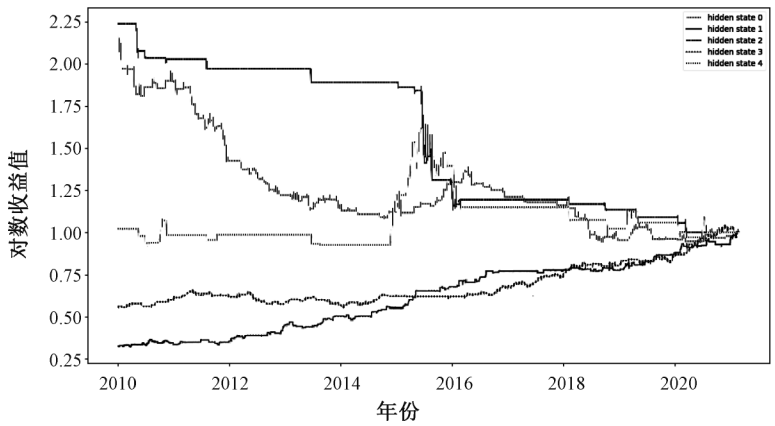

图 1　各个隐状态在不同时间的收益情况

上图的横坐标代表不同时间,纵坐标代表的则是该股票每日收益(收盘价减去开盘价)的对数值。根据图 1 可知,Hidden state0(状态 0)和 Hidden state2(状态 2)代表主要处于股市下跌状态,很容易会出现阶段性的“股灾”现象,并且其中 Hidden state2(状态 2)有断崖式下跌阶段,而 Hidden state0(状态 0)处于平稳下跌。Hidden state1(状态 1)和 Hidden state3(状态 3)均处于缓慢上涨阶段,说明这个阶段为股市的牛市,根据数据显示该阶段涨幅达到 50% 左右。Hidden state4(状态 4)特征并不显著,意味当前阶段除了中期有小幅度的波动外,整体并未有显著发展趋势,整体处于震荡调整阶段。通过隐状态走势显示,投资者能够从宏观上很好地把握一段时间内的市场状态识别,并且较为详细的解释出不同状态代表的含义。由此说明隐马尔科夫模型在股价走势预测方面是相对准确的,能够对不同市场状态的变化给出及时的反映。在能够对市场状态识别的前提下,进一步地预测分析的效果才是最佳的。

2.2　XGBoost 模型

XGBoost 是 boosting 算法的一种,是以决策树为基础的一种梯度提升算法。通过多轮迭代,每轮迭代产生一个弱分类器,每个分类器在上一轮分类器的残差基础上进行训练。对弱分类器的要求一般是足够简单,并且是低方差和高偏差的。因为训练的过程是通过降低偏差来不断提高最终分类

器的精度。

弱分类器一般会选择为 CART TREE。由于上述高偏差要求每个分类回归树的深度并不会很深，最终的总分类器是将每轮训练得到的弱分类器加权求和得到的。因此，模型目标函数为：

$$L(\Phi)=\sum_i l(y_i,\hat{y_i})+\sum_k \Omega(f_k)$$

$$\Omega(f_k)=\gamma T+\frac{1}{2}\lambda \| w \|^2$$

其中：$L(\Phi)$为线性空间上的表达式；i 是代表第 i 个样本；k 是指第 k 棵树；γT 为叶子节点数量；后一项为叶子节点权重向量的 l_2 范数。

如果不结合 HMM 模型，单独使用 XGBoost 模型来进行股价测[14]，效果并不是太理想。根据表 2 所示，仅运用 XGBoost 模型对于上涨状态预测精确率 65%，对于下跌预测精确率仅为 68%。

表 2　XGBoost 模型预测结果

预测结果	精确率	支持度
下跌	0.68	398
上涨	0.65	415

2.3　HMM－XGBoost 模型

根据上述 HMM 模型和 XGBoost 模型，我们提出了 HMM-XGBoost 模型。该模型首先从 HMM 模型中预测出所处的隐状态，接着对于不同的状态进一步预测分析。

