文章编号：2095-2163(2019)06-0055-05 中图分类号：TP391.41 文献标志码：A

基于卷积和循环神经网络模型融合的股票开盘价预测研究

赵浩博，李锡祚
（大连民族大学 计算机科学与工程学院，辽宁 大连 116000）

摘 要：本文提出了一种利用股票价格和相关新闻数据，基于卷积和循环神经网络模型融合的股票开盘价预测研究方法。针对股票开盘价预测的问题，考虑到股票相关信息的时序性以及新闻影响的持续性特点后，首先使用向量表示方法将新闻数据转换成向量，再利用卷积神经网络模型提取出股票相关的新闻文本特征，同时使用循环神经网络模型对股票价格数据进行训练，最后将新闻特征向量和价格训练后得到的向量合并，得到股票信息的低维向量表示并输入到深度神经网络中，利用深度神经网络对股票开盘价进行预测。本文实验中使用的数据是美股道琼斯指数与相关新闻，实验结果表明，本文所提出的方法在股票开盘价预测上具有明显的优越性。

关键词：股票开盘价预测；卷积神经网络；循环神经网络；深度学习

Forecast of stock opening price based on fusion of convolution and recurrent neural network models

ZHAO Haobo, LI Xizuo
(College of Computer Science and Engineering, Dalian Minzu University, Dalian 116000, China)

【Abstract】 This paper proposes a research method for stock opening price prediction with stock price and related news data based on convolution and recurrent neural network model fusion. In view of the issue of stock opening price prediction, after taking into account the timing of stock-related information and the persistence of news influence, the news data is first converted into a vector using vector representation, and then the convolutional neural network model is used to extract stock-related characteristics of the news text, and at the same time use neural network model to train the stock price data, and finally combine the news feature vector and the vector obtained after the price training recurrent toger the low-dimensional vector representation of the stock information and input it into the deep neural network, using the deep neural network forecast the opening price of the stock. The data used in this experiment is the US stock Dow Jones index and related news. The experimental results show that the method proposed in this paper has obvious superiority in the forecast of stock opening price.

【Key words】 stock opening price forecast; convolutional neural network; recurrent neural network; deep learning

0 引 言

金融市场是国家金融体系的重要部分，对于一、二级市场的参与者来说股票价格的分析预测是其做出正确判断与决定的重要参考，因此预测其价格也让大量的专家学者为之着迷[1]。在全球化的股票市场中，市场的行情与国家经济大环境、法律法规、企业经营情况、投资者信心、新闻舆情等都有所关联，股市行情具有高度的波动性与不确定性，使其成为金融与计算机领域研究中的一大难题[2]。

由于公司报表、报刊和舆论媒体等文本信息的快速增长与积累，可用于分析的数据样本也在逐渐丰富，数据数量也在不断地增加。在股票价格的预测中，如何使用文本数据来让模型的表现得到提升，在近些年的股市预测中一直是关注的热点。资本市场相关的数据信息通常可以反应股票价格波动，并且数据信息分析相比传统的 K 线分析更具有广度和深度。同时，随着 AI 领域的持续发展，机器学习和深度学习等人工智能技术在众多研究领域和实际场景中得到了广泛的应用[3]，自然语言处理领域也因为深度学习的兴起得到了发展和进步，这些技术上的突破均使得股票预测模型的建立有了更大的上升空间。

在过往的研究中，线性回归[4]、遗传算法[5]、SVM[6]、决策树[7]这些机器学习算法[8]以及深度学习网络模型都被大量用在股票预测的研究之中。在文献[9]中作者将多种机器学习算法与卷积神经网络(CNN)在股票预测中的表现进行了比较，证明了卷积神经网络模型在股票预测上的准确率优于传统的机器学习算法。而在文献[10]中，作者利用 tensorflow

作者简介：赵浩博(1994-)，男，硕士研究生，主要研究方向：机器学习、计算机技术；李锡祚(1963-)，男，博士，副教授，硕士生导师，主要研究方向：机器学习和推荐系统。

收稿日期：2019-09-08

框架搭建了多层神经网络(MLP)来对股票的价格进行预测,最终通过与传统的 BP 神经网络方法对比,说明了合适的神经网络结构有利于提高网络模型预测的准确率,同时还能有效减少预测耗时。

基于深度学习在股票预测中的优良表现和循环神经网络在序列数据预测中的特殊性,本文提出了一种基于卷积和循环神经网络模型融合的股票开盘价预测研究方法。在股票的数据选取方面包含了历史价格和相关新闻,新闻的特征提取用到了 word2vec[11] 和 CNN[12] 方法。在训练模型上,由于股票价格是时间序列数据,具有时序性,同时新闻对股价的影响具有持续性,所以本文采用的训练模型是卷积神经网络和循环神经网络。

