DOI:10.13878/j.cnki.jnuist.20221008002

韩金磊[1]　熊萍萍[1]　孙继红[2]

基于 LSTM 和灰色模型的股价时间序列预测研究

摘要

影响股价的因素错综复杂,因此在考虑多变量情形下,对时间序列中常用的长短期记忆网络(LSTM)进行修正,并选取股票价格进行预测.首先,采用方差膨胀因子(VIF)进行变量的筛选,再结合自适应提升法(Adaboost)模型查看特征变量的重要程度.其次,用爬虫对投资者情绪进行文本分析,计算情绪指数等指标并揭示其与股价的关系.然后,对格力电器、飞科电器、美的集团 3 支股票进行股价预测,对比多层感知器(MLP)模型、LSTM 模型,并选择适当的模型作为基准模型,在基准模型的基础上加上情绪指数、投资者关注度等指标构建了 LSTM-EM 模型.进一步,在考虑了投资者情绪后对残差项使用 GM(1,1)模型进行修正.实证结果表明,该模型能对股价进行较为精确的预测.

关键词

股价预测;综合预测;文本分析;误差修正;长短期记忆网络(LSTM);灰色模型

中图分类号 F222.3;F830.91;TP18

文献标志码 A

收稿日期 2022-10-08

资助项目 国家社科基金一般项目(23BGL232);教育部人文社会科学研究规划项目(22YJA630098);江苏省社会科学项目(22GLB022);大学生创新创业训练计划项目(202210300036Z)

作者简介

韩金磊,男,硕士生,研究方向为时间序列分析、公司金融.1825971908@qq.com

熊萍萍(通信作者),女,博士,教授,研究方向为时间序列分析、金融统计预测分析、灰色系统建模.xpp8125@163.com

1 南京信息工程大学 管理工程学院,南京,210044

2 中国教育科学研究院 高等教育研究所,北京,100088

0 引言

股权融资是直接融资的一种重要渠道,能够将现实生活中多余的资金筹集起来,缓解融资约束.在一个成熟的市场上,股票的价格应当反映股票的内在价值.然而,内在价值的判定涉及折现率等一系列问题,很难做出精确的度量.在投资时,多数投资者所关注的仅是些常见的技术性指标和基本财务指标,但即使在资本市场发达的美国依然很难达到强有效或者半强有效市场,股票价格不免遭到国家政策、地缘政治、投资者情绪等突发性宏观或微观因素的影响.因此,在进行投资时,对股价做一个较为综合的考量,能帮助投资者降低投资风险,具有一定现实意义.

股价数据具有非线性、非平稳、高噪声、强时变性等非常显著的特征,对股价的预测具有一定挑战.纵观已有研究,早期的研究主要运用技术分析或经典时间序列模型.其中,技术分析是结合股票的成交量、成交价格等常见的市场指标来判定股价走势.研究者常结合时间序列模型来预测股价,常用的模型有差分整合移动平均自回归(ARIMA)模型[1]、针对金融数据波动聚集效应的广义自回归条件异方差(GARCH)模型[2]及 ARIMA 模型的变种——向量自回归(VAR)模型[3].除传统的计量模型外,灰色模型[4]、BP 神经网络模型[5]以及模糊理论[6]在股价预测中也有较多应用,但这些模型存在一定的缺陷,即对非线性、长期时间序列的效果较差.

随后,学者们对各种模型进行了结合与改进.针对模型中存在的多重共线性问题,使用主成分分析(PCA)[7]或 LASSO 方法[8]进行变量的降维筛选.参数的优化也是其中的改进方向之一,智能优化算法借鉴自然中常见的现象设计算法,因原理简单、收敛速度快成为较常用的工具,其中,细菌群体趋药性、果蝇、遗传优化等算法在处理模型的最优权值结构中有广泛的应用,且多应用于超参数较多的 BP、Elman 神经网络以及非线性支持向量机等机器学习模型[9-12].鉴于模型各有针对性,有学者运用 ARIMA、GM(1,1)、RBF 神经网络等多个模型构成了一个集成预测结果[13].有研究提出一种联合卷积神经网络(CNN)和长短期记忆网络(LSTM)方法的预测模型,使用 CNN 提取股价的图像特征,使用 LSTM 提取股价的时序特征[14].针对股价非线性非平稳的特征,研究者使用小波分析对股价自身或者是股价的影响因素先进行分解,再建立 ARIMA 模型、BP 神经网络或支持向量

回归机(SVR)模型进行预测与重构,也获得了不错的效果[15-18].同小波分解一样,经验模态分解(EMD)原是工程领域用于分解复杂信号的一种方法,由于其对非线性非平稳序列更好的适用性广受研究者青睐,与长短期记忆网络(LSTM)、非线性孪生支持向量回归机(TSVR)等模型结合可以有效地预测股价[19-20].

随着人工智能和机器学习的发展,更多研究涉及深度学习[21].长短期神经网络(LSTM)能有效提取股价序列中的信息且在一定程度上缓解循环神经网络(RNN)梯度消失和梯度爆炸的问题[22],在股价时间序列中得到了广泛应用.有研究选取技术性、基本面指标基于 LSTM、门控循环单元结构(GRU)模型构建混合神经网络,同时,多延迟嵌入张量处理技术(MDT)与注意力机制(CBAM)也被较好地与 LSTM 模型相结合[23-24].近年来,研究者开始关注一些投资者心理方面的因素.百度搜索指数、新浪微博情绪指数被证实对股价的短期预测具有一定的作用[25-27].机器学习为这种非结构化数据的处理提供了技术支持,媒体报道、公司新闻、微博评论等非结构化数据被用于提取情绪时间序列[28-30].随后,有研究在此基础上考虑了财务指标、技术性指标和网络舆情 3 种信息来源,使用支持向量分类器(SVM)对股价的涨跌进行预测[31].

结合已有文献,不难发现研究中还存在着值得改进的方面.首先在考虑特征变量时存在过于随意或考虑不足的问题.其次,结合文本对股价进行分析的文献相对较少,且对股价的文本分析往往只采用词典法,然而所用的金融词典并不完善,在反映投资者情绪时效果可能欠佳.另外,残差项作为预测值与真实值的误差往往包含了许多未被利用到的有用信息,但大多数研究者对此关注甚少.针对上述问题,本文提出了以下解决方案:1)考虑基本面和技术分析及投资者情绪等多层面指标作为特征对股价问题进行分析和预测,从多个层面选取特征变量;2)尝试创建股市语料库,并使用朴素贝叶斯的方法进行训练,对投资者的每日情绪指数进行较为精确的测算,以便更好地衡量投资者情绪;3)运用对小样本常用的灰色 GM(1,1)模型对预测与真实值的残差项进行修正,更加充分地挖掘股价内在信息.

1 方法与原理

1.1 LSTM 模型

LSTM(Long Short-Term Memory)是 RNN 的变种,在每一个重复的模块中有 4 个特殊的结构,以一种特殊的方式进行交互.在图 1 中每一条细线传输着一整个向量,黄色的圆圈代表着一种运算操作,蓝色的矩阵代表着学习到的神经网络层.LSTM 模型的核心思想是“细胞状态”.“细胞状态”类似于传送带,直接在整个链条上运行,只有一些少量的线性交互,信息损失相对较少.图 1 中,$\boldsymbol{i}_t$,$\boldsymbol{f}_t$,$\boldsymbol{o}_t$,$\boldsymbol{c}_t$,$\boldsymbol{h}_t$,$\tilde{\boldsymbol{c}}_t$ 分别表示输入门、遗忘门、输出门、细胞态、记忆体以及候选态[32].

