文章编号：2095-2163(2022)05-0107-07 中图分类号：F83 文献标志码：A

基于 LSTM 和新闻情感的股票价格预测方法

许 丽，张 利，李桂城，肖一凡，陈丽绵，唐 艳
（贵州大学 大数据与信息工程学院，贵阳 550025）

摘 要： 股票预测研究对于经济发展具有重要意义，也是困扰投资者的难题。本文提出了一种基于 LSTM 和新闻股票情感分析的组合优化模型 SVM_LSTM。首先将 XGBoost 和利用交叉验证优化的 LSTM 应用于预测中国银行、中国联通以及浦发银行的每日收盘价上，通过对比二者的性能，选择较优的 LSTM 对中国银行股票历史价格进行最终的时序预测；然后，使用 SVM 对中国银行的股票新闻进行情感倾向预测；最后，采用加权的方式将 SVM 的预测结果与 LSTM 的预测的结果进行融合。实验结果表明：第一，利用交叉验证优化的 LSTM 较 XGBoost 具有更优的评价指标，针对中国银行数据集，其 *RSEM*、*MAE*、*MSE* 比 XGBoost 分别减少了 0.234、0.173、0.011；第二，采用加权的方式将 SVM_LSTM 的预测结果调和，实验结果较原 LSTM 而言，评价指标 *RSEM*、*MAE*、*MSE* 分别减少了 7.5%、6.4%、10.8%。

关键词： SVM；XGBoost；交叉验证；LSTM；股票价格预测；新闻情感预测

Stock price prediction method based on LSTM and news sentiment

XU Li, ZHANG Li, LI Guicheng, XIAO Yifan, CHEN Limian, TANG Yan
(College of Big Data and Information Engineering, Guizhou University, Guiyang 550025, China)

【Abstract】 Research on stock forecasting is of great significance to economic development, as well as a difficult problem for investors. This paper proposes a combination optimization model SVM_LSTM based on LSTM and sentiment analysis of news stock. Firstly, XGBoost and LSTM optimized by cross validation are applied to predict the daily closing prices of Bank of China, China Unicom and Shanghai Pudong Development Bank. By comparing the performance of the both, the better LSTM is selected to make the final time series prediction of the historical stock prices of Bank of China. Then, SVM is used to predict the emotional tendency of Bank of China′s stock news. Finally, the prediction results of SVM and LSTM are fused by weighted method. The experimental results show that: firstly, the optimized LSTM with cross validation has better evaluation index than XGBoost. Compared with XGBoost, the *RSEM*, *MAE* and *MSE* of The Bank of China data set are reduced by 0.234, 0.173 and 0.011, respectively. Secondly, the prediction results of SVM _ LSTM are reconciled by weighted method. Compared with the original LSTM, the experimental results show that the evaluation indices, such as *RSEM*, *MAE* and *MSE* are reduced by 7.5%, 6.4% and 10.8% respectively.

【Key words】 SVM; XGBoost; cross validation; LSTM; stock price forecast; news sentiment prediction

0 引 言

股票市场是国家经济市场的重要组成之一。随着中国经济的腾飞和大数据的发展，一方面，关心股市问题的不再只有上市公司，越来越多的普通民众将目光转向股票投资；另一方面，影响股票价格的因素有经济、政治和公司自身的管理等，均使得股票价格走势变得难以预测。因此，众多研究者开始深入研究影响股票价格的因素，并提出了许多经典模型。

文献[1]提出了一种 ARIMA 和 SVM 的组合预测模型，首先使用 ARIMA 模型对华泰证券一年的股票价格进行线性预测，接着使用 SVM 模型对其进行非线性预测，实验表明，组合预测模型得到的综合结果精度高于单一模型。文献[2]改进了 XGBoost 模型，使用网格搜索算法对模型进行参数优化，基于原 XGBoost、GBDT、SVM 以及改进的 XGBoost 模型对 5 个短期数据集，如中国平安、中国建筑等进行预测，实验结果表明，XGBoost 表现出了最优的评价指标和最好的拟合性能。文献[3]提出了一种粒子群和支持向量机的组合模型，传统的支持向量机在股票预测中取得了较好的效果，但是根据人为经验选取支持向量机的参数和核函数存在很大的弊端，而该模型利用粒子群算法可以自动更新速度和位置且各粒子之间共享信息的特性，对 SVM 的参数进行优