1）隐状态的确定

初始化：

$$\delta_1(i)=\pi_i b_i(o_1),i=1,2,\cdots,N$$

$$\psi_1(i)=0,i=1,2,\cdots,N$$

递推，对 $t=2,3,\cdots T$：

$$\delta_t(i)=\max_{1\leqslant j\leqslant N}[\delta_{t-1}(j)a_{ij}]b_i(o_t),i=1,2,\cdots,N$$

$$\psi_t(i)=\arg\max_{1\leqslant j\leqslant N}[\delta_{t-1}(j)a_{ji}],i=1,2,\cdots,N$$

终止：

$$P^*=\max_{1\leqslant i\leqslant N}\delta_T(i)$$

$$i_T{}^*=\arg\max_{1\leqslant i\leqslant N}[\delta_T(i)]$$

最优路径回溯。

对 $t=T-1,T-2,\cdots,1$：

$$i_t{}^*=\psi_{t+1}(i_{t+1}{}^*)$$

2）HMM-XGBoost 模型的目标函数

由于具有多种状态的存在，需要重新定义不同状态下的目标函数，P_c 为状态 C 时对应的观测结果对应观测序列的概率。不同状态用 c 表示，且 c 的取值范围为 1 至 5：

$$L(\Phi_c)=\sum_i l(y_{i,c},\hat{y}_{i,c})+\sum_k \Omega(f_{k,c})$$

$$\Omega(f_{k,c})=\gamma T_c+\frac{1}{2}\lambda \| w_c \|^2$$

$l(y_{i,c},\hat{y}_{i,c}^{(t-1)}+f_t(x_{i,c}))$的二阶泰勒展开式为：

$$l(y_{i,c},\hat{y}_{i,c}^{(t-1)}+f_t(x_{i,c}))\approx l(y_{i,c},\hat{y}_{i,c}^{(t-1)})+g_{i,c}f_t(x_{i,c})+\frac{h_{i,c}}{2}f^2{}_t(x_{i,c})$$

带回目标函数：

$$L(\Phi_c)=\sum_{i=1}^{n}[l(y_{i,c},\hat{y}_{i,c}^{(t-1)})+g_{i,c}f_t(x_{i,c})+\frac{h_{i,c}}{2}f^2{}_t(x_{i,c})]+\sum_k \Omega(f_{k,c})$$

正则化项展开：

$$\sum_k \Omega(f_{k,c})=\sum_{k=1}^{t}\Omega(f_{k,c})=\Omega(f_{t,c})+\sum_{k=1}^{t-1}f_{k,c}$$

代入原目标函数得到：

$$L(\Phi_c)=\sum_{i=1}^{n}[l(y_{i,c},\hat{y}_{i,c}^{(t-1)})+g_{i,c}f_t(x_{i,c})+\frac{h_{i,c}}{2}f^2{}_t(x_{i,c})]+\Omega(f_{t,c})$$

合并一次项、二次项，再加上化简后得到分状态的最终目标函数：

$$L_c{}^t=\sum_{j=1}^{T}[G_{j,c}w_{j,c}+\frac{1}{2}(H_{j,c}+\lambda)w_{j,c}{}^2]+\gamma T_c$$

其中：$G_j=\sum_{i\in I_j}g_{i,c}$，$H_j=\sum_{i\in I_j}h_{i,c}$

由于上式子为 $w_{j,c}$ 的一元二次函数。由于 $(H_{j,c}+\lambda)>0$，则 $w_{j,c}=-\frac{G_{j,c}}{(H_{i,c}+\lambda)}$ 时，取最小值的概率 P 为：$P=p_c\times p(w_{j,c})$，最小值为 $-\frac{1}{2}\frac{G_{j,c}{}^2}{H_{i,c}+\lambda}$。其中：$p(w_{j,c})=f(w_{j,c})$，其中 f 为激活函数。