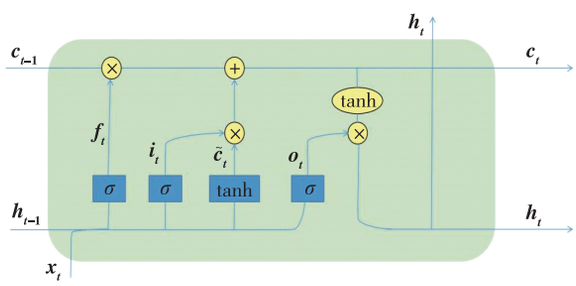


图 1 LSTM 工作原理

Fig. 1 LSTM working principle

步骤 1:决定细胞中丢弃的信息,该操作由遗忘门来完成.首先读取当前输入 $\boldsymbol{x}_t$ 和前神经元信息 $\boldsymbol{h}_{t-1}$,由遗忘门 $\boldsymbol{f}_t$ 来决定丢弃的信息,具体计算公式如下:

$$\boldsymbol{f}_t=\sigma(\boldsymbol{W}_f[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_f). \tag{1}$$

步骤 2:确定细胞状态存放的新信息,其中,sigmoid 层作为“输入门层”,决定更新值,tanh 层创建新的候选值向量 $\tilde{\boldsymbol{c}}_t$ 加入到状态中,具体计算公式如下:

$$\boldsymbol{i}_t=\sigma(\boldsymbol{W}_i[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_i), \tag{2}$$

$$\tilde{\boldsymbol{c}}_t=\tanh(\boldsymbol{W}_c[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_c). \tag{3}$$

步骤 3:更新细胞状态,将 $\boldsymbol{c}_{t-1}$ 更新为 $\boldsymbol{c}_t$.将旧状态和 $\boldsymbol{f}_t$ 相乘,丢弃需要丢弃的信息,并确定新的候选值 $\boldsymbol{i}_t\cdot\tilde{\boldsymbol{c}}_t$,具体计算公式如下:

$$\boldsymbol{c}_t=\boldsymbol{f}_t\cdot\boldsymbol{c}_{t-1}+\boldsymbol{i}_t\cdot\tilde{\boldsymbol{c}}_t. \tag{4}$$

步骤 4:确定输出,使用 sigmoid 层确定细胞状态中输出的部分,接着将细胞状态通过 tanh 进行处理,并将其和 sigmoid 层输出相乘,具体计算公式如下:

$$\boldsymbol{o}_t=\sigma(\boldsymbol{W}_o[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_o), \tag{5}$$

$$\boldsymbol{h}_t=\boldsymbol{o}_t\cdot\tanh(\boldsymbol{c}_t). \tag{6}$$

1.2 GM(1,1)模型

灰色系统理论是一种针对小样本、贫信息的数据挖掘方法,在部分信息已知,部分信息未知的灰状态下,具有十分优良的性能.GM(1,1)是经典灰色模

型,能够简单有效地挖掘出数据的内在信息[33],主要的建模步骤如下:

设:$\boldsymbol{X}^{(0)}=(x^{(0)}(1),x^{(0)}(2),\cdots,x^{(0)}(n))$ 为系统特征变量序列,其中:$x^{(0)}(k)\geqslant 0,k=1,2,\cdots,n$;$\boldsymbol{X}^{(1)}$ 为 $\boldsymbol{X}^{(0)}$ 的一阶累加生成(1-AGO)序列;$\boldsymbol{Z}^{(1)}$ 为 $\boldsymbol{X}^{(1)}$ 的紧邻均值生成序列,见式(7)—(8).

$$x^{(1)}(k)=\sum_{i=1}^{k}x^{(0)}(k),k=1,2,\cdots,n, \tag{7}$$

$$z^{(1)}(k)=\frac{1}{2}[x^{(1)}(k)+x^{(1)}(k-1)],\quad k=2,3,\cdots,n. \tag{8}$$

令参数列 $\boldsymbol{\beta}=[a,b]^{\mathrm{T}}$,其中,$a$ 为模型的发展系数,b 为灰色作用量,即:

$$\boldsymbol{B}=\begin{bmatrix}-z^{(1)}(2) & 1\\ -z^{(1)}(3) & 1\\ \vdots & \vdots\\ -z^{(1)}(n) & 1\end{bmatrix},\quad \boldsymbol{Y}=\begin{bmatrix}x^{(0)}(2)\\ x^{(0)}(3)\\ \vdots\\ x^{(0)}(n)\end{bmatrix}. \tag{9}$$

由最小二乘估计可得 $\boldsymbol{\beta}=(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B})^{-1}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{Y}$,称 $\frac{\mathrm{d}x^{(1)}}{\mathrm{d}t}+ax^{(1)}=b$ 为 GM(1,1)的白化微分方程,则 GM(1,1)模型的近似时间响应式为

$$\hat{x}^{(1)}(k)=\left(x^{(0)}(1)-\frac{b}{a}\right)\mathrm{e}^{-a(k-1)}+\frac{b}{a},\quad k=1,2,\cdots,n. \tag{10}$$

进一步,对式(10)进行累减还原,并求出对应 $\boldsymbol{X}^{(0)}$ 的时间响应式,计算过程见式(11)—(12),其中 $\alpha^{(1)}$ 表示一阶累减生成算子.

$$\hat{x}^{(0)}(k)=\alpha^{(1)}\hat{x}^{(1)}(k)=\hat{x}^{(1)}(k)-\hat{x}^{(1)}(k-1),\quad k=1,2,\cdots,n, \tag{11}$$

$$\hat{x}^{(0)}(k)=(1-\mathrm{e}^{a})\left(x^{(0)}(1)-\frac{b}{a}\right)\mathrm{e}^{-a(k-1)},\quad k=1,2,\cdots,n. \tag{12}$$

1.3 综合预测与残差修正的主要步骤

本文所采取的综合预测与残差修正的主要步骤如下:

1)获取基本面指标;

2)爬取东方财富网的股评、百度指数;

3)使用 SnowNLP 模型进行情绪指数的计算;

4)使用自适应提升法(AdaBoost)模型进行特征变量的提取,并参考方差膨胀因子(VIF)进行取舍;

5)使用多变量 LSTM 模型对股价进行预测;

6)对预测结果的残差项进行修正;

7)评估模型.

预测模型的技术路线如图 2 所示.

