

基于 Bert 股吧舆情分析的特征融合股价走势预测研究

李 梦^a, 黄政祺^b

重庆工商大学 a. 数学与统计学院; b. 经济社会应用统计重庆市重点实验室, 重庆 400067

摘 要:传统股价预测模型往往只考虑时序性数据且局限于模型自身机制,而忽略舆情对股价的影响,导致预测精度不高,针对该问题,提出基于 Bert 股吧舆情分析的特征融合预测模型对股价收盘价进行涨跌幅预测。首先,采用 Bert 自然语言处理对股吧舆情以及公司公告政策进行情感分类,并转化为虚拟变量,构建金融舆情情感特征库;然后将金融舆情特征库和时序性数据合并构建特征融合矩阵;最后输入长短期记忆网络模型(LSTM)进行股价收盘价预测,并得出股价的涨跌结果。以华银电力(600744.SH)为例进行实证分析,实验结果表明:引入股票情感特征后的模型,得到的股价走势准确率上升了 8.63%,预测收盘价的回归指标 F_{MAPE} 、 F_{RMSE} 分别下降了 23.59%、22.9%, R^2 提高了 8.11%,证明引入新的舆情情感特征在实际预测中能提高股价预测的准确率,可以作为精准预测股价走势的手段。

关键词:文本情感分析;股价预测;Bert 自然语言处理模型;长短期记忆网络模型

中图分类号:TP183 **文献标识码:**A **doi:**10.16055/j.issn.1672-058X.2023.0004.011

Feature Fusion for Stock Price Trend Prediction Based on Bert Stock Forum Opinion Analysis

LI Meng^a, HUANG Zhengqi^b

a. School of Mathematics and Statistics; b. Chongqing Key Laboratory of Economic and Social Applied Statistics, Chongqing Technology and Business University, Chongqing 400067, China

Abstract: Traditional stock price forecasting models often only consider time-series data and are limited to the model's own mechanism, ignoring the impact of public opinion on stock prices, resulting in poor forecasting accuracy. To solve this problem, a feature fusion prediction model based on Bert stock forum public opinion analysis was proposed to predict the trend of stock price. Firstly, Bert natural language processing was used to classify the sentiment of stock forum opinions and company's announcement policies, and the emotion after classification was converted into virtual variables to construct the emotion feature database of financial public opinions. Then, the feature fusion matrix was constructed by merging the financial public opinion feature database and the time-series data. Finally, a long short-term memory (LSTM) network model was used to predict the closing price of the stock and to derive the upward and downward results of the stock price. Taking Huayin Power (600744.SH) as an example for empirical analysis, the experimental results showed that after the introduction of stock sentiment features into the model, the accuracy of the stock price trend increased by 8.63%, the regression indicators of F_{MAPE} and F_{RMSE} for predicting the closing price decreased by 23.59% and 22.9%, respectively, and R^2 increased by 8.11%. This demonstrates that the introduction of new sentiment features of public

收稿日期:2022-10-24 **修回日期:**2022-12-08 **文章编号:**1672-058X(2023)04-0077-10

基金项目:重庆市科学技术委员会自然科学基金计划资助项目(CSTC2020JCYJ-MSXMX0162)。

作者简介:李梦(1973—),女,四川开江县人,副教授,博士,从事大数据分析技术及应用研究。

通讯作者:黄政祺(1998—),男,重庆九龙坡人,硕士研究生,从事数据挖掘、金融风险预测、自然语言处理研究。Email:760981047@qq.com.

引用格式:李梦,黄政祺.基于 Bert 股吧舆情分析的特征融合股价走势预测研究[J].重庆工商大学学报(自然科学版),2023,40(4):77—86.

LI Meng, HUANG Zhengqi. Feature fusion for stock price trend prediction based on bert stock forum opinion analysis[J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2023, 40(4): 77—86.

opinion can improve the accuracy of stock price prediction in practical forecasting and can be used as a means to accurately predict the trend of stock price.

Keywords: text sentiment analysis; stock prediction; Bert natural language processing model; long and short-term memory network model

1 引言

股票价格作为反映金融市场的风向标,具有高噪声、高风险、波动频繁等特点,因此想要精准预测股价的难度比较大。股价预测的传统模型只考虑股票内在的时序性数据,没有考虑外界因素的影响,导致股价预测精度很难得到提升。传统的股票价格预测方法,主要采用的方法为以下三类。

基于历史数据的时间序列模型。时间序列模型最主要的是差分自回归移动平均模型(ARIMA),且股价是时序性数据,能够很好地适用于时间序列预测。但时间序列模型必须要满足平稳性以及数据具有自相关性才能进行建模,而我国投资者众多,股市交易频繁,股价容易出现较大幅度变动,因此仅凭股票的历史数据无法准确预测股价走势。为提高预测精度,陈博闻^[1]以布林带与 MACD 为技术指标,使用 ARIMA 模型对中国平安的收盘价进行短期预测;尚未平等^[2]提出基于指数平滑法的 ARIMA-LS-SVM 组合定价模型预测股价,将时间序列模型与机器学习相结合,以此提高预测精度。

基于机器学习的股价预测方法。由于机器学习在非线性问题中表现出色,算法分为回归、分类、聚类,用于股价预测的主要是分类算法。王燕等^[3]基于网格搜索算法对 XGBoost 模型进行参数优化,用于股票的短期走势预测;李泽远^[4]基于 LightGBM 与卷积神经网络在贷款违约风险预测的研究,根据客户基本信息与贷款信息,构建了贷款违约分类模型预测贷款违约概率。机器学习虽然能够提取非线性特征,但参数设置非常不灵活,且当数据维度较多时容易产生过拟合现象。

基于深度学习的股价预测方法。近年来,深度学习在股票预测中表现出色,它引入含有大量隐藏层的神经网络来缓解过拟合现象,越来越多的学者尝试用深度学习思想解决金融市场的预测问题。李腾^[5]基于 BP 神经网络和正交小波变换去噪相结合进行股价预测研究;肖国荣^[6]依赖 BP 神经网络的非线性逼近能力和自学习能力对金融序列价格进行预测,指出利用自适应学习速率动量梯度下降反向传播算法,动态调整

BP 神经网络的学习速率,能够较好地预测股价。但 RNN 以及 BP 神经网络处理时序数据时会面临着梯度爆炸,陷入局部最优解的问题。为解决该问题,长短期记忆网络模型(LSTM)作为 RNN 的改进被提出,并且在股价预测领域中优于传统的机器学习以及神经网络。