2.4　HMM-XGBoost 模型预测效果

根据上述建立的 HMM-XGBoost 模型对股票市场进行预测分析。股票当日收盘价大于昨日收盘价则判定为上涨用 1 表示；股票当日收盘价小于等于昨日收盘价则判定为下跌用－1 表示。分类

效果如下表 3 和表 4 所示，对于上涨状态预测精确率高达 70%，对于下跌预测精确率更是高达 74%，模型整体预测的准确率能达到 77%，召回率 70%，灵敏度和特异度分别为 0.70 和 0.74。

表 3　HMM-XGBoost 预测结果

预测结果	精确率	支持度
−1	0.74	363
1	0.70	450

表 4　模型评价指标

准确率（Precision）	召回率（Recall）	灵敏度（Sensitivity）	特异度（Specificity）
0.77	0.70	0.70	0.74

为了验证 HMM-XGBoost 模型的预测效果，在此引用了机器学习中常用的随机森林模型作为对比，进行了对照实验。随机森林预测结果如表 5 所示。根据实验结果显示，HMM-XGBoost 模型预测准确率 77% 远大于随机森林的预测准确率，说明 HMM-XGBoost 模型在基于隐状态前提下的分类预测效果要明显高于随机森林预测效果。从而说明文中建立的 HMM-XGBoost 模型效果较为良好。

表 5　随机森林预测结果

预测结果	精确率	支持度
−1	0.57	363
1	0.66	450

2.5　投资建议

由于国内股市实行“T＋1”交易制度即当日买进的股票，要到下一个交易日才能卖出。因此根据交易规则以及 HMM-XGBoost 模型的预测状况，制定了如下详细的交易投资策略：

1）当整个市场判断后处于状态 0 和状态 2 时，无论时机多好都建议不要入场，因为从整体而言市场处于下跌状态，此时并不适合投资者进场。

2）当市场判断后处于状态 4 时，说明股市处于震荡阶段，此时风险较大，也同样不建议投资者入场。

3）当市场判断后处于状态 1 和状态 3 时，此时如果预测结果为 1，则表明下一个交易日的股价大概率就会上涨，选择以当天的收盘价买入，并且在下一个交易日择时卖出；如果预测结果为−1，表示下一个交易日的股价大概率会下跌，选择根据今日走势择时抛出。

根据上述投资策略，投资者会大概率获得收益，并且不用承担过多风险。经实际测试得到，如果按照上述交易投资策略进行投资时，其具体收益状况如下图 2 所示。由于上文中提出的投资策略，并未考虑到任何仓位控制和止盈止损的措施，如何合理通过添加风险控制措施来提高模型收益也是接下来要研究的主要内容。

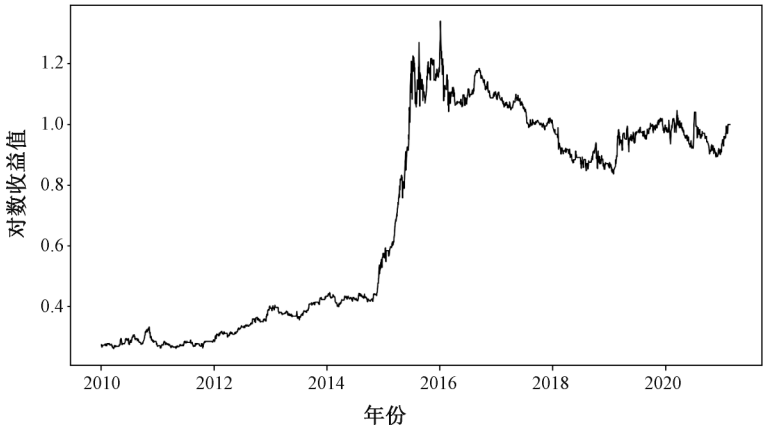

图 2　基于 HMM-XGBoost 模型的收益状况

上图的横坐标代表该股票所处的不同时间段，而纵坐标代表了基于模型预测出来的对应时间股票收益的对数值。根据图 2 显示，当采用上述模型进行投资时，投资收益会从 2015 年开始出现大幅上涨，这其中与国家政策、国内外大环境等方面密切相关。虽然 2016 年至 2018 年有一定程度的下