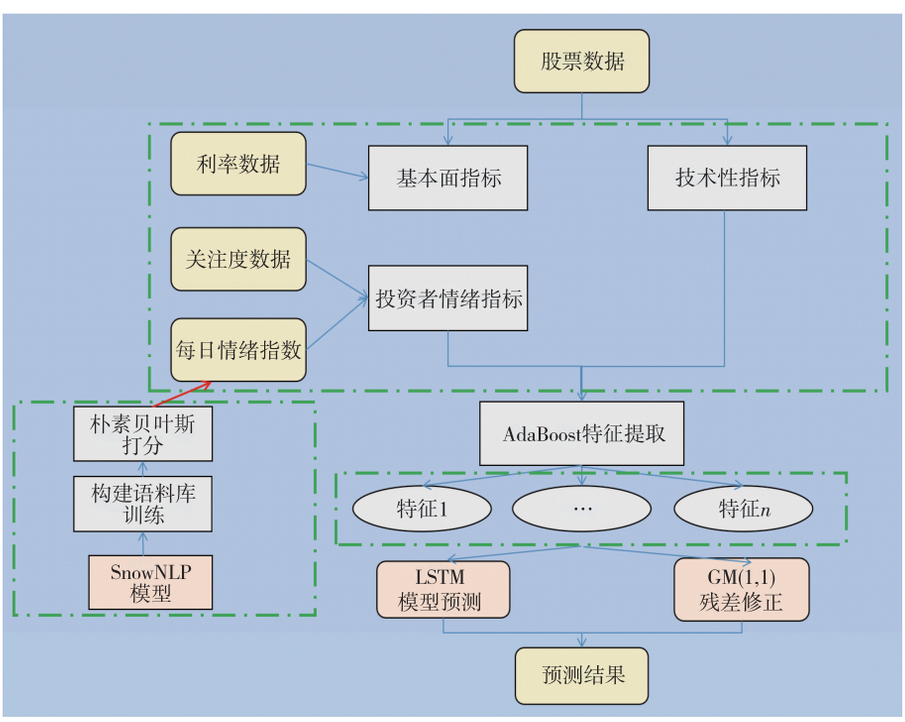


图 2 预测模型的技术路线

Fig. 2 Technical route of stock price time series prediction model

2 模型构建

2.1 数据来源

本文所选的数据大致有 3 个来源,其中:股价用 R 语言中的 pedquant 包进行获取,计算相关技术性指标和基本的财务指标(财务指标如权益净利率等仅在年报、半年报、季报才能获得,而股价是日度数据,因此,仅选取市盈率等指标);利率选取了上海银行间同业拆放利率作为市场无风险利率,其对应的数据来源于官网(https://www.shibor.org/shibor/),汇率的数据来源于中国货币网(https://www.chinamoney.com.cn/),选取的是人民币兑美元的汇率;投资者情绪相关的数据源于东方财富网(http://guba.eastmoney.com/)及百度指数官网(https://index.baidu.com/),这些数据主要使用 Python 爬取及处理,将在下一部分进行详细描述.数据时间范围为 2021 年 7 月 12 日至 2022 年 4 月 25 日,剔除不交易的日期,共计 191 天.为验证模型的稳健性,选取上证指数(1A0001)以及格力电器(000651)两组数据,并着重对格力电器的股价进行分析及预测.

2.2 变量筛选

为防止过拟合、多重共线性等问题,按照以下步骤进行变量的筛选.首先进行相关性及显著性检验;接着,为防止多重共线性,使用计量中常用的方差膨胀因子(VIF)进行判断,进行变量的筛选;随后,使用 AdaBoost 模型观察变量的重要程度.限于篇幅,相关性及方差膨胀因子的相关数据这里不作展示.在图 3 中,按照变量的重要程度进行排序,发现对上证指数影响较大的特征变量分别为 adx_adx、macd_signal、cci_20、emv_maemv、obv_、atr_atr、ex_rate、r_f.格力电器(图 3)亦按照变量的重要性由小到大的顺序进行排列,发现影响较大的特征变量分别为 pe_trailing、atr_atr、obv_、ex_rate、dpo_20、r_f.表 1 是变量的含义及描述性统计[34].对特征变量绘制变量的依赖关系,如图 4 所示.

从图 4 中可以看出反映超买超卖的 cci_20 以及趋势型指标 macd_signal 两个指标和上证指数 SSEC 大致呈现出正向相关的态势,而反映趋势强度的 adx_adx 指标与上证指数之间的关系大致是负向,其中,反映成交量与人气的 emv_maemv 指标和上证指数 SSEC 之间的关系存在显著的非线性关系.投资者在进行投资时可以着重关注这些影响较大的技术性指标.与此同时,在图 5 中也能看出反映公司价值的基本面指标 pe_trailing 与格力电器 Gree 的股价是正向的,因此投资者在进行投资时,不应该过于追求安全性,仅选市盈率较低的股票进行投资.反映震荡幅度的指标 atr_atr、人气型指标 obv_及汇率 ex_rate 与格力电器 Gree 的股价之间的关系比较复杂,在一定程度上反映了市场上不同投资者的态度以及产品的销售状况.

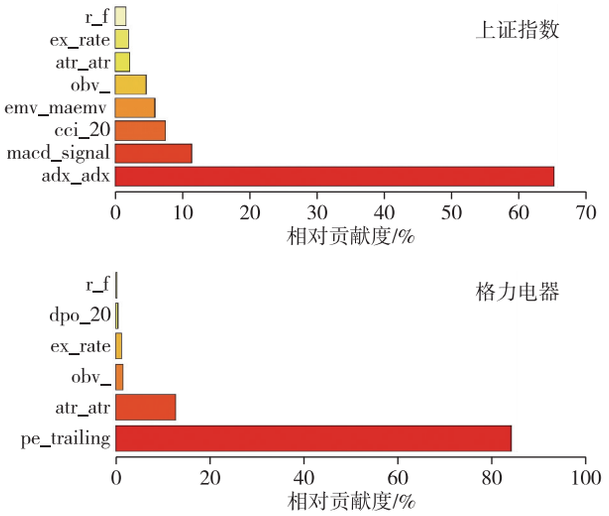


图 3 上证指数及格力电器重要特征变量

Fig. 3 Key characteristic variables of SSE index and Gree Electric Appliances

从表 1 能够看出:1)平滑异同平均(macd_signal)、简易波动指标(emv_maemv)、区间震荡线(dpo_20)与人气指标(obv_)、真实震幅(atr_atr)、上证指数(SSEC)在量纲上差距较大,后续在处理特征变量和响应变量时将进行归一化处理;2)数据都存在一定程度的左偏或者右偏,其中,上证指数相对正态分布呈现出“尖峰厚尾”的特征;3)量纲接近的指标中平均趋向指标(adx_adx)以及真实振幅(atr_atr)两个指标的方差较大,数据较为分散.

2.3 评价指标

本文采用如下的评估指标来评价模型预测效果:

1)均方根误差(RMSE):

$$e_{\mathrm{RMSE}}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}e_i^2}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i-\hat{X}_i)^2}. \quad (13)$$

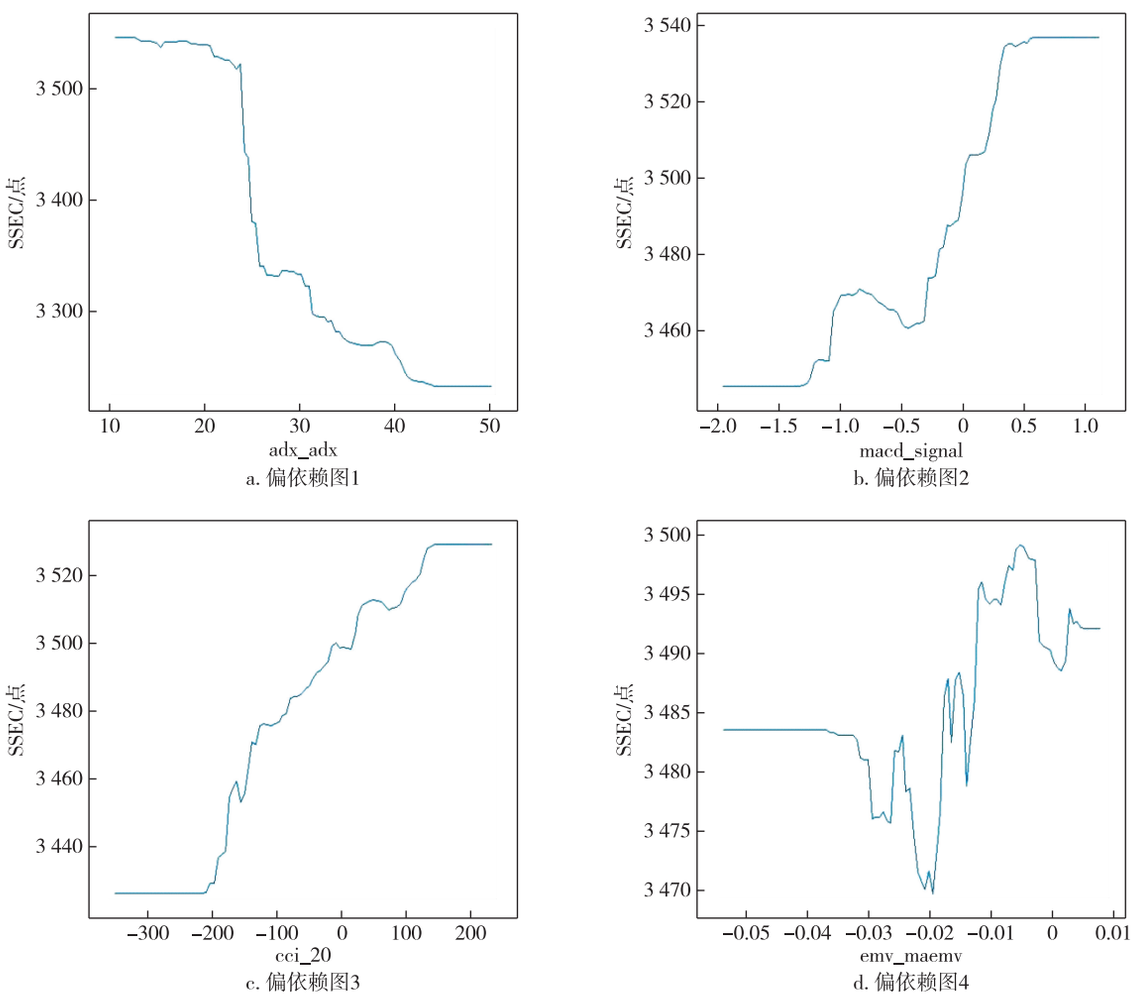
2)平均绝对百分比误差(MAPE):

