

CStock:一种结合新闻与股价的股票走势预测模型

陈可心,黄 刚

(南京邮电大学 计算机学院,江苏 南京 210023)

摘 要:股票是一种高风险、高收益的常见理财产品,为了更好地进行股票投资分析,获得有效的选股方案,文中提出了一种预测股票走势的模型 CStock。与现有的股票走势预测模型相比,CStock 模型结合新闻和股价走势进行预测,不但利用了股票市场中的交易数据,同时考虑到财经以及政治新闻对于股票市场的影响。CStock 模型主要由 BiLSTM 和 CLSTM 混合构建,BiLSTM 提取股票交易数据的相关特征,CLSTM 对新闻的语境特征进行整合和处理,最终通过全连接层输出预测结果。在实验模型中,对股票走势采用分类方法进行实验,得到分类为股票上升的概率和股票下降的概率。实验使用美股数据作为数据集。通过准确率和收益率进行预测效果评估,实验结果表明,CStock 模型在一定程度上能够准确有效地对股票走势进行预测。

关键词:股票预测;深度学习;LSTM;BiLSTM;CLSTM

中图分类号:TP391

文献标识码:A

文章编号:1673-629X(2020)09-0018-05

doi:10.3969/j.issn.1673-629X.2020.09.004

Cstock: A Stock Trend Forecasting Model Combining News and Stock Price

CHEN Ke-xin, HUANG Gang

(School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210023, China)

Abstract: Stock is a high-risk, high-yield common financial product. In order to better conduct stock investment analysis and obtain an effective stock selection plan, we propose a model CStock for predicting the stock trend. Compared with the existing stock trend forecasting model, the CStock model combines news and stock price trend to predict. It not only makes use of trading data in the stock market, but also takes into account the influence of financial and political news on the stock market. The CStock model is mainly constructed by mixing BiLSTM and CLSTM. BiLSTM extracts the relevant characteristics of stock trading data, CLSTM integrates and processes the contextual features of news, and finally outputs the predicted results through the fully connected layer. In the experimental model, the stock trend is tested by a classification method, which is classified as the probability of stock rise and the probability of stock decline. The US stock data is used as a data set in the experiment. The prediction results are evaluated by the accuracy rate and the rate of return. The experiment shows that the CStock model can accurately and effectively predict the stock trend to a certain extent.

Key words: stock prediction; deep learning; LSTM; BiLSTM; CLSTM

0 引言

股票预测研究是金融大数据的一个应用研究方向,随着中国经济的快速增长和金融市场的不断扩大,越来越多的投资者开始关注提高投资回报率并能有效地避免一定风险的方法,其中对股票走势预测在商业和金融领域具有重要的意义。面对股票价格的涨跌,投资者会获得难以预测的收益甚至亏损,因此预测股票的走势,选取值得投资的股票成为投资者关心的问题。鉴于股票市场的复杂性、不稳定性,预测股票走势所需要考虑的变量和信息来源的数量巨大,预测股票

走势是一个非常艰难的任务,如今依旧是各领域重点关注与讨论的对象。传统的分析方法主要是利用既有的股票数据和相关技术图表,结合投资者自身经验对股票走势进行预测。但是这种方法在当今日益庞大且复杂的股票市场中并不适用。除了效率低、过于依靠人工经验以外,还存在股票内容信息完整性差、特征数据冗余等一系列问题,对股票数据的利用率低,效果不佳,难以满足市场发展的需要。

随着机器学习技术的不断发展,越来越多的投资者开始使用机器学习技术对股票数据进行分析,从价

收稿日期:2019-09-08

修回日期:2020-01-09

基金项目:国家自然科学基金(61171053);南京邮电大学基金(SG1107)