作者简介：许 丽(1998-)，女，硕士研究生，主要研究方向：时序预测、金融大数据；张 利(1987-)，男，博士，特聘教授，主要研究方向：天文大数据、医疗大数据、金融大数据、图像处理；李桂城(1998-)，男，硕士研究生，主要研究方向：时序预测、金融大数据。
通讯作者：张 利 Email：lizhang.science@gmail.com

收稿日期：2021-12-09

化,实验结果显示,该算法在很大程度上提高了预测结果,对比原始的 SVM,准确率有所提高。文献[4]提出了 Bagging-SVM 模型,算法的思想是利用 bagging 将数据集划分为若干的子训练集去训练传统的 SVM,最后通过对每一个子 SVM 模型的预测结果进行投票,票数最多的项即为最终的预测结果。文献[5]在对输入数据进行预处理后,将其送入层数不同的 LSTM 和层数相同、但神经元不同的网络结构中,通过评价指标来选择合适的结果,并对苹果公司的股价进行预测分析,验证了其精度可提高 30%,证明了该方法的可行性。文献[6]将 LSTM 网络用于股票价格预测,通过对模型进行不断调优,获得最优预测模型后将其预测结果与 BP 神经网络、RNN、CNN,人工神经网络的预测结果进行对比,结果表明,LSTM 评价指标最优且预测值和真实值的曲线拟合得最好。文献[7]提出了一种基于粒子群算法改进的 LSTM 模型,传统 LSTM 网络常常根据自身经验确定其中的重要参数,由于主观性极强无法确定最优值,粒子群算法的提出解决了这一问题,通过算法寻优构建了完善的预测模型,使得准确率大大提高且获得了普遍的适用性。文献[8]将不同的输入特征送入 LSTM 模型来预测股票的收入,验证了特征不同的输入对股票价格的影响。文献[9]提出了 Adv-ALSTM 模型,通过添加扰动来模拟价格的随机性,以此来训练模型在干扰数据下的泛化能力,实验表明,在训练模型阶段,采取对抗性网络的方式大大提高了网络性能和拟合能力。为了同时满足捕获长时间依赖关系和选择相关驱动进行预测的要求,文献[10]提出了 DA-RNN 网络,该网络是一个 2 层 LSTM 结构,包含解码和编码环节,将该模型应用在 SML2010 数据上,结果表明该网络不仅可以进行有效预测,而且具有很强的可解释性。文献[11]提出了基于 TCN 和文本情感分析方法的股票价格预测方法,通过爬虫的方式获取相关数据集,使用 LSTM 对新闻文本进行情感极性分析,并将其结果与股票数据一同输入 TCN 网络,改进后模型的评价指标表现良好,验证了其可行性。

本文的结构大致如下:第一节综述了股票价格预测的经典模型和算法;第二节介绍了本文使用的一些相关算法机理;第三节概述了本文算法的原理;第四节给出了本文网络在相关数据集上的评价指标和拟合曲线的实验结果;第五节总结概括了该文章,并提出该领域未来可能的发展方向。

1 相关算法

1.1 交叉验证

泛化能力是评价一个模型好坏的指标之一。因此,常常使用交叉验证来提高模型的泛化能力。常见的交叉验证有简单交叉验证、留一交叉验证以及 k 折交叉验证。其中,简单交叉验证仅仅需要将数据集划分为 2 份,选取数据量小的一份作为验证集来评价模型的好坏,该算法虽然可以用来评价模型,但是划分的比率不同会产生较大的差异。留一法,顾名思义,将 n 个样本划分为 n 份,进行 n 次训练,每次只留一份为验证集,该方法在数据量小的时候可以最大限度地利用每一个样本,其弊端在于如果数据量很大时会造成欠拟合。k 折交叉验证把数据集划分为 k 份,每一份逐一作为验证集去评价模型,基于此再取平均值作为最后的评价指标。一般情况下,k 值越大则模型性能越好,但是有研究表明,其范围在 5~10 之间或能取得最佳效能,本文使用 10 折交叉验证对 LSTM 模型进行训练。

1.2 XGBoost

XGBoost 是梯度提升决策树算法(DBDT)的改进版本。由叶节点为分类变量的决策树和叶节点为连续变量的回归树两种增强树组成。DBDT 梯度提升树的表达式如(1)所示:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{1}$$