$$e_{\mathrm{MAPE}}=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{e_i}{X_i}\right|\times 100\%=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{X_i-\hat{X}_i}{X_i}\right|\times 100\%. \quad (14)$$

表 1　重要特征变量以及响应变量的描述性统计[34]

Table 1　Descriptive statistics of key characteristic variables and response variables

特征名称	描述	均值	方差	最小值	最大值	偏度	峰度
adx_adx	平均趋向指标	20.60	105.58	10.61	50.18	1.57	1.22
macd_signal	平滑异同平均	−0.32	0.52	−1.95	1.13	−0.25	−0.36
cci_20	商品路径指标	−29.34	13 060.10	−350.82	230.10	−0.21	0.20
emv_maemv	简易波动指标	−0.01	0.00	−0.06	0.01	−1.55	1.94
obv_	人气指标	1.70e+9	1.50e+18	−9.57e+8	4.20e+9	−0.17	−0.86
atr_atr	真实震幅	44.39	66.02	33.51	67.45	0.86	0.18
ex_rate	人民币兑美元汇率	6.40	0.00	6.30	6.50	0.29	−1.25
r_f	上海同业拆放利率/%	1.94	0.058	1.20	2.30	−1.07	0.90
pe_trailing	历史市盈率	9.77	1.53	7.69	12.84	0.62	−0.17
dpo_20	区间震荡线	0.14	0.81	−1.98	2.67	0.16	−0.30
SSEC	上证指数收盘价/点	3488	21661.22	2929	3715	−1.19	7.69
Gree	格力电器收盘价/元	38.70	25.92	30.71	51.48	0.78	−0.14



a. 偏依赖图1　b. 偏依赖图2　c. 偏依赖图3　d. 偏依赖图4

图 4　上证指数重要特征变量偏相关依赖图

Fig. 4　Partial correlation dependence of key characteristic variables for SSE Index

2.4　文本情绪指数计算

东方财富网股吧(https://guba.eastmoney.com/)是我国最大的股市投资者交流贴吧,本文选择其中的帖子来计算情绪指数.对爬取的文本进行

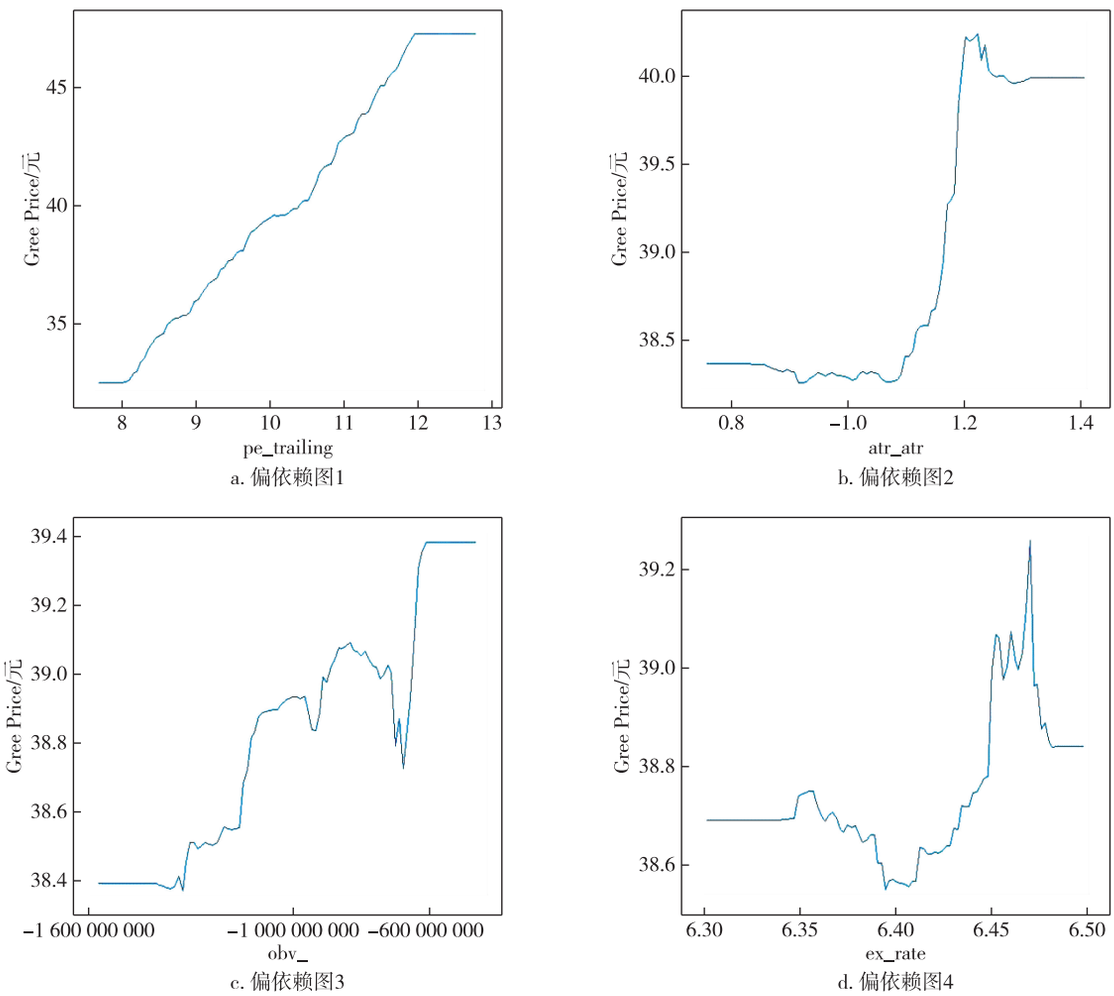


图 5　格力电器重要特征变量偏相关依赖图

Fig. 5　Partial correlation dependence of key characteristic variables for Gree Electric Appliances

预处理,去除一些没有意义的图片、数字及标点符号,通过整理每天的帖子计算每日的情绪指数.

常见的文本情绪指数计算有两种方法:一种是机器学习的方法,即先对文本进行分类选出积极消极的文本,通过支持向量机、朴素贝叶斯等模型进行训练,然后用训练完的模型进行应用,计算每天的情绪指数;另一种是运用词典的方法进行判定,构建情感词典,运用构建的情感词典筛选出每日的积极情感词和消极的情感词[35],其关键是情感词典的构建,构建一个详尽的中文金融情感词典十分重要.本文采用前一种方法,通过搜集一些已有文本,再加上作者标注的文本,构建积极和消极的语料并使用朴素贝叶斯进行分类,部分打标签的语料如表 2 所示.