作者简介:陈可心(1996-),女,硕士研究生,研究方向为深度学习;黄刚,教授,研究方向为计算机软件理论及应用。

格历史数据中学习以预测未来价格,搭建股票市场预测模型。常见的机器学习算法有 Logistic 回归、遗传算法、支持向量机等,并取得了不错的结果。而随着神经网络技术的兴起,通过搭建深度的神经网络来刻画股票价格并预测股票的走势,受到了人们的广泛关注,一些学者对这方面也展开了深入的研究^[1-2]。NairBB 等人^[3]基于决策树构建去噪混合股价预测模型。该模型首先对股票数据的相关特征进行特征提取,然后使用决策树算法对提取过的特征进行特征选择并使用 PCA 算法进行降维处理,降维后的数据输入到模糊模型中预测股价。Ticknor 等人^[4]构建了一个贝叶斯神经网络模型,不需要对数据进行预处理操作和周期分析,仅把市场价格和技术指标作为预测模型的输入,来预测未来股票的收盘价格。苏治等人^[5]构建了一种通过遗传算法对数据进行降维优化的 SVM 模型,采用量化选股方法分别从短期和中长期对其选股性能和预测精度进行了实证分析。程昌品等人^[6]采用了先对股票价格序列使用小波分解,分离出非平稳时间序列中的低频信息和高频信息,然后对高频信息构建 ARIMA 模型,对低频信息使用 SVM 模型进行拟合的方法,得到了较好的结果。郝知远等人^[7]基于舆论情报数据并进行自然语言处理及挖掘的建模预测分析的研究工作。其主要方法是依据最大化收益思想,提出了根据 ROC 曲线下的面积 AUC 值进行遗传参数寻优的支持向量机,解决传统方法在预测中可用性不高的问题。韩山杰等人^[8]基于谷歌人工智能学习系统 TensorFlow,构建多感知器 MLP (multi-layer perceptron) 神经网络模型,用于预测每日收盘股价。并就股价预测问题将 TensorFlow 与传统 BP (back propagation) 神经网络进行性能对比。Chen K 等人^[1]分析在加入不同数量的特征及不同的数据预处理状况下,使用长短期记忆网络 LSTM 对预测结果的影响;与随机预测方法相比 LSTM 模型提高了股票收益预测的准确率。王子玥等人^[9]使用 LSTM 进行股票价格的预测,提出变步长集成方法及改进的 MSE 损失函数,预测上能取得较为可观的提升,但未得出通用的最优步长范围。Minh DL 等人^[10]提出了一种利用财经新闻和情感词典预测股票价格走向的框架,将股票价格趋势预测的双流门控循环单元 (TGRU) 和在股票新闻和情绪词典上训练出 Stock2Vec 嵌入模型相结合。

不同于现有方法的是,文中提出了一种 CStock 量化选股模型,利用 contextual long short-term memory (CLSTM) 以及 bi-directional long short-term memory (BiLSTM) 结合股票相关的新闻信息和已知股价走势信息,从而有效地对股票走势进行预测。

CLSTM 和 BiLSTM 模型起源于循环神经网络^[11]

(recurrent neural network, RNN)。RNN 是一种节点定向连接成环的人工神经网络,可以利用它内部的记忆来处理任意时序的输入序列。传统的 RNN 存在梯度消失和梯度爆炸问题,因此, Hochreiter 等人提出了一种基于 RNN 的优化,即长短期记忆网络 (long short-term memory, LSTM), 一种时间递归神经网络, 常用于处理和预测时间序列中间隔和延迟相对较长的重要事件。

LSTM 在传统 RNN 的基础上,通过添加门控,使其变成门控 RNN,可以有效减少梯度消失(爆炸)等问题。LSTM 循环网络除了外部的 RNN 循环外,还具有内部的“LSTM 细胞”循环。其门控包括输入门、遗忘门和输出门。LSTM 网络比传统 RNN 更适合学习长期依赖,即可以减少梯度消失(爆炸)等问题。对于股票这类带有很强时间序列特性的数据,选择循环神经网络可以更好地结合历史信息。

BiLSTM 则是将两个不同方向的 LSTM 结合,形成双向循环神经网络,以同时提取数据的正、反向信息。而相较于 LSTM, BiLSTM 能同时利用两个方向上的时序信息,更容易挖掘出潜在模式。CLSTM 是 Shalini Ghosh 等人在 2016 年提出的一种基于话题的 LSTM。在 LSTM 的基础上, CLSTM 考虑了不同的话题下,输入门、输出门、遗忘门的权重状态。使得 LSTM 关注到话题之上,给 LSTM 一种指导。

1 选股模型

现有预测股票市场的模型主要是利用了股票市场中的交易数据。而影响股票市场波动的因素有很多,比如与股票相关的财经新闻或政治事件等,这些股票市场数据以外的信息都会影响到市场的波动。随着文本挖掘技术^[12]的出现,使得获取相关文本数据来预测股票市场走势成为现实。