为了避免正则化,通过在原网络的基础上增加正则项来简化模型,XGBoost 包括回归树和决策树,并且采用了 blocks 的存储结构,使得算法可并行,提高了运算速度。该算法的最终目标是优化函数 $Obj^{(t)}$,其中包含损失函数 l 和表示复杂度的正则项 Ω。网络在每一次迭代之后增加一棵树,来优化 $Obj^{(t)}$。函数的数学定义可写为如下形式:

$$Obj^{(t)} = \sum_{i=1}^{n} l((y_i, \hat{y}^{(t-1)}) + f_t(x_i)) + \Omega(f_t) + C \tag{2}$$

其中,$\hat{y}^{(t-1)}$ 表示保留前面 $t-1$ 次的预测结果。将 DBDT 的表达式代入公式(2)的目标函数,第 t 次迭代的目标是为了使得目标函数最小。

损失函数在第 $t-1$ 次迭代时获得的预测值的一次偏导 g_i 和二次偏导 h_i,如式(3)所示:

$$g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)}), h_i = \partial^2_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)}) \tag{3}$$

将目标函数按泰勒二阶公式展开，消除常数项带来的影响。二阶导数使得损失函数更加精确，并将式(3)带入，得到新的目标函数如式(4)所示：

$$Obj^{(t)} \approx \sum_{i=1}^{n} [l(y_i, \hat{y}^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) + C \tag{4}$$

进一步地，研究中可推得：

$$G_j = \sum_{i \in I_j} g_i, \quad H_j = \sum_{i \in I_j} h_j \tag{5}$$

在训练目标函数时，$f_t(x) = w_j$，即把函数(树)拆分为叶子的权重向量部分 w 和树的结构部分 q，目标函数可以表示为：

$$Obj^{(t)} = \sum_{j=1}^{t} [\left(\sum_{i \in I_j} g_i\right) w_j + \frac{1}{2}\left(\sum_{i \in I_j} h_i + \lambda\right) w_j^2 + \lambda T] \tag{6}$$

将式(5)带入式(6)合并一次项系数和二次项系数，得到最终的目标函数如式(7)所示：

$$Obj^{(t)} = \sum_{j=1}^{t} \left[G_j w_j + \frac{1}{2}(H_j + \lambda) w_j^2\right] + \lambda T = -\frac{1}{2} \sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T \tag{7}$$

目标函数又可称作打分函数，在一定范围和条件下，目标函数的值最小，表明此时得到了最优解。总而言之，XGBoost 具有很多优点。不仅可以减少过拟合并且处理正则项，也可提升运算速度、可自定义目标函数 $Obj^{(t)}$ 以及模型评价指标，具有高度灵敏性。除此之外，XGBoost 具有自动处理数据集缺失值的功能，允许在 boosting 迭代中使用交叉验证。

2　算法机理及流程

2.1　SVM_LSTM 模型的搭建

随着人们生活水平的提高，在满足日常衣食住行之余，已有很多人开始利用手中的闲钱来做投资，由于股票的低投资以及高回报性，使其成为大众的重要选择，基于此，如何选择最优股即已成为备受关注的热点话题。以往的研究者要么针对股票历史数据进行时序预测，要么采取分类网络进行文本情感分析，忽略了二者直接的内在联系。在此背景下，本文提出了 SVM_LSTM 网络。

在时序预测通道上，本文首先采用归一化的手段处理股票数据，对股票数据进行归一化处理可以削弱量纲带来的影响，将数据统一到一个相差不大的范围，以此来减少较大数据和较小数据的影响，本文采取的归一化公式具体如下：

$$x_i' = \frac{x_i - \bar{x}}{\max(x_i) - \min(x_i)} \tag{8}$$

其中，x_i 表示第 i 个变量，$\bar{x}$ 表示均值。

其次，本文使用 10 折交叉验证划分经过预处理的股票数据，来训练 LSTM 模型；最后，调用训练良好的模型预测中国银行、中国联通、浦发银行的股票数据。LSTM 机理详见 2.2 节。

在文本预测通道上，对于中国银行股票新闻的文本数据集，本文首先在百度智能云网站注册一个 AipNlp 账号，这是自然语言处理的 Python SDK 客户端，通过调用该网站可以对本文爬取的文本数据集做一个初步的处理，由于使用该接口对新闻文本标题预测得到的结果为每一条新闻词条对应的积极的可能性和消极的可能性，本文在 Python 中指定当积极的可能性大于消极的可能性时，就将该新闻文本标注为 1，代表利好，反之，将文本标注为 0，表示利空；随后采用处理好的文本数据集训练 SVM 模型；最后，调用 SVM 预测文本新闻情感极性。SVM 原理详见 2.3 节。