从时间序列模型—机器学习模型—深度学习模型的变换,学者们将其运用于股价特征的预测,并不断取得成果,但股价本身具有不确定性以及高噪声等特征,且大多数都是基于模型的改进,未考虑其他影响特征因素。事实上,股票价格的走势受许多非线性因素影响,如国家政策宏观经济、突发事件、舆情因素、国际金融环境等会间接作用于投资者的行为,产生情绪效应,从而影响股票价格走势^[7]。如徐月梅等^[8]从宏观角度考虑新闻对股价的影响,基于 CNN-BiLSTM 对新闻进行划分并进行情感分类,再使用 LSTM 进行股价预测,但没有考虑到个人因素对于股价的影响。由于我国的金融市场规模庞大,有很多投资散客,如果不考虑舆情因素,会忽略股民与股价之间的关联性,间接降低股民对市场的信心。为此,本文考虑将非线性的舆情情感特征因素融合股票时序性数据来预测股价的走势。

首先,确定个股,利用爬虫提取东方财富网中的股吧文本信息以及公司公告政策,再利用 Bert(Bidirectional encoder representations from transformer)模型进行情感分类并打上标签,构建情感特征矩阵;然后结合股票的数值数据构建特征融合矩阵;最后使用 LSTM 网络对特征融合矩阵进行建模预测。模型的优点在于:第一,构建了基于 Bert 的金融舆情情感特征库,并通过实证分析证明引入情感特征能够提高股价预测走势的准确率;第二,现有研究多为直接人为设定参数值,本文探究了不同参数对模型预测的影响,并通过迭代参数法找到了最优参数。

2 基于股吧舆情分析的股价预测模型

基于 Bert 情感分析和 LSTM 网络模型,提出融合金融情感特征和数值特征的股票价格走势预测流程图,其结构框架如图 1 所示。

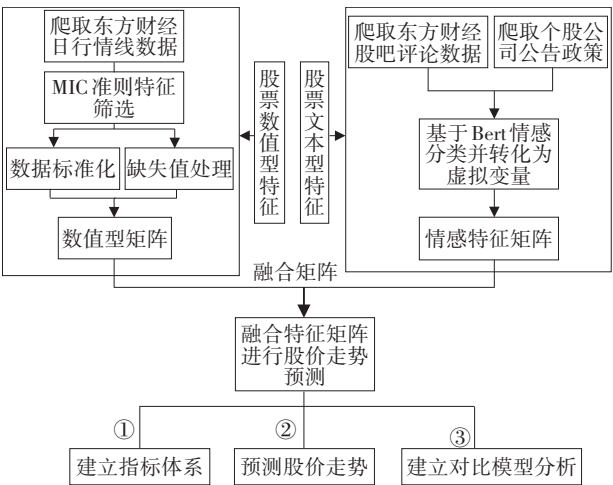


图 1 融合情感特征与数值特征的股票价格走势预测流程图

Fig. 1 Flow chart of stock price trend prediction combining emotional and numerical features

具体思路如下:对于爬取的文本型数据,首先进行人工标注;再使用 Bert 模型进行情感分类,Bert 输出的情感标签为数值型变量 0、1、2,情感类型分别为消极、中立、积极;最后将其转化为虚拟变量,构建情感特征库。对于爬取的股票数据,它的变量有多个,若将其全部输入 LSTM 网络层,会增加神经网络的负担,也会影响模型的效果,因此通过 MIC (MaxiMal Information

Coefficient)最大信息系数来对变量进行筛选,再对数据进行预处理。将情感特征矩阵与股票数值型数据融合,构建特征融合矩阵。使用 Keras 框架中的 LSTM 网络模型对特征融合矩阵进行建模预测,且使用迭代参数法寻找最优参数进行预测,并比较不同的模型,最后预测股票走势。

2.1 构建 Bert 金融舆情情感特征库

基于金融舆情信息与公告政策提出的 Bert^[8]模型结构图如图 2 所示,构建基于 Bert 股吧舆情的金融舆情情感特征库。

为分析舆情情感特征,将金融舆情信息以及公告政策,通过“Embedding”进行嵌入,转化为 $E_1, E_2, \dots, E_n$ 作为模型的文本输入向量, $T_1, T_2, \dots, T_n$ 为模型的输出向量。对于输入的舆情信息 $E_1, E_2, \dots, E_n$,首先给句子两端打上特殊符号“CLS”与“SEP”,类似于 word2vec 里单独给句子训练向量,但股吧里面的评论信息简短精炼,需要更精确地考虑句子之间的语义关联,因此引入 Transformer 双向编码器^[9]。Transformer 是一个 encoder-decoder 的结构,即“编码-解码”结构,由 Multi-Head Attention 和一个全连接组成,用于将舆情信息转化成特征向量,其中 Multi-Head Attention 层为最核心的部分^[10],如图 2 所示。