文中运用了文本挖掘和深度学习的相关知识,结合嵌入式词向量技术^[13],采用 Bi-LSTM 双向循环神经网络、CLSTM 和 CNN 卷积神经网络对和股票有关的时序数据、文本数据进行分析,挖掘数据的深层次特征,构建选股模型。

文中对股票的数据分为数值型数据和文本型数据两部分,各部分使用不同的网络结构。其中数值型数据包括开盘价,最高价,最低价,收盘价,变化率。文本型数据包括股票名称,股票相关新闻。

1.1 模型介绍

文中使用了一种 BiLSTM 和 CLSTM 相结合的模式,对股票走势进行预测。将文本型数据作为 CLSTM 的 Contextual 信息输入。同时,将数值型数据作为 BiLSTM 的输入数据。

1.1.1 字符型数据

(1) 新闻信息提取。

股票相关新闻信息描述了该上市公司的运营状态。因为新闻信息与股票走势存在较大相关性,人们常常根据新闻信息对股票走势进行预测。为了对股票相关新闻信息进行编码,对股票相关的新闻信息通过 tf-idf^[14] 的方法,提取出相关关键词集合 K 。使用 Embedding 的方法,将一个大小为 n 的集合 K 中的每一个词 w_i 映射为对应的词向量 w_i 。对于向量集合 w ,使用全连接神经网络,对其信息进行提取。

$$\vec{\text{new}} = a\vec{w} + b \quad (1)$$

其中, a 为每个词向量的权重大小, a 属于 R^n , b 为偏置向量的权重大小, b 属于 R^n , $\vec{\text{new}}$ 为新闻信息提取结果向量。

(2) 股票信息。

对每个股票进行 Embedding, 将其转化成对应的股票向量 S 。对字符型数据通过 Embedding 的方法, 使得其变为相应的词向量 v_c 。将生成的词向量送入 CLSTM, 将股票名称 (Name) 作为 Topic, 对股票新闻关键词 Keys 进行处理, 特别的, 这里使用 CLSTM 的输出矩阵作为输出, 得到相应的隐含层矩阵信息 h_c 。

$$h_c = \text{CLSTM}(\text{Embedding}(\text{Name}), \text{Embedding}(\text{Keys})) \quad (2)$$

将 CLSTM 的输出矩阵作为卷积层 Conv1D 的输入, 然后使用最大池化层 Maxpool1D 对卷积结果进行池化操作。采用最大池化的方法提取特征值的最大特征来代替整个局部特征并大幅降低特征向量的维度。处理后得到特征向量 h_a 。

1.1.2 数值型数据

文中对数值型数据 x 使用 BiLSTM 进行处理, 使用 BiLSTM 输出矩阵的最后一维作为其输出, 得到相应的隐含层信息 h_i 。

1.1.3 全连接层

通过对数值型数据得到的输出 h_i 和字符型数据得到的输出 h_a 进行连接, 通过全连接层进行计算, 得到输出 f_c 。

$$f_c = \text{Concat}(h_i, f_c) * W_{f_c} + b_{f_c} \quad (3)$$

其中, W_{f_c} 为全连接层的权重矩阵, b_{f_c} 为偏置向量。

1.1.4 Softmax 分类器

通过对全连接层的数据进行分析, 得到分类为股票上升的概率和股票下降的概率 P :

$$P = \text{Softmax}(f_c) \quad (4)$$

即, 完成分类。

1.2 模型设计

网络模型如图 1 所示。该模型的输入层包括数据

信息和新闻信息两大部分, 模型的主体部分首先使用 BiLSTM 对数据信息方面的特征分别进行处理, 将文本型数据作为 CLSTM 的 Contextual 信息输入, 而后将输出矩阵结合 CNN 再次处理。最后使用多层全连接神经网络对所有数据进行处理。

