深 圳 大 学

本 科 毕 业 论 文（设计）

题目: 基于 LSTM 的股票价格预测方法研究

姓名: 武子越

专业: 信息与计算科学

双学位专业：计算机科学与技术

学院: 数学与统计学院

学号: 2015190160

指导教师: 陈剑勇

职称： 教授

2019 年 4 月 21 日

深圳大学本科毕业论文（设计）诚信声明

本人郑重声明：所呈交的毕业论文（设计）,题目《基于 LSTM 的股票价格预测方法研究》 是本人在指导教师的指导下,独立进行研究工作所取得的成果.对本文的研究做出重要贡献的个人和集体,均已在文中以明确方式注明.除此之外,本论文不包含任何其他个人或集体已经发表或撰写过的作品成果.本人完全意识到本声明的法律结果.

毕业论文（设计）作者签名：

日期：　　年　　月　　日

目录

基于 LSTM 的股票价格预测方法研究

武子越 2015190160

【摘要】

在深度学习领域中,长短期记忆（Long Short-Term Memory,LSTM）网络非常适用于时间序列的分析与预测,而股票数据属于非常典型的时间序列数据.本文的目的在于通过 LSTM 算法去准确预测股票价格的走势,作为股票买卖的判断,帮助投资者达到盈利.首先,本文使用了股票历史数据作为 LSTM 网络输入,通过实验发现数据归一化和价格比例化可以提高 LSTM 模型的预测精度.其次,本文提出了多单元预测模型,通过实验发现该方法在预测股票中短期趋势上有很好的利用价值.最后,本文使用了两种基于 LSTM 预测结果的股票交易策略,发现它们都可以从股市中获取较为稳定的收益,尤其在 T+0 策略上的风险较低,说明其拥有非常好的商业价值.

【关键词】

股票预测；LSTM；深度学习；量化投资

1 引言

1.1 股票价格预测的意义及发展现状

1.1.1 量化投资背景介绍

长期以来,准确地预测出股票未来的价格是量化投资领域一直追求的一个目标.1949年至1974年间,由于许多投资者为了在金融市场中盈利,量化投资技术就此诞生[1].在随后过去的40年里,量化投资在国外发展迅速并且逐渐成熟.而在学术界中,量化投资并没有一个严格的定义,但可以将其笼统地概况为：量化投资是一种将投资者的投资智慧与思想具体转变为数学模型的过程,投资者利用该模型来判断金融市场的趋势,并使用计算机来决策和执行投资操作,从中获利.在过去的40年中,最为成功的量化投资案例当属华尔街的文艺复兴科技公司（Renaissance Technology Corporation）,该公司通过自己研究出来的数学模型,在1988年至2015净年均收益率约为40%,而同期标准普尔500指数平均收益为每年10.27%,巴菲特也只超出该指数6个百分点,而该公司却超出了几乎30个百分点.这样的成功说明量化投资是具有非常大的商业意义的.

在量化投资中,一般都是采用计算机去执行分析与决策,这样可以很好地克服人的主观情绪和心理对投资决策的影响[2].在过去,关于投资分析和决策相关的工作,都是交给人工去完成的,但是人是无法绝对理性的,在投资决断时候,大多数人都会受市场因素的影响,做出一些不够理性的操作.而通过对历史数据和当前市场数据进行分析预测的量化投资模型,在推出运行之后,一般情况下是拒绝人工干预的,这样可以将人类的感性对投资的影响最小化,整个投资过程更加理性和客观.

量化投资还有一个非常大的优势,其对信息能做出快速反应和决策.对于股票市场而言,很多机会都是稍纵即逝的,股票的涨停和跌停可能就在几秒内发生,人脑过慢的反应速度和犹豫

心理都有可能错失最佳的操作时机,而量化投资往往使用计算机根据当前实时的信息进行相应的快速操作,把握市场机会.其中在一方面最为成功的模型就是高频交易[3].其在极短的时间内,通过响应速度非常快的计算机对市场的变化做出反应并完成交易,从中通过几秒内的差价变化,获取利润.