本文抓取的上证指数以及格力电器的帖子时间跨度为 2021 年 7 月 12 日至 2022 年 4 月 25 日,除去其中没有交易的天数,共计 191 天,上证指数和格力电器帖子的条数分别为 1 万余条及 12 万余条,部分数据内容及打分如表 3 所示,效果较好,优于一般的情感词典法(情绪指数范围为 0~1).

表 2　部分语料归类

Table 2　Examples of corpus classification

序号	描述	分类类型
1	基本面太差劲摇摇欲坠卖	消极语料
2	抄底啊兄弟	积极语料

通过计算整理得出每日的情绪指数.由于量纲的差异,先对变量进行标准化处理,并计算文本情绪指数 em 与响应变量(收盘价)之间的灰色关联度[36],经计算,得出上证指数与格力电器文本情绪指数与各自收盘价的灰色关联度分别为 0.71 和 0.70,有较大关联性.股价关注度也是投资者情绪的一种体现[37],因此计算了投资者的关注度 att,计算

表 3　部分帖子内容

Table 3　Content of some posts

序号	描述	打分
1	垃圾格力.两年没到就不制冷了,祝你 st	0.019
2	明天格力冲击涨停,拭目以待	0.999
3	明天开始掉头向上,我已满仓格力,准备吃肉[抄底]	0.992
4	卖给小米吧,既能丰富小米的生态,格力也可以借小米品牌提高销量	0.560

方式见式(15)[38],其中,AbbrSVL 表示股票简称搜索量,CodeSVL 表示股票代码搜索量.发现投资者关注度 att 与对应的收盘价也有较强的关联,因此将投资者情绪 em 以及投资者关注度 att 共同作为特征变量.

$$att=\ln(AbbrSVL+CodeSVL). \tag{15}$$

2.5　LSTM 模型参数设置

使用多变量 LSTM 模型在深度学习平台 Tenorflow 上搭建神经网络.构建 3 层的神经网络,其中 2 层为隐藏层,第 3 层为输出层,第 1 层包含 80 个神经元,第 2 层包含了 100 个神经元,用表 1 中选择的变量作为特征变量,使用默认的学习率 0.01,使用 Adam 优化器,选择均方误差作为损失函数,迭代次数 epoch 以及每次喂入的数据 batchsize 分别为 50 和 64.

2.6　误差修正

为进一步减小模型的误差,首先选出误差较小的基准模型,再使用滚动 GM(1,1)模型进行修正,用前 7 天的误差预测第 8 天的误差,充分挖掘残差项的信息.对于其中出现的负的残差项,首先对数据加上一定的正数进行建模,再对残差进行相应的预测,预测出对应的数据后再减去原来加上的正数还原,得出最终预测的残差项.

3　实证分析

3.1　股价时间序列预测

本文选择格力电器的收盘价作为响应变量进行预测,因为格力电器作为 A 股市场的白马股具有一定的代表性.从 191 天的数据中,选取样本中前 140 天的数据作为训练集,后 51 天的数据作为测试集.采用滚动预测的方法,用前 7 天特征变量的数据预测第 8 天的标签,即收盘价的数据.根据前面的工作,选择的特征变量分别为 pre_trailing、atr_atr、obv_、ex_rate、dpo_20、r_f 以及后面加入的情绪相关变量 att、em 及其自身的收盘价共计 9 个特征变量.具体的预测结果如图 6 所示.为了说明模型的稳健性,本文加上了同为电器行业的飞科电器(603868)以及美的集团(000333)的股价,采用同样的方法在同一时间段内进行预测,预测结果分别如图 7 和图 8 所

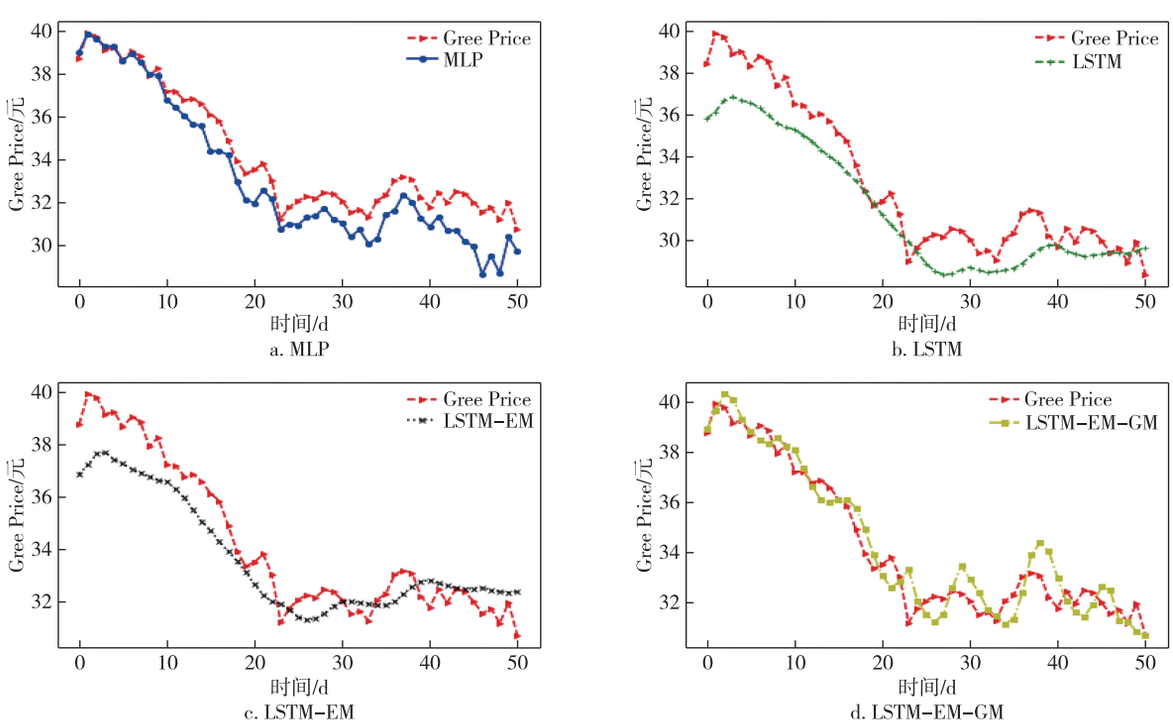


图 6　不同模型格力电器股价预测比较

Fig. 6　Comparison of Gree Electric Appliances stock prices forecasted by different models