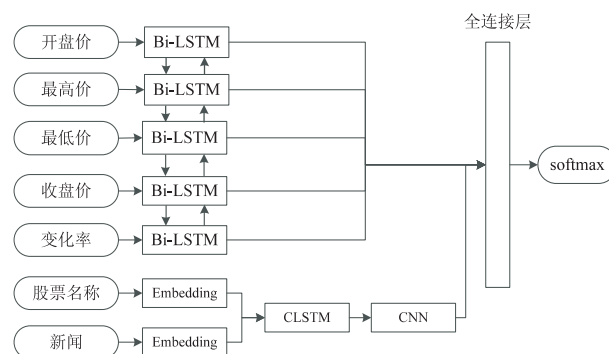


图 1 网络模型

1.3 选股策略

通过神经网络预测模型的输出, 可以得到每个股票上升的概率 P 。通过对 P 进行排序, 选出最高概率的 10 只股票 S_1, S_2, \dots, S_{10} , 通过对其概率进行求和。

$$\text{Sum} = \sum_{i=1}^{10} P(S_i) \quad (5)$$

然后按照如下公式进行选股, 对每股的投入 $\text{Inv}(S_i)$ 如下:

$$\text{Inv}(S_i) = P(S_i) / \text{Sum} \quad (6)$$

则每次投资的收益率为:

$$\text{Gain} = \sum_{i=1}^{10} \text{Inv}(S_i) G(S_i) \quad (7)$$

其中, $G(S_i)$ 表示 S_i 股当天的实际价格变化率。

1.4 原型系统

文中设计的原型系统主要包括四层, 首先是实时数据获取层, 接着是数据存储层, 接着是数据分析层, 最后是输出层。原型系统结构如图 2 所示。

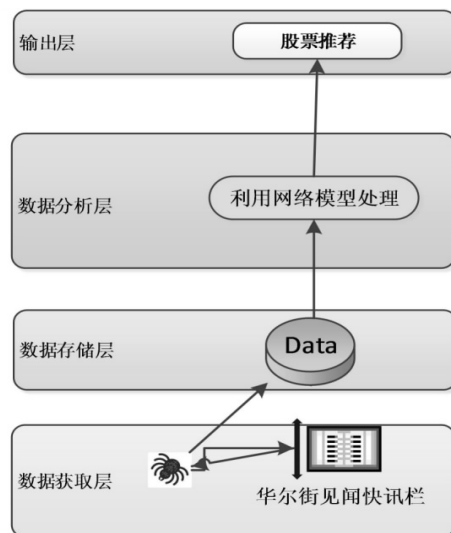


图 2 原型系统结构

1.4.1 数据获取层

网络爬虫 (web crawler), 又称为网络蜘蛛 (web spider) 或 web 信息采集器, 是一种按照一定规则, 自动抓取或下载网络信息的计算机程序或自动化脚本, 是目前搜索引擎的重要组成部分。狭义上理解: 利用标准的 HTTP 协议, 根据网络超链接 (如 <https://www.baidu.com/>) 和 web 文档检索的方法 (如深度优先) 遍历万维网信息空间的软件程序。功能上理解: 确定待爬的 URL 队列, 获取每个 URL 对应的网页内容 (如 HTML/JSON), 解析网页内容, 并存储对应的数据。

网络爬虫按照系统架构和实现技术, 大致可以分为以下几种类型: 通用网络爬虫 (general purpose web crawler)、聚焦网络爬虫 (focused web crawler)、增量式网络爬虫 (incremental web crawler)、深层网络爬虫 (deep web crawler)。实际的网络爬虫系统通常是几种爬虫技术相结合实现的。

通用网络爬虫: 爬行对象从一些种子 URL 扩充到整个 web, 主要为门户网站搜索引擎和大型 web 服务提供商采集数据。通用网络爬虫的爬取范围和数量巨大, 对于爬行速度和存储空间要求较高, 对于爬行页面的顺序要求较低, 通常采用并行工作方式, 有较强的应用价值。

聚焦网络爬虫, 又称为主题网络爬虫: 是指选择性地爬行那些与预先定义好的主题相关的页面。和通用爬虫相比, 聚焦爬虫只需要爬行与主题相关的页面, 极大地节省了硬件和网络资源, 保存的页面也由于数量少而更新快, 可以很好地满足一些特定人群对特定领域信息的需求。