至此，研究中采用了加权的方式将 SVM 的预测结果与 LSTM 的预测结果进行融合。其算法流程如图 1 所示。图 1 中，第 i 天的股票最终预测价格 y_pred_{last} 的计算方法如式(9)所示：

$$y\_pred_{last} = y\_pred_i - y\_pred_i * weight \tag{9}$$

其中，y_pred_i 表示时序预测第 i 天的结果。在本文中，由于在时序预测阶段，真实值与预测值在整体走势上拟合得很好，只需要微小的调整就可以使得预测值更接近于真实值，故而本文将 *weight* 设置为 0.01。

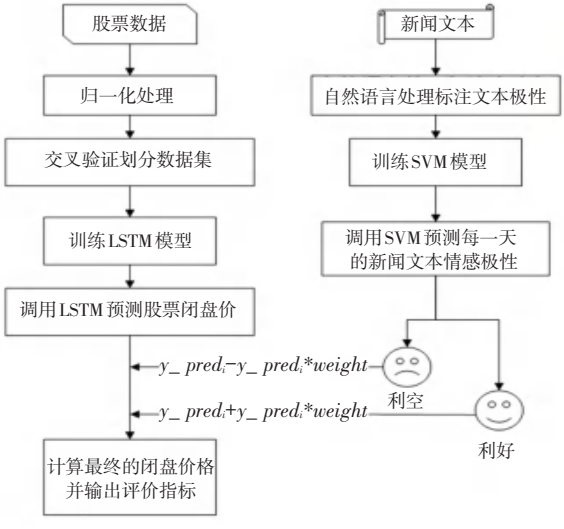


图 1　算法流程图

Fig. 1　Algorithm flow chart

2.2 LSTM

循环神经网络 RNN 能够处理时序数据,时序数据、即数据之间具有时间先后性,数据整体呈现某种趋势。循环神经网络具有跨越时间节点的自连接层,能够建立当前时刻与序列上一时刻的关系。但是随着网络层的加深,由于循环神经网络只会获取上一节点的信息,无法储存距离当前时刻较远的网络的输出,会造成梯度消失。为了解决梯度消失和梯度爆炸,最早由 Zhao 等人[12]提出了 LSTM 网络,在此后的一段时间里,该网络被广泛地应用在时序预测邻域中。

LSTM 是对 RNN 模型的改进,以达到解决梯度消失的效果。这一改进主要表现为:在 RNN 隐藏层中添加了长短期记忆单元;通过添加门控结构、引入 *sigmoid* 激活函数结合 RNN 原有的 tanh 激活函数,来控制网络对历史信息的输入保存和输出;输入信息由添加的细胞状态记忆。综上所述可知,LSTM 解决了短期依赖和长距离依赖问题。

LSTM 的 3 个门控单元分别为输入层、隐藏层、输出层,共同构成了模型的输入部分。LSTM 网络结构如图 2 所示。由图 2 可知,该结构中各主体组成部分的设计原理及数学表述详见如下。

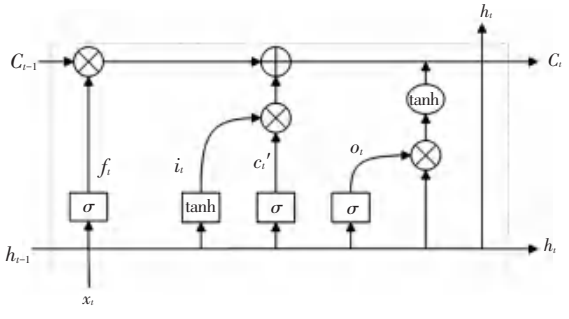


图 2 LSTM 网络结构

Fig. 2 LSTM network structure

(1)输入层。为全连接层,该网络首先对数据进行预处理,以达到输入数据格式要求。输入门的值 i_t 和输入细胞的候选状态值 $\tilde{C}_t$ 可由如下方式计算得到:

$$\boldsymbol{i}_t = \delta(\boldsymbol{W}_i * (\boldsymbol{X}_t, \boldsymbol{h}_{t-1}) + \boldsymbol{b}_i) \tag{10}$$

$$\tilde{\boldsymbol{C}}_t = \tanh(\boldsymbol{W}_c * (\boldsymbol{X}_t, \boldsymbol{h}_{t-1}) + \boldsymbol{b}_c) \tag{11}$$

其中,$\boldsymbol{W}$ 和 $\boldsymbol{U}$ 表示计算时的权重矩阵;$\boldsymbol{b}$ 表示偏置向量;σ 表示激活函数 *sigmoid*。