1.1.2 LSTM 股票预测发展现状

在很多领域中,时间序列是非常常见的输入数据,如语言处理、电信号分析、语音识别、流量分析预测、天气预报,就业率失业率分析、通货膨胀分析等等.更抽象来说,时间序列数据是指在特定时间段内,按照时间顺序规则地进行的任何测量或观察序列.时间序列分析主要包括表达特征和建立数据模型这两种分析.金融市场中的各种活动通常具体为特殊类型的时间序列来表示和分析,股票数据就是其中的一种.

现有的股票量化投资方法主要是基于股票的技术数据、基本面等指标选出优良的股票,而后续的操作只能根据投资者自行判断,涨幅范围是无法较为准确的预测的.如果结合一些其他的技术,在允许的误差范围内直接预测股票的未来走势,帮助投资者判断未来股票的价格区间,让投资者获得更大的收益.

然而,传统方法对股票数据的预测和分析是非常困难的,主要是因为其噪声较多导致的.传统方法仅依赖于如线性回归和参数估计,无法准确地识别和捕捉重要的特征[4].

在过去几年中,一些初步的证据表明了机器学习技术能够识别金融市场数据中的（非线性）结构[5],并且由于深度学习的空前进步,许多科学领域利用其高精度性能为各种问题建立了有效的解决方案.深度神经网络（DNN）在许多其他应用领域表现出了卓越的性能,例如信号处理,语音识别和图像分类.因此,使用深度学习技术去解决金融时间序列的预测问题是一个很好的方案.

目前,已经有许多研究者将不同的人工神经网络（ANN）应用于时间序列预测.与计量经济学模型不同的是,人工神经网络没有严格的模型结构和一系列的假设,只要拥有大量的数据,就可以进行建模.其中深度循环神经网络（RNN）是许多研究者所采用的预测方法之一.这种技术可以记住先前的数据输入,同时根据当前的数据来学习知识.Hornik 等人提出了一种将 RNN 与计量经济模型相结合去预测股市趋势的方法[6],他们首先使用 ARIMA 模型提取数据,然后将这些数据输入到 RNN 中训练,最后预测台湾股市的趋势.他们表明,从 ARIMA 中提取的特征后的数据比不提取特征数据的结果更为准确.Hajizadeh 等人提出了两种混合模型去预测 S&P500 指数的波动性[7].他们将 ANN 与指数广义自回归条件异方差（GARCH）模型相结合进行了实验,结果表面两种混合模型都优于单一答计量经济模型,产生的测试误差较低.Kristjanpoller 等人也利了用广义自回归条件异方差（GARCH）模型和 ANN 融合的方法[8],提出了有关拉美市场波动率的预测模型,并指出该模型在平均绝对误差上优于 GARCH 模型.Lendasse 等人提出了基于径向基函数神经网络的 Bel 20 股票市场指数预测的系统[9].他们表明,该系统可以捕捉金融时间序列数据中的非线性关系.为了预测印度的股票市场, Perez 等人提出了 ANN 与随机森林和支持向量回归（SVR）相结合的混合模型[10], ANN-SVR 组合模型得到了最好的预测结果.王杰等人提出了一种用于预测的指数平滑混合系统[11],他们将 ARIMA 和反向传播神经网络相结合,在道琼斯工业平均指数开盘价和深圳综合指数收盘价上进行了测试.实证结果表明,混合系统的性能优于所有的单个系统的性能,可以提供更准确的预测结果.2015 年,Adhikar 和 Ratnadip 使用了一种综合方法预测股市[12],该方法结合了多种模型,包括 ARIMA 和 ANN 模型.跟据观察,他所提出的混合模型优于单独的预测模型.然而,

Agarwal 和 Sastry 通过两个线性模型：ARMA 和指数平滑模型,再结合 RNN 去预测股市[13].他们发现预测性能的提高主要归功于 RNN.