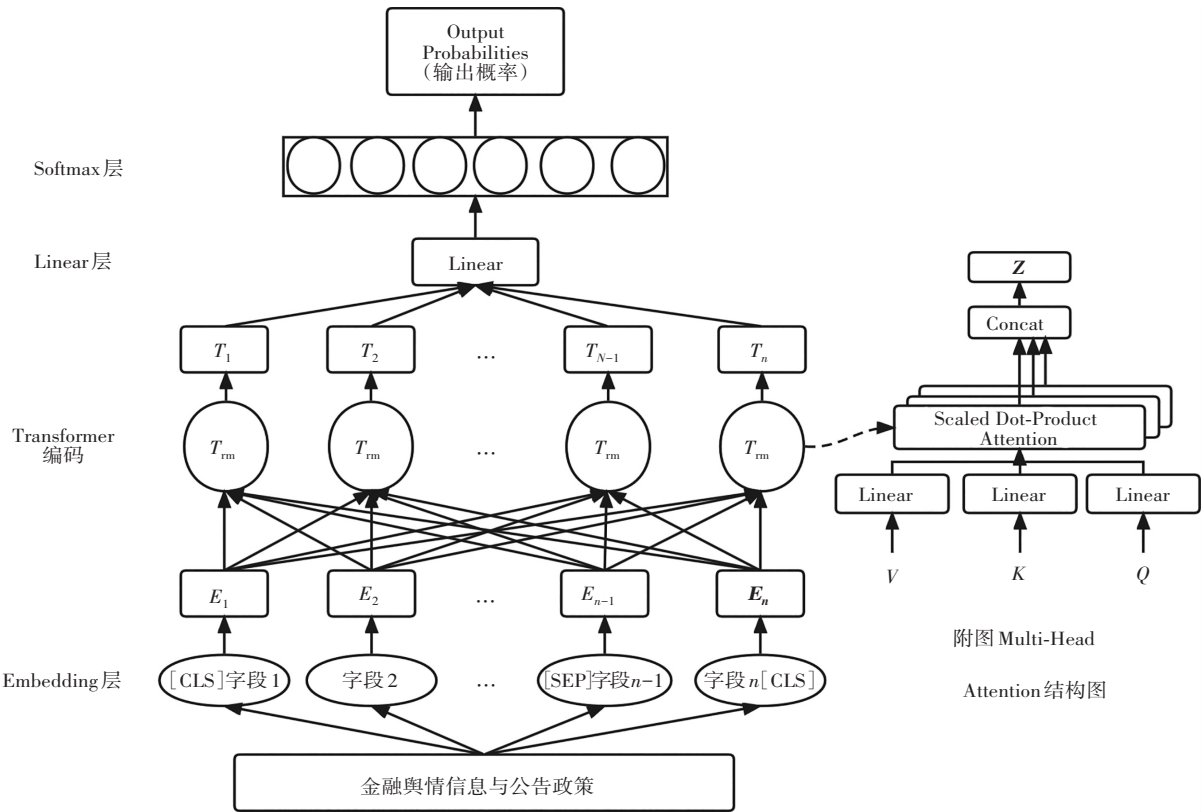


图 2 基于 Bert 的舆情情感分类流程图

Fig. 2 Flow chart of public opinion sentiment classification based on Bert

编码层提出了最为核心的机制即注意力机制(Self-Attention),在 Transformer 的 encoder 中,数据会经过注意力机制模块得到一个加权特征向量 Z ,这个 Z 即为 $Attention(Q, K, V)$:

$$Attention(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)$$

其中, d_k 表示向量的维度, Softmax 为归一化函数。该机制的计算过程:先将输入单词转化成嵌入式向量,根据嵌入式向量得到 q, k, v 等 3 个向量,计算每一个向量的 score 值。为避免梯度爆炸,Transformer 对 score 进行归一化处理,即除以 $\sqrt{d_k}$,再对 score 用 Softmax 函数激活,得到加权分数 v ,再得到结果 $Z = \sum v$,最后输出条件概率。

经过上述步骤,最后得到舆情信息向量在不同条件下(情感极性 $\{0, 1, 2\}$)的条件概率分布矩阵,其中 $T_1, T_2, \dots, T_n$ 为输出向量,表示在当前情感类别下的条件概率,具体如表 1 所示。

表 1 情感条件概率分布矩阵

Table 1 Emotion probability distribution matrix

	0	1	2	Sentiment
T_1	P_{10}	P_{11}	P_{12}	P_1
T_2	P_{20}	P_{21}	P_{22}	P_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
T_n	P_{n0}	P_{n1}	P_{n2}	P_n

得到情感条件概率分布矩阵后,根据对应概率最大的情感特征 P_i 生成情感分类结果,计算步骤如下:

$$P_i = \text{Max}(P_{i0}, P_{i1}, P_{i2}) \quad (1 \leq i \leq n) \quad (1)$$

将得到的分类结果转化为虚拟变量,再与金融舆情信息进行合并并同步,最后将情感特征矩阵与股票数值型数据融合,构成多变量的特征融合矩阵,作为 LSTM 网络的输入数据。

2.2 基于 LSTM 的股价走势预测

LSTM 作为 RNN 的变种,能够解决 RNN 的梯度弥散或梯度爆炸等问题,如 RNN 在信息间隔较远时,就容易丧失学习能力^[11],因此 LSTM 就被提出了^[12],即 LSTM 就是为了解决长期依赖问题而生的。LSTM 包含 3 个特殊的“门控单元”,如图 3 所示,分别是遗忘门(Forget Gate, f_t)、输入门(Input Gate, i_t)、输出门(Output Gate, o_t)。

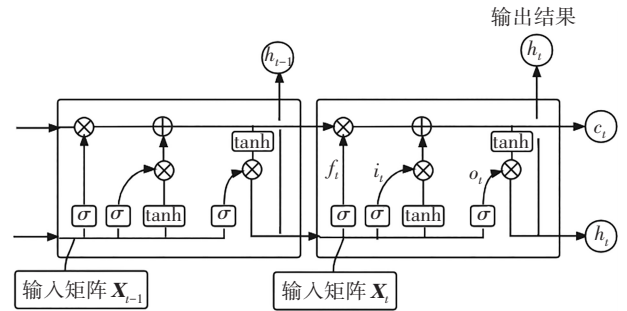


图 3 LSTM 股价预测网络结构

Fig. 3 Network structure of LSTM stock price prediction

其中的遗忘门和输入门是能够让 LSTM 长期依赖的关键所在,它能够控制信息的遗忘和更新。遗忘门会读取 h_{t-1} 与 X_t ,通过输入门层(Sigmoid)输出 $0 \sim 1$ 之间的数值给细胞中的 c_{t-1} 进行相乘,若为 1,则表示完全保留,若为 0 则表示完全舍弃。输入门也类似,首先通过 Sigmoid 层得到 i_t ,同时将 h_{t-1} 与 X_t 进行 tanh 处理,得到一个 $-1 \sim 1$ 的值,并将它和 i_t 相乘,得到输出部分。

X_t 为 t 时刻该细胞的输入值, h_t 为 t 时刻该细胞的状态输出, f_t 为遗忘门公式, i_t 为输入门公式, o_t 为输入门公式, $\hat{c}_t$ 为更新的输入信息, c_t, h_t 分别为长期和短期记忆传递模型, w_i, w_f, w_c, w_o 分别为输入门、遗忘门、更新门、输出门的权重矩阵, b_i, b_f, b_c, b_o 分别为其对应的偏置, σ 为激活函数,作用在各个门空单元上,产生 0 与 1 之间的值。激活函数有多种,如 Relu、Sigmoid 等,激活函数用在状态和输出上,将其输出在 $-1 \sim 1$ 之间,具体公式如式(2)一式(7)所示:

$$i_t = \sigma(w_i * [h_{t-1}, X_t] + b_i) \quad (2)$$

$$f_t = \sigma(w_f * [h_{t-1}, X_t] + b_f) \quad (3)$$

$$\hat{c}_t = \tanh(w_c * [h_{t-1}, X_t] + b_c) \quad (4)$$

$$o_t = \sigma(w_o * [h_{t-1}, X_t] + b_o) \quad (5)$$

$$c_t = f_t * c_{t-1} + i_t * \hat{c}_t \quad (6)$$

$$h_t = o_t * \tanh(c_t) \quad (7)$$

针对股票价格进行走势预测,具体流程:先爬取股吧舆情消息以及公告政策进行数据预处理,再使用 Bert 进行情感分类,然后生成虚拟变量与股票的数值型数据融合,最后使用 LSTM 网络进行股价预测。因 LSTM 网络层无法确定合适的参数,因此先根据经验设置初始参数,再通过迭代法寻找最优参数进行调参,使用最优参数进行股票价格走势预测,具体流程图如图 4 所示。