(2)隐藏层。是包含多个 LSTM 神经元的循环神经网络。在该网络结构中,激活函数选取 σ 和 tanh,t 时刻遗忘门的 *sigmoid* 和当前时刻细胞的状态的更新值的数学表达式可写为:

$$\boldsymbol{f}_t = \delta(\boldsymbol{W}_f * (\boldsymbol{X}_t, \boldsymbol{h}_{t-1}) + \boldsymbol{b}_f) \tag{12}$$

$$\boldsymbol{C}_t = \boldsymbol{i}_t * \tilde{\boldsymbol{C}}_t + \boldsymbol{f}_t * \boldsymbol{C}_{t-1} \tag{13}$$

(3)输出层。将隐含层的多个输出结果映射到全连接层来获得模型的最终输出。在该层网络结构中,计算得到的最终输出结果如下:

$$\boldsymbol{O}_t = \delta(\boldsymbol{W}_O * (\boldsymbol{X}_t, \boldsymbol{h}_{t-1}) + \boldsymbol{b}_O) \tag{14}$$

$$\boldsymbol{h}_t = \boldsymbol{O}_t * \tanh(\boldsymbol{C}_t) \tag{15}$$

2.3 SVM

假设股票输入样本为 n 维空间的一个点 X 为 $(x_1, x_2, \cdots, x_n)$,y_1 和 y_2 分别表示股票的涨跌,SVM 的重点是建立一个超大平面的决策曲面,设超平面 $g(\boldsymbol{x})$ 为 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b = 0$。当该值大于 0 时,分类为 y_1;反之,分类为 y_2,支持向量的平面范围值大小为 $[-1, 1]$,落在这之间的样本点即为支持向量机。从而得到最大间隔 $d = \dfrac{2}{\| w \|}$。SVM 分类原理则如图 3 所示。

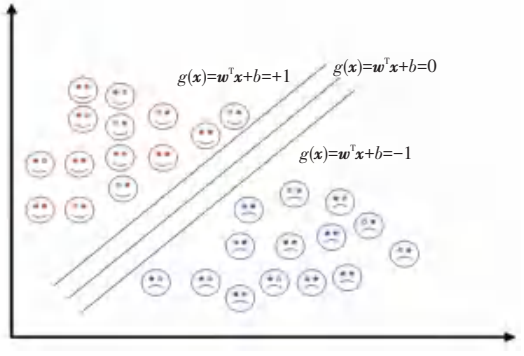


图 3 SVM 分类示意图

Fig. 3 SVM classification diagram

利用拉格朗日优化最大间隔,最终得到决策函数具体如下:

$$f(x) = \mathrm{sgn}\Big(\sum_{i=1}^{N} y_i\, \alpha_i\Big) \cdot K((x_i\, x_j) + b) \tag{16}$$

$$N = y_1 + y_2 \tag{17}$$

其中,N 表示分类总数。

二次规划问题,求解约束最优化的问题变成公式(18):

$$Max \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i\, \alpha_j \cdot y_i\, y_j \cdot K(x_i \cdot x_j) \tag{18}$$

$$\text{s.t. } 0 \leqslant \alpha_i \leqslant C$$

$$\sum_{i=1}^{N} \alpha_i\, y_i = 0, i = 1, 2, \cdots, N$$

其中,c 为惩罚函数,该值与模型对噪声的容忍

度成正比，与模型的泛化能力成反比。

常见的 SVM 核函数有线性核函数、多项式核函数、径向基核函数，对应的数学公式可顺次表示为：

$$K(x_i,x_j)=(x_i,x_j) \tag{19}$$

$$K(x_i,x_j)=(x_i\cdot x_j+1)^d \tag{20}$$

$$K(x_i,x_j)=\exp(-\gamma\parallel x_i-x_j\parallel^2) \tag{21}$$

其中，d 和 γ 是常数项。

核函数的选择决定了模型的预测效果，通过核函数网络，SVM 将输入的向量映射到高维特征的空间，将原本的非线性问题变换为高维度特征空间的线性可分问题。

3　实验

3.1　数据集介绍

本文通过 baostock 库爬取了中国银行、中国联通、浦发银行三只股票从 2020 年 8 月 30 号到 2021 年 10 月 27 号的相关数据，包括开盘价 *open*，收盘价 *close*，最高价 *high* 以及最低价 *low*。收盘价表示股票当天的最后成交价，一般将其作为股票价格预测中唯一的因变量。图 4 是中国银行的收盘价随时间的变化曲线，横坐标表示不同的日期，纵坐标表示相应日期对应的股票价格。从图 4 中可以看到该时序数据整体价格分布在 2~4 之间，某一天的闭盘价周围一段时间内不会发生太大的突变；但是，图 4 中黄色方框处股票急剧上升，在绿色方框处股票又突然下降。