1 相关技术

1.1 Word2Vector

在神经网络等机器学习和深度学习模型中,无法直接处理字符串类型的数据,因此需要将其转换为纯数字信息。在转换过程中,应尽可能保留数据原始信息。

Word2Vector 与 One-hot 类似,是一种将文本数据转换为矢量的模型,广泛用于自然语言处理(NLP)中。One-hot 对文本中的所有单词进行计数,然后对于每个词汇表编号,为每个单词创建 N 维向量。向量的每个维度代表一个单词,因此对应的数字位置中的维度值为 1,其它维度均为 0。虽然此方法保留原始单词信息,但在文本数量多的情况下维度太高,而且不能反映两个词之间的关系。例如,猫和小猫明显比猫和珊瑚更接近,但其却在单词向量表示中无法得到体现。相比于 One-hot 的编码方式,Word2Vector 通过学习文本,使用单词向量来表示单词的语义信息,通过将单词向量"嵌入空间"(嵌入就是将原始单词所在的空间映射到新空间),达到语义相似的单词之间距离接近的目的。这样便可以降低维度并反映单词和单词之间的关系。

在 Word2Vector 方法中,主要有 Skip-Gram 和 CBOW 两种模型。从直观上理解,CBOW 的做法是,将一个词所在的上下文中的词作为输入,而词本身作为输出。Skip-Gram 的做法和 CBOW 刚好相反,其将一个词所在的上下文中的词作为输出,而词本身作为输入。具体情况如图 1 所示。

1.2 卷积神经网络(CNN)

卷积神经网络(CNN)在计算机视觉领域取得了极大的进展,与此同时,CNN 开始应用于自然语言处理(Natural Language Processing)的各种任务,也逐渐在自然语言处理领域占有了重要的地位。

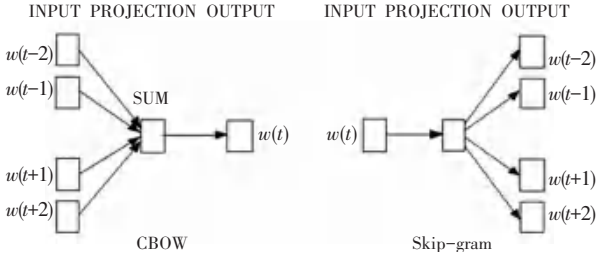


图 1 CBOW 和 Skip-Gram 模型结构图

Fig. 1 CBOW and Skip-Gram model structure

之所以用 CNN 来进行自然语言处理的工作,是因为其解决了传统词袋模型和连续词袋模型句子中词语的顺序被忽略、训练参数非常大的问题。在图像中卷积核通常是对图像的一小块区域进行计算,而在文本中,一句话所构成的词向量作为输入。每一行代表一个词的词向量,所以在处理文本时,卷积核通常覆盖上下几行的词,所以此时卷积核的宽度与输入的宽度相同。通过这样的方式,就能够捕捉到多个连续词之间的特征,并且能够在同一类特征计算时共享权重。

1.3 循环神经网络

循环神经网络(RNN)提出后,被广泛用于分析预测序列数据[13],但经过大多数学者研究发现,随着时间的推移 RNN 模型会存在忘记之前状态信息的问题,之后便提出了长短期记忆循环神经网络(LSTM)[14]。LSTM 是一种时间递归的神经网络,由于其特殊的模型结构,使得 LSTM 具有适合处理和预测时间序列中间隔和延迟较长的重要事件的特性。LSTM 的网络结构采取控制门的机制,其核心结构是由 3 个门构成,分别是遗忘门、输入门和输出门。LSTM 的关键在于运行在上方的细胞状态,这是其能保留记忆的原因。具体结构如图 2 所示。

LSTM 模型的计算原理如下:

首先,计算的是输入门 i_t 的值和在 t 时刻输入细胞的候选状态值 $\widetilde{C_t}$,公式如下:

$$i_t = \sigma(W_i * [h_{t-1}, X_t] + b), \tag{1}$$

$$\widetilde{C_t} = \tan h(W_c * [h_{t-1}, X_t] + b), \tag{2}$$

其次,要计算的是在 t 时刻,遗忘门的激活值 f_t,公式如下:

$$f_t = \sigma(W_f * [h_{t-1}, X_t] + b), \tag{3}$$

通过上面两个步骤的计算,就可以得到 t 时刻细胞状态的更新值 C_t,公式如下:

$$C_t = i_t * \widetilde{C_t} + f_t * C_{t-1}, \tag{4}$$

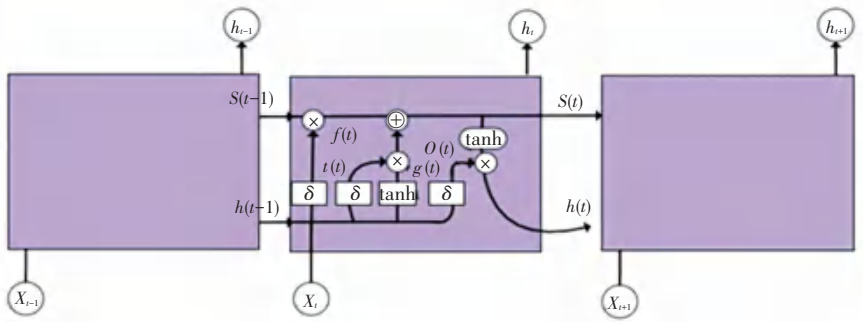


图 2　LSTM 网络结构图

Fig. 2　LSTM network structure

得到了新的细胞状态之后，就可以进行输出门值的计算，公式如下：

$$O_t = \sigma(W_o * [h_{t-1}, X_t] + b), \tag{5}$$

$$h_t = O_t * \tan h\ (C_t). \tag{6}$$

其中，W_i、W_c、W_f、W_o 是 4 个不同权重。在 t 时刻下，f_t 表示遗忘门，目的是计算 $t-1$ 时刻的状态在 t 时刻保留多少。i_t 表示输入门，计算出该时刻的输入信息，其中，X_t 是输入的特征。O_t 表示输出门，通过计算可以得到下一层的输入值。$\widetilde{C_t}$ 表示细胞状态的候选集，h_t 表示输出特征，h_{t-1} 表示前一时刻隐层的输入特征。通过以上的计算，LSTM 模型就可以有效利用输入的时间序列数据来使其具有长时期记忆的功能。

2　模型构建

本文所使用的模型是基于 CNN 和 LSTM 的融合模型。因为股票历史交易信息具有时序性，是属于时间序列数据，同时相关新闻对股票价格的影响具有持续性。综合考虑以上几点，本文采用 CNN 和 LSTM 作为训练数据的主要模型。模型框架如图 3 所示。

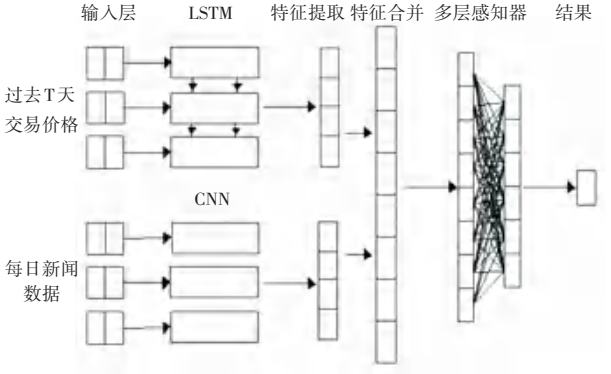


图 3　基于卷积和循环神经网络的股票开盘价预测模型

Fig. 3　Stock opening price prediction model based on convolution and recurrent neural networks

2.1　输入层

该模型的输入层由两部分组成：每天的新闻标题（News），以及过去连续 T 天的交易信息（Price）。

2.2　隐层

模型的隐层分为 2 部分：首先将清洗后的新闻标题进行词向量嵌入，将得到的新闻数据输入到 CNN 模型中并进行 2 次卷积和池化的操作，之后经过 Flatten 层压平，再进行全连接的计算，从而得到新闻数据的特征向量。同时，另一边的 LSTM 模型接收股票价格的数据，经历 3 个隐藏层的计算之后把数据输入到全连接层，最终得到价格的特征向量。最后，将两个模型得到的特征向量进行合并输入到新的全连接网络模型中，并使用 Dropout 方法来防止模型出现过拟合问题。

2.3　输出层

因为本文研究的目标是股票每日的开盘价格，属于回归问题，所以最后网络模型的输出结果为一维向量。

3　实验

3.1　数据集

本文所选用的数据是美股道琼斯指数和相关新闻的数据集，包括从 2008 年 8 月 8 日到 2016 年 7 月 1 日近八年的股票交易信息和相关新闻数据，其中新闻数据共有 73 609 条。实验中将股票的数据按照交易时间排序，其中前 70% 的股票数据作为训练集，后 30% 的数据作为测试集。

3.2　模型实现

本实验在 Windows 系统环境下，使用 Python3.6 作为编程语言，开发工具使用 JetBrains PyCharm 和 Anaconda3，运用 Keras 构建网络模型结构，底层应用 Tensorflow 框架。实验相关参数设置见表 1。