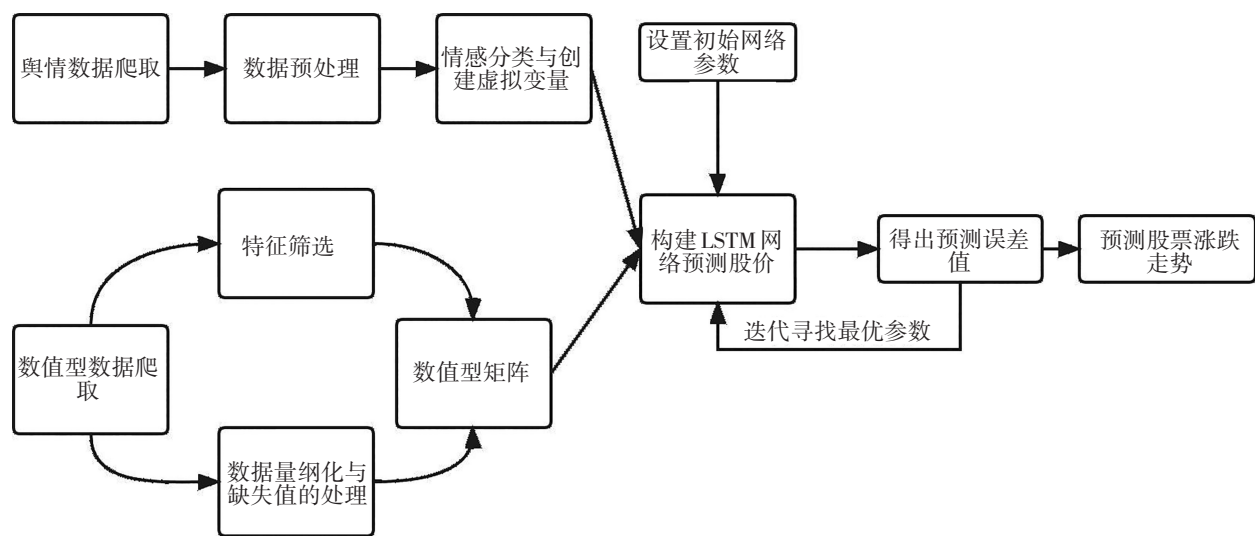


图 4 股价预测流程图

Fig. 4 Stock price forecast flow chart

3 数据准备

3.1 环境配置

本文基于真实数据进行实证分析,实验环境见表 2。

表 2 实验环境

Table 2 Experimental environment	
实验环境名称	参数
CPU	Intel(R) Core(TM) i7-10875H
CPU 核心线程数	10 核 16 线程
GPU	NVIDIA GeForce RTX 3070Ti
GPU 显存	8GB
软件	Python3. 8, Tensorflow1. 13

3.2 数据收集

“碳中和”政策出台后,许多与能源相关的企业龙头纷纷参与进来。能源一直以来都是股票交易市场的热点,相关的舆论以及话题热度都很高,而华银电力作为“碳排放”的龙头,具有较好的代表性。本文数据分为两类:第一类股票从东方财富网上爬取华银电力(600744.SH)2016-01-04—2022-03-14 共 1 491 个交易日的行情日线作为股票数值型数据;第二类文本数据通过东方财富网中的华银电力股吧爬取评论,共 12 645 条。

3.3 文本情感分类

3.3.1 文本数据预处理

爬取的部分舆情数据如表 3 所示。

表 3 部分原始数据

Table 3 Partial raw data

评 论	作 者	最后更新
最后悔买了这个股,一路绿到我发慌。套了 40%	我是超人 v	2022/3/14
电力三傻,还有哪个来着	我怕谁	
希望 2022. 12. 31 日之前,回本,注意,是只回本。	羽翼 dh	2022/3/13
⋮	股友麻麻滴神	2022/3/11
还是华银电力牛	⋮	⋮
2:30 暴力拉升	暴狂风干	2016/1/6
华银今天绝对高开! 点赞	Bruce0511	2016/1/5
	辉煌 1111	2016/1/4

考虑同一天内存在多个评论以及非股票交易日的评论,本文进行了数据筛选:将同一天的评论进行情感分类之后,再计算同一天内情感标签最多的情感类别作为当天的情感值输入。

3.3.2 建立情感标签库

基于爬取的文本数据,进行人工标记,建立预训练的金融舆情情感标签库,部分情感标签如表 4 所示。表格中的 comment 为评论内容,sent 为情感标签,0、1、2 分别代表消极、中性、积极。

表 4 金融舆情情感标签库

Table 4 Emotional tag library of financial opinions

comment	sentiment
\$ 华银电力(SH600744) \$ …30 个点,啥时能回本啊	1
下周看你表演	2
下周电力板块必涨	2
⋮	⋮
太垃圾了,说好的碳中和呢???	0
下午两点半高开! 冲	2
选错股了,黔源电力和你,选了你我真是无语	0
大家赶紧买,十倍牛股,砸锅卖铁买它,抓紧,抄底了	2

3.3.3 基于 Bert 的情感分类模型

针对标记好的数据集进行划分,用于训练模型,划分比例为 4:1。为了验证基于 Bert 的舆情情感分类器性能,将增加 XGBoost、Multinomial NB 模型进行对比,验证 Bert 在情感分类模型上的有效性。

XGBoost 作为决策树的衍生模型,使用精度更高的损失函数且增加并行结构,在分类任务上表现优异^[13]。Multinomial NB 基于原始贝叶斯理论基础之上,假设先验分布为多项式分布的朴素贝叶斯模型^[14],将文本特征通过 Word2vec 转化为词向量后进行预测,适用于离散型特征的分类,能够很好地适用于股吧评论数据的分类问题。将划分好的数据集进行情感分类并得出分类结果,如表 5 所示。