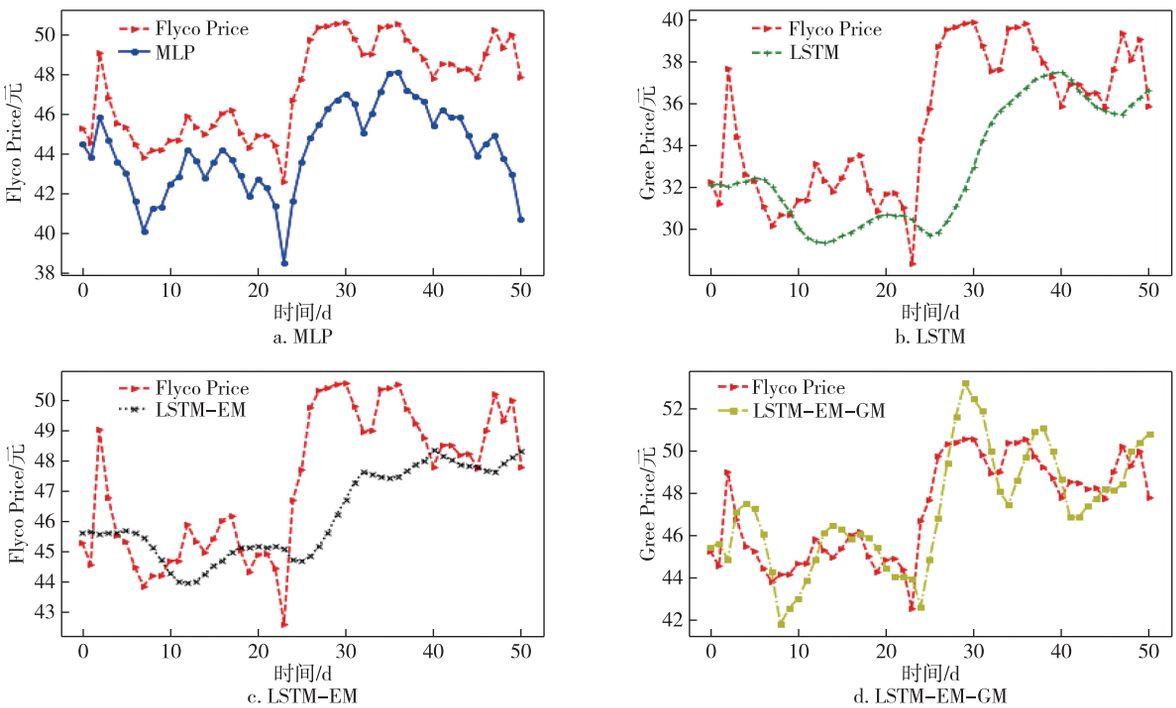


图 7　不同模型飞科电器股价预测比较

Fig. 7　Comparison of Flyco Electric Appliance stock prices forecasted by different models

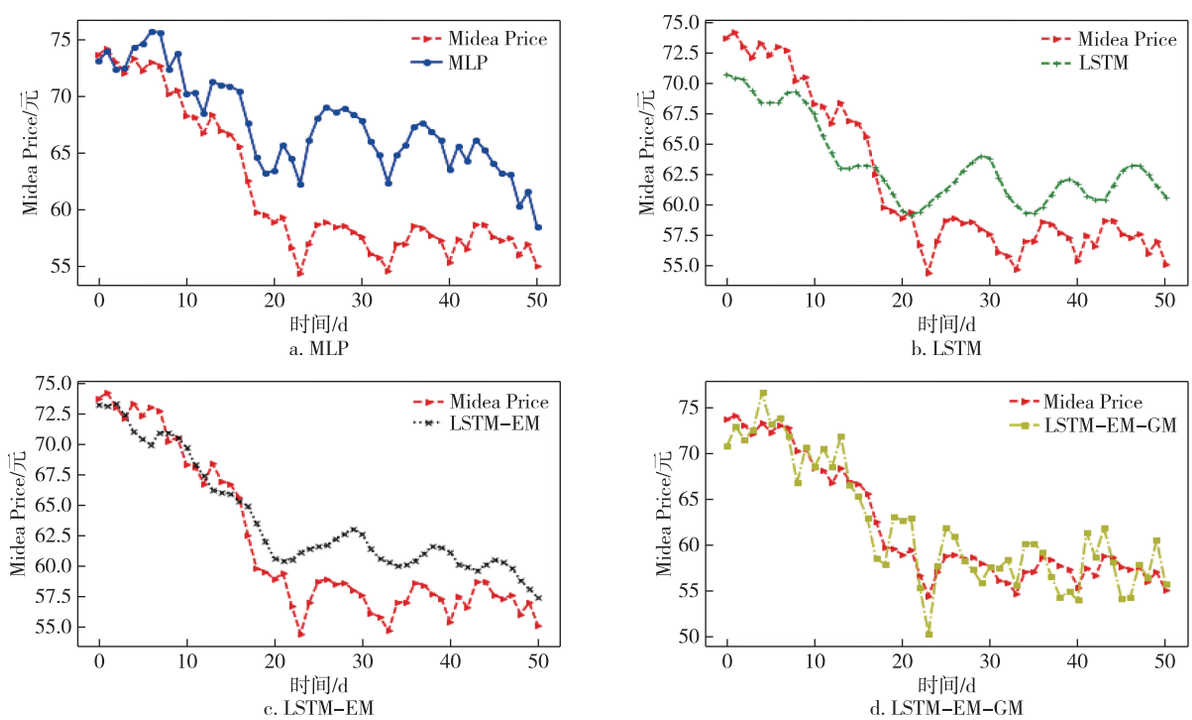


图 8　不同模型美的集团股价预测比较

Fig. 8　Comparison of Midea Group stock prices forecasted by different models

示,其中 Gree、Flyco、Midea 分别表示格力电器、飞科电器、美的集团三支股票.

本文首先选择的是 MLP(Multilayer Perceptron)基础神经网络[39]与 LSTM 模型进行比较,发现即使没有添加情绪相关的指标,LSTM 的预测效果也相对较好;接着分别对比加入情绪指数、投资者关注度的

模型 LSTM-EM 以及在此基础上对误差修正的模型 LSTM-EM-GM,具体评测指标的数值如表 4 所示.

表 4　不同模型预测结果对比

Table 4　Comparison of prediction results of different models

模型	样本	MAPE/%	RMSE
MLP	Gree	3.265	1.217
	Flyco	6.527	3.386
	Midea	9.840	6.572
LSTM	Gree	3.124	1.278
	Flyco	3.406	2.289
	Midea	4.201	2.991
LSTM-EM	Gree	3.125	1.130
	Flyco	3.007	1.949
	Midea	5.855	3.996
LSTM-EM-GM	Gree	1.862	0.770
	Flyco	2.900	1.662
	Midea	2.950	2.171

从表 4 中能够看出:在格力电器(Gree)的股价预测中,没有加入情绪指数的 LSTM 模型的 MAPE、RMSE 分别为 3.124%、1.278,加入了情绪指数的 MLP 模型的 MAPE、RMSE 分别为 3.265%、1.217.加入情绪指数 em、投资者关注度 att 的模型 LSTM-EM 模型的 MAPE 以及 RMSE 分别为 3.125%、1.130,相较于 MLP 模型,无论是 MAPE 还是 RMSE 指标都有了显著的提高,因此选用 LSTM 模型为基准模型.而相对于 LSTM 模型,LSTM-EM 模型虽然 MAPE 没有显著变化,但 RMSE 有了较大下降.进一步,使用 GM(1,1)模型对 LSTM-EM 进行修正后的模型 LSTM-EM-GM 模型的 MAPE 及 RMSE 分别为 1.862%、0.770,都相对之前的模型有了更为显著的下降,为最优模型.在飞科电器(Flyco)及美的集团(Midea)的股价预测中,LSTM-EM-GM 也是最优的预测模型,反映其预测误差大小的指标中 MAPE 分别为 2.900%及 2.950%,RMSE 指标的值分别为 1.662 及 2.171,相对于对比模型 MLP 以及原始模型 LSTM,误差都有所下降.

从上述的对比中能够发现,相较于 MLP,具有记忆性的 LSTM 模型能够对股价数据进行更好的预测,且投资者情绪与投资者关注度与格力电器的股价之间有较为明显的关联,将其作为特征变量加入到模型中能在一定程度上提高模型的预测精度.同时,由于选用灰色 GM(1,1)模型,充分挖掘了残差项中的信息,使得模型的预测精确度有了较为显著的提高.结合所选取的电器行业的 3 个案例,发现模型对于波动剧烈且下降的格力电器股价、波动下降但降幅略小的美的集团以及波动上升的飞科电器股价,都能进行较好的预测,验证了模型的稳健性.

3.2　模型普适性研究

上述研究预测了近 2 个月的股票收盘价.为更好地说明模型的适用性,缩短预测时间,仅预测 2 天的数据量,选择上证指数 SSEC 作为响应变量进行验证.同样使用 LSTM-EM-GM 模型,除了迭代轮数 epoch 变为 100,滚动窗口设置为 5,其余参数同上.选择的特征变量分别为 adx_adx、macd_signal、cci_20、emv_maemv、obv_、atr_atr、ex_rate、r_f 以及投资者情绪指数 em、投资者关注度 att 和自身收盘价共计 11 个特征变量.具体预测结果如表 5 所示.由于数据量较小,这里不再绘图展示.