表 1 实验相关参数设置

Tab. 1 Experimental related parameter settings

相关参数	值
迭代次数	200
处理历史价格的神经元个数	80
处理历史价格的隐层个数	3
词向量维度	16
初始学习率	0.001
Dropout	0.7
卷积核大小	3 * 16
卷积核个数	8

3.3 实验流程

使用卷积和循环神经网络融合模型进行股票开盘价预测的实验具体如下：

(1)对新闻文本进行预处理操作，如：分词、去除停用词等；

(2)应用 Word2Vec 模型生成 16 维度的向量矩阵(词向量嵌入)；

(3)将向量输入到 CNN 模型中，进行特征提取操作；

(4)将股票价格数据输入到 LSTM 模型中，进行特征提取操作；

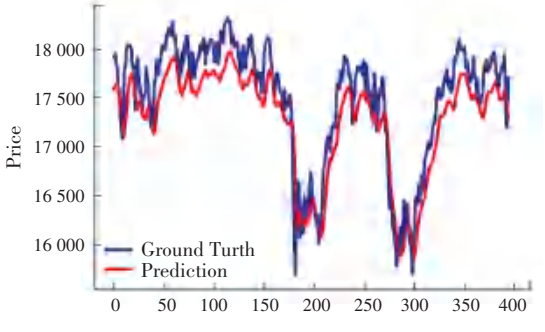
(5)将新闻特征向量和价格特征向量进行合并；

(6)利用深度神经网络训练合并后的向量；

(7)生成股票开盘价的真实值和预测值之间的对比图像，计算出模型预测的误差值。

3.4 实验结果

(1)首先进行的是仅用股票价格数据作为唯一输入特征的实验。其实验数据、训练集、测试集划分方式和所有的实验参数都与之后进行的实验相同。本实验主要是探索单一的股票价格数据输入到 LSTM 模型中的预测表现情况。具体结果如图 4 所示。



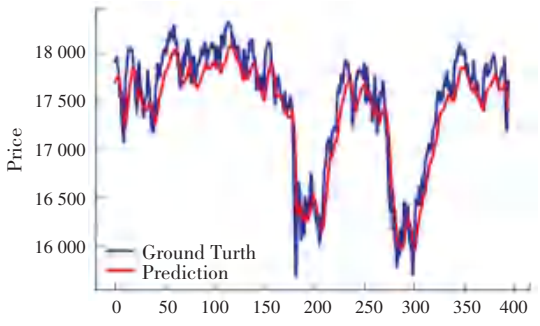
RMSE:299.77186824215084, MAE:262.93794140435614,
MAPE:1.499983938339244

图 4 单一股票价格数据实验结果

Fig. 4 Single stock price data experiment results

从图 4 的结果可以直观看出：模型的预测值和真实值存在一定的差距，并非十分理想，预测值普遍低于真实值；

(2)其次就是本文所提出的方法：利用股票价格和相关新闻数据，基于深度学习的股票开盘价预测。本实验主要是探索在添加相关新闻的情况下，LSTM 模型的预测情况是否优于单一的价格预测。具体结果如图 5 所示。



RMSE: 240.19255639947514, MAE:198.33494858238635,
MAPE:1.1384400051927723

图 5 基于卷积和神经网络模型融合的股票开盘价预测实验结果

Fig. 5 Experimental results of stock opening price prediction based on convolution neural network model fusion

由实验结果可以看出，利用股票价格和相关新闻的数据进行股票开盘价预测时，模型的精准度得到了很大地提升。其中 RMSE 与 MAE 分别减少了近 60 和 65 不等，MAPE 也有所下降，说明本文的方法对股票开盘价的预测准确度更高。因此可以说明，股票历史交易信息和相关新闻同时决定了股票的开盘价格。在模型计算复杂度上，CNN 和 LSTM 网络采用权重共享的方式，大大减少了网络中需要学习的参数个数，使其计算复杂度也随之下降。

4 结束语

对于股票投资者来说，如果能够知道未来的大盘价格和走势，就能为其股票选择提供有意义的参考价值。本文提出了一种利用股票价格和相关新闻数据，基于深度学习的股票开盘价预测方法。通过对股票相关新闻进行处理并利用卷积神经网络模型对新闻特征进行提取，充分利用所获取的数据信息，最后再将新闻特征和价格特征进行合并、拼接，共同对股票的开盘价进行预测。通过与单一的价格作为输入进行预测的实验进行分析和比较，证明本文所提出的股票开盘价预测方法有明显的优越性。

在今后的工作中，还会考虑到以下几个方面的

(下转第 64 页)