表 5 情感分类器准确率(Accuracy)对比

Table 5 Comparison of accuracy of emotion classifiers

模 型	Multinomial NB	XGBoost	Bert
训练集	79.09	70.12	86.21
测试集	77.64	68.11	87.09

由表 5 可以得出:基于 Bert 股吧评论情感分类器的准确率在训练集上达 86.21%,测试集上达 87.09%,优于 Multinomial NB 和 XGBoost,其中 XGBoost 的准确率明显低于 Bert,可能原因在于 XGBoost 进行情感分类时,未考虑词与词之间的关系,而 Bert 则是基于上下文进行双向特征提取,效果要明显高于 XGBoost。Bert 模型在测试集上的准确率高于训练集上的准确率,原因主要是 2021—2022 年 3 月之间的测试集比较完整且质量较高,而 2020 年及以前的文本数据有少许缺失,导致测试集的准确率反超训练集。需要注意的是,该股吧的评论内容集中在 2019—2022 年,2019 年之前的评论较少,即有一部分情感特征输出为空值,需要对其进行填充。填充的内容根据国泰安(CMSAR)数据库中的公告政策以及股吧舆情进行填补,然后将情感数据变更为虚拟数据,并且同步到当天的股票数值型数据中,建立金融舆情情感特征库。

3.4 股票财务数据预处理

3.4.1 特征融合以及筛选

基于舆情情感特征与数值型数据融合,对特征融合数据进行数值信息分析,采用 F_{MIC} (最大信息系数)对股价的特征变量进行筛选^[15]。它是属于 MINE(最大信息)的非参数性探索,用于计算变量之间的关联程度,特别是高维度数据,是一种优秀的数据关联性计算方式。互信息的概念公式如下式。

$$I(x;y) = \int p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)} dx dy$$

$$I[x;y] \approx I[X;Y] = \sum_{X,Y} P(X,Y) \log_2 \frac{p(X,Y)}{p(X)p(Y)}$$

其中: $p(x,y)$ 为变量 x 和 y 之间的联合概率,而 F_{MIC} 的思想是针对变量之间的互信息关系,将其用离散形式分布在二维空间中,所划分的二维空间在 x,y 方向分别划分区间,查看当前的散点在各个方格中的落入情况,并求联合概率,下面是 F_{MIC} 的计算公式:

$$F_{MICmic}(x;y) = \max_{a * b < B} \frac{I[x;y]}{\log_2 \min(a,b)}$$

a,b 分别为在 x,y 上划分的格子个数, B 是变量,本文规定 B 的大小为数据量的 0.6 次方。 F_{MIC} 的计算步骤:首先利用 Python 中的 MINE 库对 i,j 进行定阶;然后对数据的行 x,y 构成的散点图进行 i 行 j 列的网格化,求出最大互信息值并进行归一化;最后选择不同尺度(i 与 j 的组合)下互信息的最大值作为 F_{MIC} 值。融合后的股票特征进行 F_{MIC} 准则筛选后的前 10 名列表如表 6 所示。

表 6 股票最大信息相关性排序(前 10)

Table 6 Ranking of MIC correlation of stocks (top 10)

序 号	股票特征	F_{MIC}
1	收盘价	1
2	总市值	0.999
3	流通市值	0.999
4	最高价	0.981
5	最低价	0.969
6	开盘价	0.959
7	前收盘	0.946
8	换手率	0.644
9	sentiment	0.527
10	涨跌幅	0.432

3.4.2 量纲化与空值处理

基于 MIC 筛选出的 9 个变量,对其进行量纲化与空值处理,这些数据通过“Choice”金融客户端进行获取,“Choice”为东方财富网所附带的客户端,可以通过客户端查看各类高质量金融数据,能够为金融分析提供快速简洁的数据。对于股票数据而言,不存在异常值,所以只需要查看交易日有无空值。经查询,2016-04-01—04-22 有约 15 d 的缺失值,是由于公司发布非公开发行股票的议案,导致股票停牌,将这 15 d 的日行情线删除,然后对数据进行标准化处理,消除由于量纲不同带来的影响,同时也提升模型的精度与收敛速度^[16]。

设 $S_j = [s_{1j}, s_{2j}, s_{3j}, s_{nj}]$ 表示第 j 个特征向量在 n 天内组成的一维向量组, S_{ij} 表示第 j 个特征向量在第 i 天的具体数值,将所有的 S_{ij} 进行标准化:

$$S_{ij} = \frac{S_{ij} - \mu_j}{\sigma_j}$$

其中, σ_j 与 μ_j 分别代表特征向量 S_j 中的标准差与均值。
经过数据预处理后,对数据集进行划分,划分时采用不打乱顺序,划分比例为 9 : 1,测试集的时段为 2021-07-30—2022-03-14,训练集的时段为 2016-01-04—2021-07-29,数据集的统计信息如表 7 所示。

表 7 数据统计信息
Table 7 Data statistics

	训练集	测试集
数据集	1 339	149
时间段	2016/1/4—2021/7/29	2021/7/30—2022/3/14

4 实验结果分析和评价

4.1 初始参数设置

对于融合之后的数据,将其输入到 LSTM 网络层还需要设置一系列参数。由于个股类型的不同,无法直接找到最优解,本文根据搜寻到的资料提出迭代寻找最优参数方法^[17],具体如表 8。

表 8 LSTM 网络迭代参数设置
Table 8 LSTM network iteration parameter setting

参 数	LSTM
记忆天数(mem_days)	[3,5,7]
预测天数(pre_days)	1
LSTM 层数(lstm_layers)	[1,2,3]
Dense 层数(dense_layers)	[1,2,3]
神经元个数(units)	[16,32,64]
Dropout	[0.1,0.3,0.5]
Batch_size	32
epoch	50
Input_shape	(mem_days,15)
Optimizer	Adam
activation	Relu
loss	MSE
Metrics	Mape

记忆天数与预测天数的释义:根据前 mem_days 天来预测后 pre_days 天的 lable 值,LSTM 层数与 Dense 层则根据 Keras 官方文档进行调试,神经元个数根据数据所含特征个数暂定为 16、32、64,Droupout 率可以防止过拟合。Input_shape 的释义:将前 mem_days 天的融合数据打包一起输入 LSTM 网络层,其中 15 为融合数据特征个数。优化器使用 Adam,激活函数使用 Relu 为最佳,训练模型中的损失函数使用均方误差(R_{MSE}),评估标准使用平均绝对百分比误差(R_{Mape}),最后使用 R_{Mape}