跌,但是如果投资者是在 2015 年之前进行的投资并且是长期持有,那收益就会较高。与人为主观性投资相比,该模型的收益相对来说也是较为可观,可以基本满足股市上投资者的需求。

3 结论

首先对于股市所处的阶段利用 HMM 模型进行状态识别,是处于平稳上升、震荡上升、基本持平、缓慢下降、震荡下降这 5 个阶段的哪一阶段;接着对不同阶段引入 XGBoost 模型进行股价涨跌预测。为了体现出模型的预测效果,引入了随机森林模型作为对比实验,结果显示该模型的预测效果是要远高于随机森林模型。并且我们针对提出的模型进行了实测分析,以评估其收益情况。

本文运用 HMM-XGBoost 模型分析了国内股票市场,针对不同时期下市场所处的不同状态,给出了相应的投资建议,使得投资者在不用承担太多风险下,获得最大限度的收益。本文提出的 HMM-XGBoost 模型效果要远比其它的模型好,这是由于市场整体环境在不断变化,只有在针对隐状态的分析基础上进行预测才会有更好地预测效果。因此对于国内股票市场上的股价涨跌预测问题,可以将我们提出的 HMM-XGBoost 模型预测结果作为参考。

对于今后的研究还可以在以下两大方面上进行探索:一是将数据挖掘和大数据技术更好地适配国内金融市场,对于企业来说可以利用大数据和数据挖掘的技术进行风险控制和风险识别;二是在未来市场上可以引入深度学习的内容,将深度学习的相关技术手段用于金融方面的预测分析,实现我国金融智能化。

参考文献

[1] ZUO Y,EISUKE KITA.Stock price forecast using bayesian network[J].Expert Systems With Applications,2012,39(8):6729-6737.
[2] ADEBIYI A A ,ADEWUMI A O ,et al.Comparison of arima and artificial neural networks models for stock price prediction[J].Journal of Applied Mathematics,2014,2014(1):1-7.
[3] NGUYEN T T,YOON S.A novel approach to short-term stock price movement prediction using transfer learning[J].Applied Sciences,2019,9(22):4745.
[4] 贺毅岳,高妮,韩进博,等.基于长短记忆网络的指数量化择时研究[J]统计与决策,2020,36(23):128-133.
[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] 商晔.隐马尔可夫模型参数训练的改进及在股市预测中的应用[D]上海:上海交通大学,2011.
[7] LI Y M. Application of deep reinforcement learning in stock trading strategies and stock forecasting[J].Computing,2019,102(prepublish):1-18.
[8] XU Y, et al.A hybrid two-stage financial stock forecasting algorithm based on clustering and ensemble learning[J].Applied Intelligence,2020,50(prepublish):1-16.
[9] LI H, HUA J, LI J, et al.Stock forecasting model fs-lstm based on the 5g internet of things [J].Wireless Communications and Mobile Computing,2020,2020(6):1-7.
[10] BAUM L E,PETRIE T.Statistical inference for probabilistic functions of finite state markov chains[J].Annals of Mathematical Statistics,1966,37(6):1554-1563.
[11] CHEN T,GUESTRIN C.xgboost: A scalable tree boosting system[C]//Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining .2016: 785-794.
[12] 李航.统计学习方法[M]北京:清华大学出版社,2019:193-213.
[13] 韩晴.基于 HMM 的股指期货交易策略及优化研究[D]西安:西北大学,2019.
[14] 伯毅.基于 XGBoost 模型的短期股票预测[D]哈尔滨:哈尔滨工业大学,2018.

(责任编辑 陈葵晞)