对于新闻文本数据集，本文利用 lxml 库爬取了“新浪财经网”和“金融界”网站中有关中国银行的新闻文本。lxml 是 XML 和 HTML 文件的解析器，还可以用于 Web 爬取。

图 4　中国银行股票跌涨曲线

Fig. 4　Stock decline and rise curve of Bank of China

3.2　评价指标

本文采取均方根误差（*RMSE*）、平均绝对误差（*MAE*）以及均方误差（*MSE*）来衡量模型的性能。各指标值的数学公式可表示如下：

$$MSE=\frac{1}{n}\sum_{i=1}^{n}(y^{(i)}-\hat{y}^{(i)})^2 \tag{22}$$

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y^{(i)}-\hat{y}^{(i)})^2} \tag{23}$$

$$MAE=\frac{1}{n}\sum_{i=1}^{n}|y^{(i)}-\hat{y}^{(i)}| \tag{24}$$

其中，n 是样本数；$\hat{y}^{(i)}$ 是模型预测值；$y^{(i)}$ 是真实值。

对于分类模型 SVM，本文采取预测准确率（*Accuracy*）来衡量效果的好坏，研究后可推得的数学公式如下：

$$Accuracy=\frac{TP+TN}{TP+FP+TN+FN} \tag{25}$$

其中，具体参数含义见表 1。

表 1　预测与实际情况类型表

Tab. 1　Forecast vs. actual type table

实际/预测	积极(1)	消极(0)
上涨	*TP*	*FN*
下跌	*FP*	*TN*

3.3　XGBoost 和 LSTM 模型的预测结果

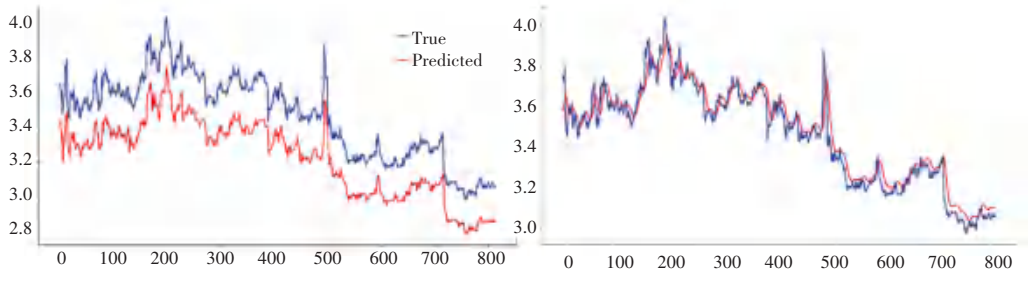
图 5(a)~图 5(c)中，左侧为 XGBoost 模型的预测值与真实值的拟合曲线，右侧为 LSTM 网络的预测值与真实值拟合曲线。本文中，如果不做特殊说明，统一使用红色曲线表示预测曲线，蓝色曲线表示真实值曲线。可以看到：

(1)XGBoost 模型拟合得到的预测值与真实值

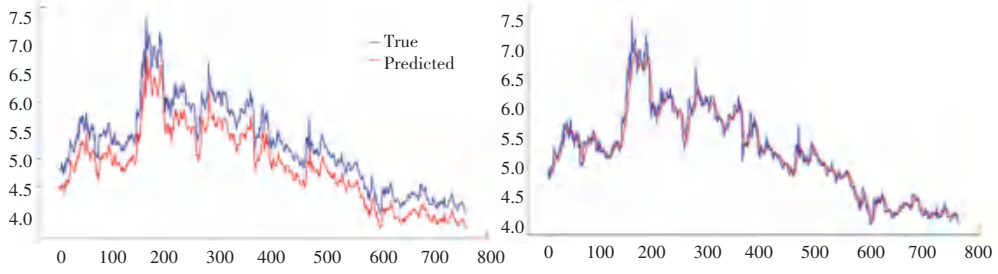
相比,其涨幅趋势大致相同,并且在第 N 天出现峰值时也能够进行较好的预测,然而预测值曲线整体处于真实值下方,说明对于数值的预测存在偏差,因此该模型性能仍存有提升空间。