最小值的模型参数为最优参数。

4.2 评价指标

本文使用精确率($P_{recision}$),准确率($A_{ccuracy}$),召回率(R_{ecall})以及 $F(\alpha)$ 为预测效果评价指标,精确率指被判定为正类的所有样本中,真实的正类样本所占比率;准确率为所有判断正确的样本比例;召回率为正类中,被判定为正类的比例, $F(\alpha)$ 为 $P_{recision}$ 与 R_{ecall} 的加权调和平均,一定程度上可以消除两者在大规模数据中的相互制约, $F(\alpha=1)$ 则是当 α 值为 1 时的特殊形式,具体计算公式如式(8)—式(12):

$$P_{recision \cdot i} = \frac{F_{TP \cdot i}}{F_{TP \cdot i} + F_{FP \cdot i}} \tag{8}$$

$$A_{accuracy \cdot i} = \frac{F_{TP \cdot i} + F_{TN \cdot i}}{F_{TP \cdot i} + F_{FP \cdot i} + F_{TN \cdot i} + F_{FN \cdot i}} \tag{9}$$

$$R_{ecall \cdot i} = \frac{F_{TP \cdot i}}{F_{TP \cdot i} + F_{FN \cdot i}} \tag{10}$$

$$F(\alpha) = \frac{(\alpha^2 + 1)P_{recision \cdot i} \times R_{ecall \cdot i}}{\alpha^2(P_{recision \cdot i} + R_{ecall \cdot i})} \tag{11}$$

$$F(\alpha=1) = \frac{2P_{recision \cdot i} \times R_{ecall \cdot i}}{P_{recision \cdot i} + R_{ecall \cdot i}} \tag{12}$$

针对股票预测模型而言, i 应当取值为 1, F_{TP} 为正确预测上涨的样本数, F_{TN} 为正确预测下跌的样本数, F_{FP} 为错误预测上涨的样本数, F_{FN} 为错误预测下跌的样本数。对于本文的股吧评论情感分类一共有 3 种情形,即 0(消极)、1(中立)、2(积极),那么 i 的取值为 1、2、3, $F_{TP \cdot i}$ 表示被正确分类到第 i 类的样本数量, $F_{FP \cdot i}$ 、 $F_{FN \cdot i}$ 、 $F_{TN \cdot i}$ 以此类推,仅是类别不同。

4.3 预测结果分析

4.3.1 分类指标对比

仿效情感分类的预测模型对比,本节股票预测模型也将引入其他算法来对比来说明引入情感特征之后的有效性。逻辑回归假设因变量 y 服从伯努利分布,它采用 Sigmoid 函数将线性问题变成非线性问题,通过极大似然函数,运用梯度下降法来求解参数,最后达到二分类问题,很适合预测股票涨跌的二分类问题^[18],因此本文引入逻辑回归与 LSTM 神经网络模型进行比较:仅引入行情线数据的 LR(Logistic Regression)模型、引入情感特征的逻辑回归模型、仅引入行情线数据的普通 LSTM 模型、引入情感特征的 LSTM 模型。

由于逻辑回归需要设定阈值,因此将涨跌幅设置为 1%~3% 的区间,使用 Python 确定最优阈值进行股价预测并作为实验对照组。对于 LSTM 模型,模型输出结果为预测后一天的收盘价,则将结果延后一天与当天的前收盘价进行比较,输出涨跌结果,再计算准确率,预测结果对比见表 9。

表 9 2021 年 7 月至 2022 年 3 月测试集各模型预测比较
Table 9 Comparison of model predictions in the test set
from July 2021 to March 2022

模 型	精确率/%	召回率/%	$F(\alpha=1)$ 值
LR	63.18	63.27	63.22
LR _{sentiment}	65.82	65.93	65.87
LSTM	68.14	66.29	67.20
LSTM _{sentiment}	72.32	70.81	71.56

由表 9 可看出:在模型对比上,LSTM 模型的预测效果明显好 LogisticRegression,主要原因还是因为 LSTM 模型的长期记忆能力以及在时序数列上的超模性,本文提出的 Bert_LSTM 模型取得了最佳效果。在情感特征对比上,对于 LogisticRegression 而言,引入了情感特征的模型 $F(\alpha=1)$ 值相对于未引入情感特征的 $F(\alpha=1)$ 值提高了 4.19%,对于 LSTM 而言,同样提升了 6.49%,均有不同程度的提高,证明了情感特征对于股价预测的有效性。

4.3.2 回归指标对比

预测模型的回归指标采用 F_{MAPE} (平均百分比误差)、 R^2 (可决系数)、 F_{RMSE} (均方误差)。公式如下:

$$F_{MAPE} = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

$$F_{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

$$R^2 = \frac{F_{SSR}}{F_{SST}}$$

公式中的 y_i 为收盘价实际值, $\hat{y}_i$ 代表收盘价预测值, F_{SSR} 为回归平方和, F_{SST} 为离差平方和。

LSTM 采用迭代参数法寻找最优参数来预测收盘价格,这里主要观察 Dropout 率与记忆天数(mem_days)给模型带来的影响。为了防止模型过拟合以及迭代时间过长,Dropout^[19] 被提出,它可以作为神经网络的选择器,在每一个 epoch 训练中忽略掉一部分特征检测器,可以明显减少过拟合现象。而记忆天数则由于数据的大小维度不同,不能直接确定最佳记忆天数,本文提出[3,5,7]3 种记忆天数进行迭代计算,对比如表 10 所示。

表 10 不同记忆天数与 Dropout 率对模型误差带来的影响 (F_{MAPE})

Table 10 Influence of different memory days and Dropout rates on model errors (F_{MAPE}) %

Dropout	Mem_days		
	3	5	7
0.1	2.84	2.72	3.01
0.2	3.38	3.25	3.87
0.3	4.86	4.58	5.02

由于股市交易具有时效性,日线的交易情况会跟

前几日的行情走势具有紧密联系,因此记忆天数对于预测效果非常重要,从表 10 大致可以得出 Dropout 率设置为 0.1,记忆天数设置为 5 是最佳选择。最优的迭代参数如表 11 所示。

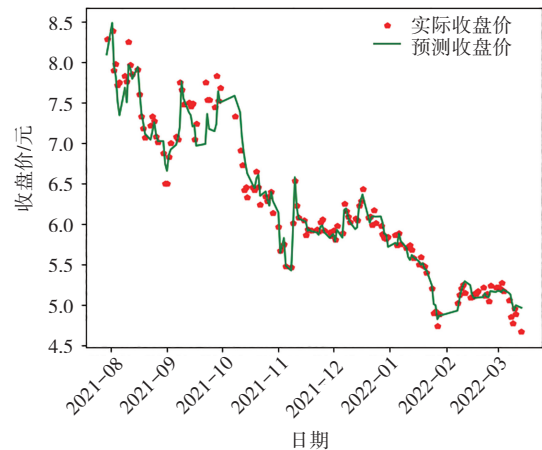
表 11 最优参数
Table 11 Optimal parameters

参 数	LSTM
记忆天数(mem_days)	5
LSTM 层数(lstm_layers)	1
Dense 层数(dense_layers)	2
神经元个数(units)	16
Dropout	0.1