表 5　不同模型短期预测结果对比

Table 5　Comparison of short-term prediction results of different models

日期	实际值	MLP	LSTM	LSTM-EM	LSTM-EM-GM
20220422	3 086.919	3 032.605	3 102.059	3 095.854	3 077.771
20220425	2 928.512	3 060.837	3 074.692	3 066.633	3 012.445
MAPE/%		3.139	2.741	2.503	1.581
RMSE		101.143	103.918	97.870	59.701

从表 5 中能够看出,对于上证指数的预测,由于量纲的问题导致 RMSE 的指标都相对较大.总的来说 MLP 模型的效果依然相对较差,其 MAPE 及 RMSE 指标的数值分别为 3.139%及 101.143,LSTM-EM-GM 模型的效果依然最好,MAPE 及 RMSE 的数值分别为 1.581%及 59.701.LSTM-EM-GM 模型在更短期的预测中取得不错的表现,验证了模型的普适性.

4　结论

基于股票市场的技术性指标、基本面指标以及投资者情绪和投资者关注度对格力电器和上证指数进行了分析,并对变量筛选后的模型进行了残差项的修正.通过实证分析得出如下结论:

1)投资者在进行投资时除了关注市场行情、了解大盘指数,也可以关注与股指关联性较大的指标,如反映超买超卖的 cci_20、反映趋势的 macd_signal 以及 adx_adx 等指标;在投资个股时除了关注汇率、市盈率等基本面的指标,也应适当关注对各股影响

较大的技术性指标,如反映震幅的指标 atr_atr、人气型指标 obv_等.

2)情绪指数以及投资者关注度与股价之间存在较强的关系,将其作为特征变量能在一定程度上提高模型的预测精度,所以,在投资时应当时刻关注市场上的投资者情绪,适时操作.

3)通过对残差项进行修正,能显著地提高模型的预测效果,说明残差项中蕴含着丰富的信息,且使用 GM(1,1)模型对于没有明确分布的时间序列具有较好的特征提取作用.

本文虽然在特征变量选取方面及情绪指数的计算方面具有一定的科学性,考虑了可能存在的多重共线性以及各个特征变量的贡献程度,整理了相对完善的语料库而未使用词典法计算情绪指数,但也存在一定的不足之处:首先,影响股市的因素错综复杂,可能还有许多影响重大的因素未得到体现;其次,文本分析及情绪指数的计算还缺乏成熟的体系.这是后续研究值得完善的方面.

参考文献

References

[1] Li Z W,Han J,Song Y P.On the forecasting of high-frequency financial time series based on ARIMA model improved by deep learning[J]. Journal of Forecasting,2020,39(7):1081-1097

[2] Engle R F, Rangel J G. The spline-GARCH model for low-frequency volatility and its global macroeconomic causes[J]. The Review of Financial Studies, 2008, 21(3):1187-1222

[3] 陈成,丁皖婧.证券业行政监管的股价风险预测与监控[J].统计与决策,2019,35(5):159-163
CHEN Cheng,DING Wanjing.Stock price risk prediction and monitoring by administrative supervision of securities industry[J].Statistics & Decision,2019,35(5):159-163

[4] 覃思乾.股价预测的 GM(1,1)模型[J].统计与决策,2006(6):22-23
QIN Siqian.GM(1,1) model of stock price forecast[J]. Statistics & Decision,2006(6):22-23

[5] 许兴军,颜钢锋.基于 BP 神经网络的股价趋势分析[J].浙江金融,2011(11):57-59,64
XU Xingjun,YAN Gangfeng.Analysis of stock price trend based on BP neural network[J].Zhejiang Finance,2011(11):57-59,64

[6] 孙泉,赵旭峰,钱存华.基于多点加权马尔可夫链模型的股价预测分析[J].南京工业大学学报(自然科学版),2008,30(3):89-92
SUN Quan,ZHAO Xufeng,QIAN Cunhua.Prediction and analysis of stock price based on multi-objective weighted Markov chain[J].Journal of Nanjing University of Technology (Natural Science Edition),2008,30(3):89-92

[7] 谢心蕊,雷秀仁,赵岩.MI 和改进 PCA 的降维算法在股价预测中的应用[J].计算机工程与应用,2020,56(21):139-144.
XIE Xinrui, LEI Xiuren, ZHAO Yan. Application of mutual information and improved PCA dimensionality reduction algorithm in stock price forecasting[J].Computer Engineering and Applications,2020,56(21):139-144

[8] 胡聿文.基于优化 LSTM 模型的股票预测[J].计算机科学,2021,48(增刊 1):151-157
HU Yuwen. Stock forecast based on optimized LSTM model[J].Computer Science,2021,48(sup1):151-157

[9] 张宁致,周佳丽,孙武军.基于优化回声状态网络的股价预测研究[J].管理工程学报,2014,28(1):94-101
ZHANG Ningzhi, ZHOU Jiali, SUN Wujun. Stock index prediction based on optimized echo state network[J]. Journal of Industrial Engineering and Engineering Management,2014,28(1):94-101

[10] 綦方中,林少倩,俞婷婷.基于 PCA 和 IFOA-BP 神经网络的股价预测模型[J].计算机应用与软件,2020,37(1):116-121,156
QI Fangzhong, LIN Shaoqian, YU Tingting. A prediction model of stock price based on PCA and IFOA-BP neural network[J]. Computer Applications and Software, 2020, 37(1):116-121,156

[11] 尹湘锋,崔浩锋,文雪婷.基于两类核函数的 TSVR 在股价预测中的比较[J].统计与决策,2021,37(12):43-46
YIN Xiangfeng, CUI Haofeng, WEN Xueting. Comparison of TSVR based on two kinds of kernel functions in stock price forecasting[J]. Statistics & Decision, 2021, 37(12):43-46

[12] 张立军,苑迪.基于 GA-Elman 动态回归神经网络的股价预测模型研究[J].华东经济管理,2008,22(9):79-82
ZHANG Lijun, YUAN Di. Stock market forecasting research based on GA-Elman neural network[J]. East China Economic Management,2008,22(9):79-82

[13] 肖祎平,刘新卫,张威.基于非负权重最优组合预测的股价预测研究[J].统计与决策,2013(18):142-145
XIAO Yiping, LIU Xinwei, ZHANG Wei. Research on stock price forecasting based on non-negative weight optimal combination forecasting[J]. Statistics & Decision, 2013(18):142-145

[14] 方义秋,卢壮,葛君伟.联合 RMSE 损失 LSTM-CNN 模型的股价预测[J].计算机工程与应用,2022,58(9):294-302
FANG Yiqiu, LU Zhuang, GE Junwei. Forecasting stock prices with combined RMSE loss LSTM-CNN model[J]. Computer Engineering and Applications, 2022, 58(9):294-302

[15] 甘昕艳,张钰玲,潘家英.基于股价指数预测的仿真研究[J].计算机仿真,2010,27(10):297-300
GAN Xinyan, ZHANG Yuling, PAN Jiaying. Study on simulation of stock price index forecasting[J]. Computer Simulation,2010,27(10):297-300

[16] 张东祥,刘英顺.盈余、资产质量与银行股价预测力:基于中国 16 家上市银行的实证检验[J].金融论坛,2017,22(1):56-66,80
ZHANG Dongxiang,LIU Yingshun.Earnings,asset quality

and the ability to predict the prices of bank stocks: an empirical test based on Chinese 16 listed banks[J]. Finance Forum,2017,22(1):56-66,80

[17] 林志勇,张维强,徐晨.基于小波变换与 MOBP 的股价预测[J].计算机工程与应用,2008,44(16):215-217
LIN Zhiyong,ZHANG Weiqiang,XU Chen.Forecasting of stoke price based on wavelet transform and MOBP[J]. Computer Engineering and Applications,2008,44(16):215-217