(2)LSTM 在保留了 XGBoost 优点的同时,预测值无限逼近真实值,并且与绝大部分的低谷值和峰值完全吻合,针对中国银行数据集,表 2 显示了 LSTM 的评价指标 *RMSE*、*MAE* 和 *MSE* 较 XGBoost 分别减少了 0.234、0.173 和 0.011,说明了本文搭建的模型预测精度更高。

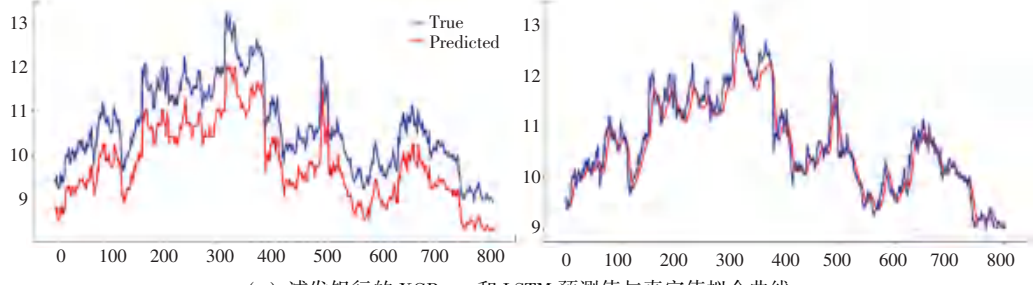
(3)中国银行无论在曲线拟合程度、还是在评价指标上,均优于中国联通和浦发银行,对于 LSTM 网络,各评价指标之间整体没有大致的差别,表明了本文模型具有较好的泛化能力。



(a) 中国银行的 XGBoost 和 LSTM 预测值与真实值拟合曲线



(b)中国联通的 XGBoost 和 LSTM 预测值与真实值拟合曲线



(c) 浦发银行的 XGBoost 和 LSTM 预测值与真实值拟合曲线

图 5 XGBoost 和 LSTM 模型的预测结果

Fig. 5 Prediction results of XGBoost and LSTM models

表 2 XGBoost 和 LSTM 在各数据集上的评价指标

Tab. 2 Evaluation indicators of XGBoost and LSTM in each dataset

数据集	评价指标	XGBoost	LSTM
中国银行	*RMSE*	0.238	0.004
	MAE	0.235	0.062
	MSE	0.057	0.046
中国联通	*RMSE*	0.390	0.023
	MAE	0.375	0.152
	MSE	0.152	0.105
浦发银行	*RMSE*	0.827	0.079
	MAE	0.812	0.280
	MSE	0.684	0.202

3.4 SVM 实验结果

使用中国银行股票新闻数据集对 SVM 模型进行训练,该模型的准确率(*Accuracy*)达到了 81%,高于文献[3]中 SVM 模型所达到的 79.11%,说明本文模型的可行性。表 3 展示了自 2021 年 9 月 1 号到 2021 年 10 月 25 号的预测结果,其中 1 代表利好,0 代表利空,分别对应了股票的上涨和下跌。表 3 中的新闻文本标题均为中国银行的个股资讯,由于不是每天都有相关新闻消息,表中的日期并不连续。没有日期则对应的时序预测股票价格保持不变。

表 3　SVM 预测结果

Tab. 3　SVM prediction results

日期	新闻标题	情感极性
10.25	中国银行:刘秋万辞去首席信息官职务	1
10.18	中国银行上海分行违法被罚 540 万:贷款资金用于购房等	0
10.17	中国银行南平分行违法被罚:贷前调查和贷后管理未尽职	0
10.02	中国银行莆田分行违法被罚:未按工程进度发放贷款等	0
09.30	中国银行(03988.HK):预计 10 亿元票据 9 月 30 日上市	1
09.24	中国银行:郑国雨因工作调动辞去副行长职务	1
09.24	中国银行副行长郑国雨辞职:上半年公司净利 1128.13 亿	1
09.24	中国银行:郑国雨因工作调动:辞去副行长职务	1
09.18	首家！中国银行横琴粤澳深度合作区分行机构落地	1
09.17	中国银行卡产业保持稳健发展,银行卡欺诈率连续四年下降	0
09.14	中国银行乌鲁木齐市分行被罚:未有效监督贷款资金用途	0
09.10	奥园美谷医美制造标杆获中国银行 10 亿授信	1
09.06	中国银行吉林 3 分支被罚:未合理审查交易单证真实性等	0
09.06	中国银行业协会:2020 年末银行业托管规模达 169.04 万亿元,同比增长 10.29%	1
09.01	2021 年度中国银行业发展报告:上半年银行业主要指标持续改善	1