文中设计的原型系统为聚焦网络爬虫, 对华尔街见闻快讯栏介绍的数据进行爬取。

1.4.2 数据存储层

系统实时地通过网络爬虫将数据存储到数据库中, 以便于模型对于数据的分析。同时经过一段时间对数据进行清理, 防止数据库出现内存不够的情况。

1.4.3 数据分析层

实时地将数据输入到已经训练好的模型当中, 本原型系统采用的是离线模型, 将已经训练好的模型直接用来分析数据。

1.4.4 输出层

输出模型最终分析的结果, 对用户进行展示, 引导用户选择模型分析得出的最优股票。

2 实验

2.1 数据集

文中利用网络爬虫来获取数据。爬取了华尔街见

闻快讯栏目中 2017 年 1 月 1 日至 2018 年 12 月 31 日的新闻标题及相应的发布时间作为财经新闻的初始样本数据, 同时从 Wind 数据库中获取了 2017 年 1 月 4 日至 2018 年 12 月 29 日中交易日的交易数据, 包括开盘价、最高价、最低价、交易量、涨跌幅。

实验使用美股数据作为数据集合, 取 100, 100, * 分别作为测试集, 验证集, 训练集的大小, 使用 train-development-test 模型进行训练。

2.2 设置

实验运行在 Ubuntu 16.04 操作系统上, 使用 Tensorflow, Python3 等工具, 设定网络 BiLSTM 和 CLSTM 的层数为 2, 隐含层大小为 64 (双向 128), 使用交叉熵作为损失函数。

2.3 数据表示

文中使用了数值型数据和文本型数据, 下面将分开讨论:

2.3.1 数值型数据

文中使用的数值型数据包括每股每日的开盘价、最高价、最低价、收盘价和价格变动率, 为了简化模型, 使用了近 50 天内的股票数据作为基础数据, 用于预测股票走势。

2.3.2 文本型数据

文中希望对股票及其相应的新闻进行提取, 从新闻中获取股票可能的走势信息, 如国家政策支持可能导致股票上升等。由于每条新闻过长, 因而采用了 Tf-idf (term frequency - inverse document frequency), 一种用于信息检索与数据挖掘的常用加权技术, 对每条新闻进行处理, 提取出新闻相关的关键词作为输入。为了和数值型数据对齐, 同样采用了近 50 天内的新闻数据同时结合相应的股票注册名信息作为输入, 用于预测股票走势。

2.4 数据输出

通过对上述数据进行训练, 预测该日股票的走势。在尝试了拟合股票走势和分类股票走势等方法之后, 采用分类方法进行实验, 即对股票走势进行二分类 (上升/下降) 来进行预测。实验表明, 将问题简化为分类问题比拟合股票走势准确率更高。

2.5 评估

准确率: 实验在测试集上预测的准确率 Acc (每股上升与否) 为:

$$Acc = 0.5232 \quad (8)$$

实验采用上述选股策略对测试集进行回测, 选择连续的一百天, 并除去回测当天数据小于 100 股的情况, 进行测试, 使用收益率作为衡量指标, 其结果如图 3 和图 4 所示。

每日收益率: 图 3 表示每次投资可获得的收益率

和天数的关系,其中横坐标表示天数,纵坐标表示收益率,点为每次投资的收益率,虚线为 0 坐标,其中收益大于 0.00 的天数为 Gdays,统计可得:

$$Gdays = 0.7041 \quad (9)$$

结果表明每天收益大于 0 的实际概率接近 0.7041。

总收益率:图 4 表示假设每天都采取模型给出的投资策略(这里忽略了手续费等支出),其收益和时间的关系,其中横坐标表示天数,纵坐标表示相比于第 0 天的收益率,其中最高点的坐标为(98,0.2849),即如果用此模型进行选股,那么在 98 天时,可以累计获得相当于本金 28.49% 的收益。