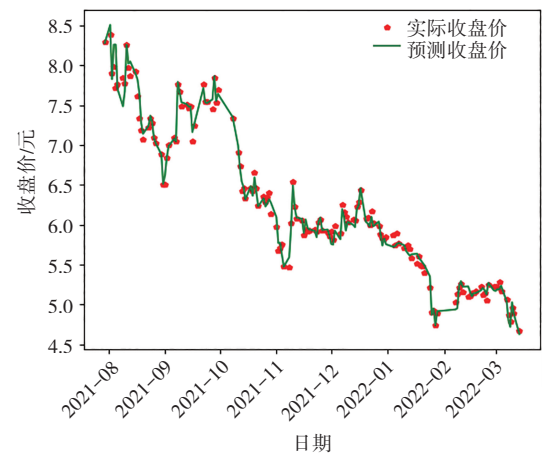
更改参数后,将 Logistic Regression 与 LSTM 模型进行比较,同时在模型的基础上引入情感特征并计算回归指标。

4.3.3 股票价格走势预测

针对测试集的股价走势预测,将模型所得出的收盘价预测值与实际收盘价进行比较,如图 5 以及图 6 所示,分别为两种模型未引入和引入情感特征之后的股价走势预测图。可以明显看出引入情感特征之后的预测收盘价格能够较好地与实际收盘价格重合。



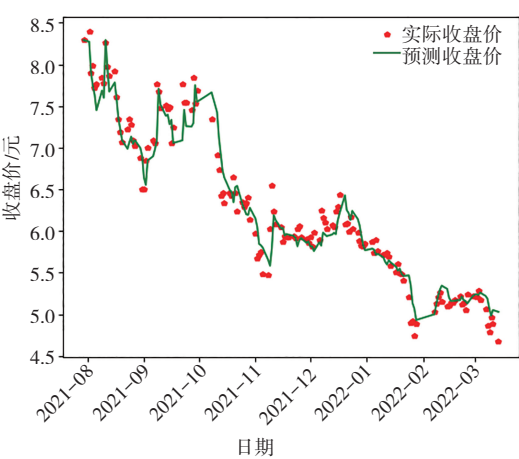
(a) 未引入情感特征的 LSTM 网络



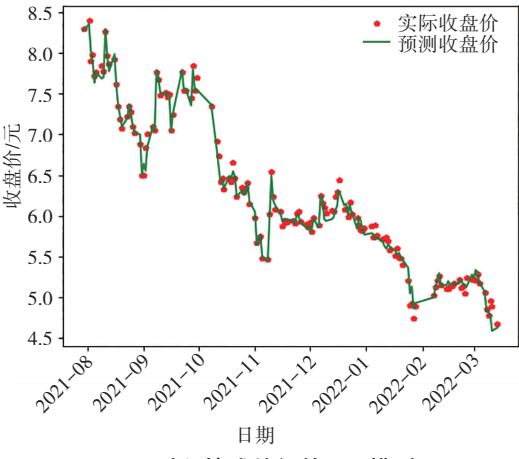
(b) 引入情感特征的 LSTM 网络

图 5 LSTM 网络股价走势预测图

Fig. 5 LSTM network stock price trend prediction chart



(a) 未引入情感特征的 LR 模型



(b) 引入情感特征的 LR 模型

图 6 LR 模型股价预测走势图

Fig. 6 Stock price forecast chart of LR model

为更具体地观察股价的预测走势,将测试集预测的收盘价与实际收盘价进行比较,表 12 为部分股价预测结果以及模型指标对比。由表 12 可得: Logistic Regression 在引入情感特征之后, F_{MAPE} 与 F_{RMSE} 分别下降了 4.3%与 8.40%, R^2 提升了 1.6%;LSTM 在引入情感特征后, F_{MAPE} 与 F_{RMSE} 分别下降了 15.5%与 8.69%, R^2 提升了 2.4%,均有提升。但相对于 Logistic Regression, LSTM 在股价预测趋势上更有优势。

表 12 测试集交易日股价预测结果

Table 12 Stock price forecast results by test set on trading days					
单位:元					
日 期	真实收盘价	LR	LRsentiment	LSTM	LSTM _{sentiment}
2022/3/12	4.67	4.618	4.631	4.642	4.629
2022/3/11	4.89	4.691	4.591	4.817	4.875
2022/3/10	4.96	4.885	4.797	4.854	5.026
2022/3/9	4.78	4.781	4.813	4.819	4.718
2022/3/8	4.86	4.891	4.829	4.721	4.803
2022/3/7	5.06	5.104	5.072	4.992	5.040
2022/3/4	5.17	5.117	5.186	5.171	5.233
2022/3/3	5.28	5.259	5.312	5.237	5.187
2022/3/2	5.21	5.241	5.337	5.301	5.192

2022/3/1	5.22	5.234	5.241	5.278	5.170
2022/2/28	5.22	5.251	5.281	5.198	5.177
...	...	...	...	...	...
2022/8/4	7.98	7.781	7.891	7.824	8.060
2021/8/3	7.9	7.792	8.019	7.952	7.828
2021/8/2	8.39	8.124	8.351	8.489	8.507
2021/7/30	8.29	8.281	8.274	8.102	8.281
模型回归指标对比					
F_{MAPE}	3.72%	3.56%	3.22%	2.72%	
F_{RMSE}	0.119	0.109	0.092	0.084	
R^2	0.912	0.927	0.962	0.986	

5 结 论

本文提出的模型,证实了情感特征对于股票预测的有效性以及 LSTM 网络模型在股票预测中的潜力。本文以华银电力(600744.SH)为例,提出 Bert_LSTM 股票预测模型,得到以下结论。

爬取股吧评论作为文本数据,通过 Bert 自然语言处理对评论进行情感极性分析并打上标签,与财务指标一起输入 LSTM 模型能够提高模型预测准确率,未来可以进一步探讨如政策,市场等因素对于股价预测的影响。

文本采取对照实验,其中分类模型采用 Multinomial NB、XGBoost 作为对比模型来说明 Bert 在文本情感分析的优势,但未尝试深度学习模型在文本情感分类的应用。股票收盘价预测模型使用 Logistic Regression 作为对比模型来说明 LSTM 在预测时序序列上的优势,表明其在处理多维度特征序列时具有更好的预测性能,

对 LSTM 网络而言,记忆天数(mem_days)与 Dropout 率等取值的不同会影响模型的准确率,需要进行迭代才能找到适合数据的最佳参数,后续可以考虑对于 LSTM 模型本身进行优化与改进。