[18] 戴稳胜,吕奇杰,David Pitt.金融时间序列预测模型:基于离散小波分解与支持向量回归的研究[J].统计与决策,2007(14):4-7
DAI Wensheng, LÜ Qijie, David Pitt. Financial time series forecasting model: research based on discrete wavelet decomposition and support vector regression[J]. Statistics and Decision,2007(14):4-7

[19] 刘铭,单玉莹.基于 EMD-LSTM 模型的股指收盘价预测[J].重庆理工大学学报(自然科学),2021,35(12):269-276
LIU Ming, SHAN Yuying. Prediction of closing price of stock index based on EMD-LSTM model[J]. Journal of Chongqing University of Technology (Natural Science), 2021,35(12):269-276

[20] 张冰,王传美,贺素香.改进的 TSVR 模型在股市高频数据上的预测[J].计算机工程与设计,2019,40(11):3241-3246
ZHANG Bing,WANG Chuanmei,HE Suxiang.Prediction of improved TSVR model on high frequency stock market [J].Computer Engineering and Design,2019,40(11):3241-3246

[21] Krollner B, Vanstone B J, Finnie G R. Financial time series forecasting with machine learning techniques: a survey [C]//18th European Symposium on Artificial Neural Networks. April 28-30, 2010, Bruges, Belgium. DBLP,2010:25-30

[22] Shah D, Campbell W, Zulkernine F H. A comparative study of LSTM and DNN for stock market forecasting [C]//2018 IEEE International Conference on Big Data (Big Data). December 10-13,2018, Seattle, WA, USA. IEEE,2019:4148-4155

[23] 曹超凡,罗泽南,谢佳鑫,等.MDT-CNN-LSTM 模型的股价预测研究[J].计算机工程与应用,2022,58(5):280-286
CAO Chaofan,LUO Zenan,XIE Jiaxin,et al.Stock price prediction based on MDT-CNN-LSTM model [J]. Computer Engineering and Applications,2022,58(5):280-286

[24] 赵红蕊,薛雷.基于 LSTM-CNN-CBAM 模型的股票预测研究[J].计算机工程与应用,2021,57(3):203-207
ZHAO Hongrui,XUE Lei.Research on stock forecasting based on LSTM-CNN-CBAM model[J].Computer Engineering and Applications,2021,57(3):203-207

[25] 卫强,赵羡,张遵强,等.基于投资者关注的股价走势预测与交易策略设计:股票间交叉模式视角[J].系统工程理论与实践,2016,36(6):1361-1371
WEI Qiang,ZHAO Xian,ZHANG Zunqiang,et al.Stock price prediction and trading strategy design based on investors' attention: cross patterns perspective between stocks[J]. Systems Engineering—Theory & Practice, 2016,36(6):1361-1371

[26] 陈晓红,彭宛露,田美玉.基于投资者情绪的股票价格及成交量预测研究[J].系统科学与数学,2016,36(12):2294-2306
CHEN Xiaohong, PENG Wanlu, TIAN Meiyu. Stock market prediction based on investor sentiment[J].Journal of Systems Science and Mathematical Sciences,2016,36(12):2294-2306

[27] Da Z, Engelberg J, Gao P J. The sum of all FEARS investor sentiment and asset prices[J].The Review of Financial Studies,2015,28(1):1-32

[28] 许雪晨,田侃.一种基于金融文本情感分析的股票指数预测新方法[J].数量经济技术经济研究,2021,38(12):124-145
XU Xuechen,TIAN Kan.A novel financial text sentiment analysis-based approach for stock index prediction[J]. The Journal of Quantitative & Technical Economics, 2021,38(12):124-145

[29] 张梦吉,杜婉钰,郑楠.引入新闻短文本的个股走势预测模型[J].数据分析与知识发现,2019,3(5):11-18
ZHANG Mengji, DU Wanyu, ZHENG Nan. Predicting stock trends based on news events[J].Data Analysis and Knowledge Discovery,2019,3(5):11-18

[30] 黄润鹏,左文明,毕凌燕.基于微博情绪信息的股票市场预测[J].管理工程学报,2015,29(1):47-52,215
HUANG Runpeng,ZUO Wenming,BI Lingyan.Predicting the stock market based on microblog mood[J].Journal of Industrial Engineering and Engineering Management, 2015,29(1):47-52,215

[31] 饶东宁,邓福栋,蒋志华.基于多信息源的股价趋势预测[J].计算机科学,2017,44(10):193-202
RAO Dongning, DENG Fudong, JIANG Zhihua. Stock price movements prediction based on multisources[J]. Computer Science,2017,44(10):193-202

[32] 孙瑞奇.基于 LSTM 神经网络的美股股指价格趋势预测模型的研究[D].北京:首都经济贸易大学,2016
SUN Ruiqi.Research on price trend prediction model of U.S. stock index based on LSTM neural network[D]. Beijing: Capital University of Economics and Business,2016

[33] 刘思峰,曾波,刘解放,等.GM(1,1)模型的几种基本形式及其适用范围研究[J].系统工程与电子技术,2014,36(3):501-508
LIU Sifeng, ZENG Bo, LIU Jiefang, et al. Several basic models of GM(1,1) and their applicable bound[J].Systems Engineering and Electronics,2014,36(3):501-508

[34] 尹宏,胡红霞.股市技术分析实战全书[M].北京:经济管理出版社,2007

[35] 姜富伟,孟令超,唐国豪.媒体文本情绪与股票回报预测[J].经济学(季刊),2021,21(4):1323-1344
JIANG Fuwei, MENG Lingchao, TANG Guohao. Media textual sentiment and Chinese stock return predictability [J].China Economic Quarterly,2021,21(4):1323-1344

[36] 刘思峰.灰色系统理论及其应用[M].8 版.北京:科学出版社,2017

[37] Barber B M,Odean T.All that glitters:the effect of attention and news on the buying behavior of individual and

institutional investors [J]. The Review of Financial Studies,2008,21(2):785-818
[38] 俞庆进,张兵.投资者有限关注与股票收益:以百度指数作为关注度的一项实证研究[J].金融研究,2012(8):152-165
YU Qingjin, ZHANG Bing. Limited attention and stock performance:an empirical study using Baidu index as the proxy for investor attention[J].Journal of Financial Research,2012(8):152-165
[39] Wang Y L,Wang L P,Chang Q,et al.Effects of direct input-output connections on multilayer perceptron neural networks for time series prediction[J].Soft Computing,2020,24(7):4729-4738

Stock price time series prediction based on LSTM and grey model

HAN Jinlei[1] XIONG Pingping[1] SUN Jihong[2]
1 School of Management Science and Engineering,Nanjing University of Information Science & Technology,Nanjing 210044,China
2 Institute of Higher Education,China National Academy of Education Sciences,Beijing 100088,China

Abstract In view of the complicated factors influencing the stock price,we revised the Long Short-Term Memory (LSTM) network,which is commonly used in time series,to predict stock prices under the condition of multivariable.First,the Variance Inflation Factor (VIF) was used to screen variables,and then the adaptive promotion (Adaboost) model was combined to check the importance of characteristic variables.Second,the crawler was used to conduct text analysis of investor sentiment,calculate indicators including sentiment index,and reveal the relationship between them and stock price.Then,prices of three stocks including Gree Electric Appliances,Flyco Electric Appliances and Midea Group were predicted by Multilayer Perceptron (MLP) and LSTM,and the appropriate model was selected as the benchmark model. Finally,indicators of sentiment index and investor concern were added to the benchmark model to construct the LSTM-EM model,and the GM (1,1) model was used to correct the residual term after considering investor sentiment.The empirical results show that the proposed model can predict the stock price accurately.

Key words stock price forecast;comprehensive prediction;text analysis;error correction;long short-term memory (LSTM) network;grey model