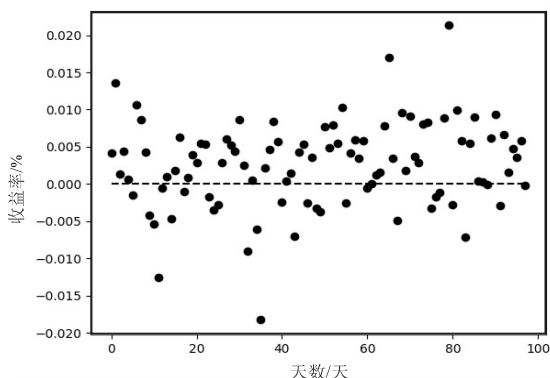


图 3 每日收益率

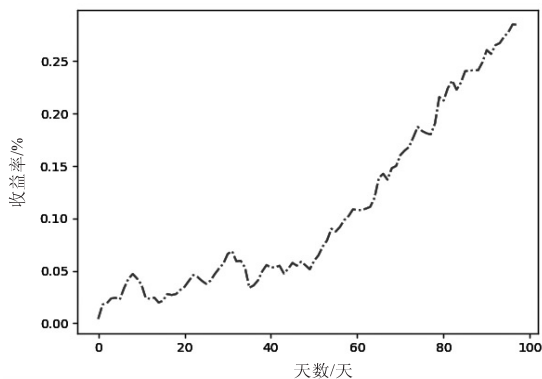


图 4 总收益率

3 结束语

文中提出了一种基于 LSTM 的神经网络模型,用于预测股票走势,同时给出了一种相应的选股策略。实验表明,考虑了数值型数据(开盘价,最高价,最低价,收盘价,价格变动)和字符型数据(股票名,新闻关键词)之后,可以获得较为不错的收益。虽然 CStock 对于股票走势预测是可行的,但是现实生活中还有很多影响股市的因素没有加入到实验的特征值中,如:股民情绪。下一步可以通过抓取投资者在 Twitter 上的讨论信息等的股评,利用 NLP 进行分析,这样可能会

得到更加精准的预测结果。在投资频率方面,将本模型运用于实际中时,交易手续费问题不容忽视,需要对预测的模型进行输出大小的修改,从而减少手续费的消耗。

参考文献:

- [1] CHEN K, ZHOU Y, DAI F. A LSTM-based method for stock returns prediction: a case study of China stock market [C]//2015 IEEE international conference on big data. Santa Clara; IEEE, 2015: 2823-2824.
- [2] NELSON D M Q, PEREIRA A C M, DE OLIVEIRA R A. Stock market's price movement prediction with LSTM neural networks [C]//2017 international joint conference on neural networks. Anchorage, AK, USA; IEEE, 2017: 1419-1426.
- [3] NAIR B B, DHARINI N M, MOHANDAS V P. A stock market trend prediction system using a hybrid decision tree-neuro-fuzzy system [C]//International conference on advances in recent technologies in communication & computing. Kottayam, India; IEEE, 2010.
- [4] TICKNOR J L. A Bayesian regularized artificial neural network for stock market forecasting [J]. Expert Systems with Applications, 2013, 40(14): 5501-5506.
- [5] 苏治, 傅晓媛. 核主成分遗传算法与 SVR 选股模型改进 [J]. 统计研究, 2013, 30(5): 54-62.
- [6] 程昌品, 陈强, 姜永生. 基于 ARIMA-SVM 组合模型的股票价格预测 [J]. 计算机仿真, 2012, 29(6): 343-346.
- [7] 郝知远. 基于改进的支持向量机的股票预测方法 [J]. 江苏科技大学学报: 自然科学版, 2017, 31(3): 339-343.
- [8] 韩山杰, 谈世哲. 基于 TensorFlow 进行股票预测的深度学习模型的设计与实现 [J]. 计算机应用与软件, 2018, 35(6): 267-271.
- [9] 王子玥, 谢维波, 李斌. 变步长 BLSTM 集成学习股票预测 [J]. 华侨大学学报: 自然科学版, 2019, 40(2): 269-276.
- [10] MINH D L, SADEGHI-NIARAKI A, HUY H D, et al. Deep learning approach for short-term stock trends prediction based on two-stream gated recurrent unit network [J]. IEEE Access, 2018, 6: 55392-55404.
- [11] 杨丽, 吴雨茜, 王俊丽, 等. 循环神经网络研究综述 [J]. 计算机应用, 2018, 38(A02): 1-6.
- [12] 翟羽佳, 王芳. 基于文本挖掘的中文领域本体构建方法研究 [J]. 情报科学, 2015, 33(6): 3-10.
- [13] 张琴, 郭红梅, 张智雄. 融合词嵌入表示特征的实体关系抽取方法研究 [J]. 数据分析与知识发现, 2017, 1(9): 8-15.
- [14] 姚海英. 中文文本分类中卡方统计特征选择方法和 TF-IDF 权重计算方法的研究 [D]. 长春: 吉林大学, 2016.