为更好地研究股票市场的发展,未来工作重点可以在以下方面:第一,模型方面,使用 Bert 自然语言处理模型与各种神经网络结合进行对比预测;第二,在进行情感分类时,可以将分类结果细分为多类,如消极、略消极、中立、略积极、积极,在进行股价走势分类时也可以将分类细分为大跌、小跌、小涨、大涨,再来比较不同结果带来的变化;第三,在情感分析的基础上,增加政策性因素以及突发事件、网络舆情带来的变化,但此类事件均属于宏观事件,是否直接影响到股市变化还有待商榷。

参考文献(References):

[1] 陈博闻. 基于技术指标及 ARIMA 模型预测股票价格——以中国平安保险集团公司股票调整后的收盘价为例[J]. 统计与管理, 2021, 36(7): 53—57.
CHEN Bo-wen. Forecast stock price based on technical

- indicators and ARIMA model: taking the adjusted closing price of Ping An Insurance Group of China as an example [J]. Statistics and Management, 2021, 36(7): 53—57.
- [2] 尚卫平, 戴昱. 基于平滑 ARIMA-LS-SVM 组合模型的股价预测研究[J]. 区域金融研究, 2018(5): 17—23.
SHANG Wei-ping, DAI Yu. Research on stock price prediction based on smooth ARIMA-LS-SVM combination model [J]. Journal of Regional Financial Research, 2018(5): 17—23.
- [3] 王燕, 郭元凯. 改进的 XGBoost 模型在股票预测中的应用[J]. 计算机工程与应用, 2019, 55(20): 202—207.
WANG Yan, GUO Yuan-kai. Application of improved XGBoost model in stock forecasting[J]. Computer Engineering and Applications, 2019, 55(20): 202—207.
- [4] 李泽远. 可超越评分卡模型么? 基于 LightGBM 与卷积神经网络在贷款违约风险预测的研究[J]. 特区经济, 2021(5): 67—69.
LI Ze-yuan. Can it go beyond the scorecard model? research on prediction of loan default risk based on light GBM and convolutional neural network [J]. Special Zone Economy, 2021(5): 67—69.
- [5] 李腾. 基于小波分析和 BP 神经网络相结合的股票波动预测方法研究[D]. 天津: 天津大学, 2018.
LI Teng. Research on stock volatility prediction method based on wavelet analysis and BP neural network [D]. Tianjin: Tianjin University, 2018.
- [6] 肖国荣. BP 神经网络在基金价格预测中的应用研究[J]. 计算机仿真, 2011, 28(3): 373—376.
XIAO Guo-rong. Application research of fund net value prediction based on BP neural network[J]. Computer Simulation, 2011, 28(3): 373—376.
- [7] 唐国豪, 姜富伟, 张定胜. 金融市场文本情绪研究进展[J]. 经济学动态, 2016(11): 137—147.
TANG Guo-hao, JIANG Fu-wei, ZHANG Ding-sheng. Research progress of textual sentiment in financial markets [J]. Economic Trends, 2016(11): 137—147.
- [8] 徐月梅, 王子厚, 吴子歆. 一种基于 CNN-BiLSTM 多特征融合的股票走势预测模型[J]. 数据分析与知识发现, 2021, 5(7): 126—137.
XU Yue-mei, WANG Zi-hou, WU Zi-xin. Predicting stock trends with CNN-BiLSTM based multi-feature integration model [J]. Data Analysis and Knowledge Discovery, 2021, 5(7): 126—137.
- [9] MANEL A, CHIRAZ L. BE-BLC: BERT-ELMO-based deep neural network architecture for English named entity recognition task [J]. Procedia Computer Science, 2021, 192: 14—18.
- [10] YU K X, LI Y C, ZHANG D S H. Improved aspect-level sentiment analysis method based on multi-head attention mechanism [J]. Journal of Physics: Conference Series, 2021, 2005(1): 8—13.
- [11] 杨松, 卫文学. Multi Head-Self Attention BiLSTM 网络应用于主观题评分的研究[J]. 软件, 2020, 41(12): 180—184.
YANG Song, WEI Wen-xue. Multi Head-Self Attention BiLSTM network applied to subjective question grading research [J]. Software, 2020, 41(12): 180—184.
- [12] 李华旭. 基于 RNN 和 Transformer 模型的自然语言处理研究综述[J]. 信息记录材料, 2021, 22(12): 7—10.
LI Hua-xu. A review of natural language processing based on RNN and Transformer models [J]. Journal of Information Record Materials, 2021, 22(12): 7—10.
- [13] 陈宇韶, 唐振军, 罗扬, 等. 皮尔森优化结合 Xgboost 算法的股价预测研究[J]. 信息技术, 2018(9): 84—89.
CHEN Yu-shao, TANG Zhen-jun, LUO Yang, et al. Research on stock price prediction based on Xgboost algorithm with Pearson optimization [J]. Information Technology, 2018(9): 84—89.
- [14] 李文丽. 基于朴素贝叶斯分类的网络谣言识别研究[J]. 计算机工程与科学, 2022, 44(3): 495—501.
LI Wen-li. Network rumor recognition based on naive Bayesian classification [J]. Computer Engineering & Science, 2022, 44(3): 495—501.
- [15] ZHENG K F, WANG X J, WU B. Feature subset selection combining maximal information entropy and maximal information coefficient [J]. Applied Intelligence, 2020, 50(2): 15—20.
- [16] 李兴奇, 高晓红. 无量纲化方法的有效性评价[J]. 统计与决策, 2021, 37(15): 24—28.
LI Xing-qi, GAO Xiao-hong. Validity evaluation of dimensionless methods [J]. Statistics and Decision, 2021, 37(15): 24—28.
- [17] 王丹凤, 李书臣, 翟春艳, 等. 基于神经网络参数优化的迭代学习控制算法[J]. 河南科技大学学报(自然科学版), 2008(6): 42—44, 110.
WANG Dan-feng, LI Shu-chen, ZHAI Chun-yan, et al. An iterative learning control algorithm of parameters optimization based on neural network [J]. Journal of Henan University of Science and Technology (Natural Science Edition), 2008(6): 42—44, 110.
- [18] 王文轩, 蔡伟宏. 基于 Logistic 回归的股价上涨概率预测研究[J]. 中国市场, 2020(6): 7—8.
WANG Wen-xuan, CAI Wei-hong. Research on probability prediction of stock price rise based on logistic regression [J]. China Market, 2020(6): 7—8.
- [19] HINTON GEOFFREY E. Machine learning for neuroscience. [J]. Neural Systems Circuits, 2011, 1(1): 1—1.