3.5　SVM_LSTM 模型实验结果

本文将中国银行的预测结果按照对应日期加权(2.1 节附有详细说明)到 LSTM 的预测结果上。经过计算,本文网络 SVM_LSTM 的最终结果见表 4,显然,各评价指标较原 LSTM 均明显减少，*RMSE*、*MAE*、*MSE* 分别减少了 7.5%、6.4%、10.8%。

表 4　SVM_LSTM 评价指标

Tab. 4　SVM_LSTM evaluation index

网络	*RMSE*	*MAE*	*MSE*
LSTM	0.004 0	0.062	0.046
SVM_LSTM	0.003 7	0.058	0.041

4　结束语

本文综合考虑了影响股票价格的双重因素,分别是新闻文本和股票历史数据,通过 SVM 和 LSTM 对 2 个通道的数据进行预测,再对其预测结果进行加权融合,使得 2 种不同的数据类型相互补充,让最终预测股价更接近于真实价格。将本文模型应用于中国移动股票的预测上,结果表明:

(1)利用交叉验证优化的 LSTM 虽然比 XGBoost 模型具有更高的预测精度,曲线拟合程度也更优,但是模型预测值和真实值之间仍然存在一定的差距。

(2)SVM 模型预测准确率虽然达到了 81%,能够给投资者提供大致的股票跌涨讯息,可是却不能够获取实际闭盘价格。

(3)本文采取加权融合的方式将 SVM 以及 LSTM 的预测结果融合,使得评价指标 *RMSE*、*MAE*、*MSE* 在原 LSTM 的基础上分别减少了 7.5%、6.4%、10.8%,证明了 SVM_LSTM 网络的实用性,能够给投资者带来一定的参考价值。

下一步工作将考虑使用爬虫获取更多与股票相关的数据和文本讯息,进行多维度、深层次的分析,同时考虑分别优化时序和文本预测网络,采取更科学的举措归并两者的预测结果。

参考文献

[1] 张青. 基于 ARIMA-SVM 组合模型的创业板股票价格预测分析[J]. 广西质量监督导报, 2019(12):131-132.

[2] 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.

[3] 张秀香, 韦文山. PSO-SVM 在股票收盘价中的研究与应用[J]. 信息与电脑, 2020, 32(20): 44-47.

[4] 陈亚男, 薛雷. 基于 Bagging-SVM 的股票趋势预测技术[J]. 电子测量技术, 2019 (14): 58-62.

[5] 彭燕, 刘宇红, 张荣芬. 基于 LSTM 的股票价格预测建模与分析[J]. 计算机工程与应用, 2019, 55(11): 209-212.

[6] 黄建华, 钟敏, 胡庆春. 基于改进粒子群算法的 LSTM 股票预测模型[J/OL]. 华东理工大学学报 (自然科学版):1-12[2021-10-14].https://doi.org/10.14135/j.cnki.1006-3080. 20210616001.

[7] CHEN Kai, ZHOU Yi, DAI Fangyan. A LSTM-based method for stock returns prediction: A case study of China stock market[C]// Proceedings of the 2015 IEEE International Conference on Big Data (Big Data). Washington, DC, United States: ACM, 2015: 2823-2824.

[8] FENG Fuli, CHEN Huimin, HE Xiangnan, et al. Enhancing stock movement prediction with adversarial training[J]. arXiv preprint arXiv:1810.09936, 2018.

[9] QIN Yao, SONG Dongjin, CHEN Haifeng, et al. A dual-stage attention-based recurrent neural network for time series prediction [J]. arXiv preprint arXiv:1704.02971, 2017.

[10] LI Xinyi, LI Yinchuan, YANG Hongyang, et al. DP-LSTM: differential privacy - inspired LSTM for stock prediction using financial news[J]. arXiv preprint arXiv:1908.10806,2019.

[11] 田红丽,金硕,闫会强. 一种基于 TCN 和新闻情感的股票预测方法[J]. 信息技术与信息化,2021(06):17-21.

[12] ZHAO Zheng, CHEN Wehai, WU Xingming, et al. LSTM network: a deep learning approach for short-term traffic forecast [J]. IET Intelligent Transport Systems, 2017, 11(2): 68-75.