近些年来,长短期记忆（Long Short-Term Memory,LSTM）网络发展迅速,它是 RNN 的变形,在 RNN 的基础上通过保留先前的网络状态,提高了捕获长期依赖性的能力,而且使其能够处理更长的输入序列[14].这些性能的提高,让我们对于股票市场的预测问题成为可能.在国际上,许多学者指出长短期记忆（LSTM）网络非常适合解决各种关于时间序列数据的任务[15].LSTM 很好地解决了传统 RNN 的一些问题,它通过一些存储单元和门操作,有效地解决了长期依赖性问题,克服了梯度消失或梯度爆炸的问题.因此,在金融领域中,很多关于时间序列分析的研究中都是用 LSTM 网络来建立模型的.

在过去的基于神经网络的研究中,一些技术性的指标会预先处理好,与输入数据相结合后传入到神经网络中,以便于学习.然而,LSTM 则不需要预先计算好技术指标,它会根据最原始的数据,自动检测出与原始数据较相关的一些模式.

2015 年，LeCun 等人提出了一个基于 LSTM 的股票预测模型[16],并且用于预测中国的股票,他们使用了股票和指数的历史价格数据进行训练,最终结果证实了 LSTM 具有良好的预测能力,在一定程度上提高了预测精度.2016 年,Akita 等人将日本经济新闻的文本数据作为 LSTM 网络的输入[20],同时也输入了股票市场的历史数据,来预测 10 家上市公司的开盘价格,同时提出了基于预测的模拟交易策略,通过对比实验,他们发现基于文本数据的模型比只有股票数据的模型获利更高（1.67 倍）.2017 年，Nelson 等人开发了一种基于 LSTM 的股票预测方法[17],通过该方法进行了模拟交易,同时与多种方法（多层感知器、随机森林和伪随机模型）进行了性能比较,实验结果表明,基于 LSTM 模型的交易风险较低.同年,Bao 等人也使用了 LSTM 网络进行了股票价格的预测[18],其模型表现出较高的预测准确率,此外,他们发现可以将不同类型的时间序列数据输入到网络之中,以扩大可用的信息空间,预测结果更好.Li 等人从论坛帖子中提取投资者的情绪数据[19],并结合股票历史数据输入到网络中,来预测沪深 300 指数和每日大众的情绪,在他们的实验中,LSTM 模型的表现优于支持向量机模型,并且加入情绪数据后,可以显著提高第二天开盘价的准确率（从 78.57%提高到 87.86%）.

1.2 本文研究内容及工作安排

本论文主要使用 LSTM 网络作为训练预测的模型,同时将股票和指数的历史数据作为数据,去预测股票未来的价格.本文研究目标包括：1、不同数据处理对于模型结果的影响.2、评估多单元预测模型的准确度和价值.3、将预测结果与股票交易策略结合,分析其实际特点和价值.本文内容安排如下:

一、概括描述了本文选题背景及国内外发展状况.

二、介绍 LSTM 网络结构与原理.

三、详细介绍了本文提出的多单元预测模型.

四、接受本文中的数据及处理方法.

五、描述了本文使用的两种交易策略.

六、实验与分析.

七、总结与展望.

2 LSTM 网络简介

2.1 RNN模型

深度学习方法主要利用一组计算层提取特征,在输入数据中学习到相应的知识.前一层输出的数据是后续层的输入,输入数据被馈送到输入层,目标输出的数据最终由输出层生成.循环神经网络（RNN）是上述深度学习架构算法之一,除了需要对数据进行预处理外,它还使用当前的输入数据来学习网络权重.当网络接收到顺序输入的数据后,它会向前传递,计算得到的特征信息会保留在网络的隐藏层中.而训练算法是反向传播（BPTT）的变体,其将每个时间的步骤连接到前面的步骤之中,根据时间顺序或系列顺序来执行计算,然后修改前面隐藏层中的权重,从而获得知识. 图 1 示出了 RNN 的典型架构和其预测过程.

但是 RNN 有一些短板,当执行一轮循环训练时,前面距离较远一些的输入数据已经早被遗忘,无法被记忆,这会有可能使得一些信息无法识别,导致隐藏层权重较小,可能导致学习梯度小时,从而使得结果的精度下降,这时候我们需要一个可以处理长期依赖的算法来解决这个问题.

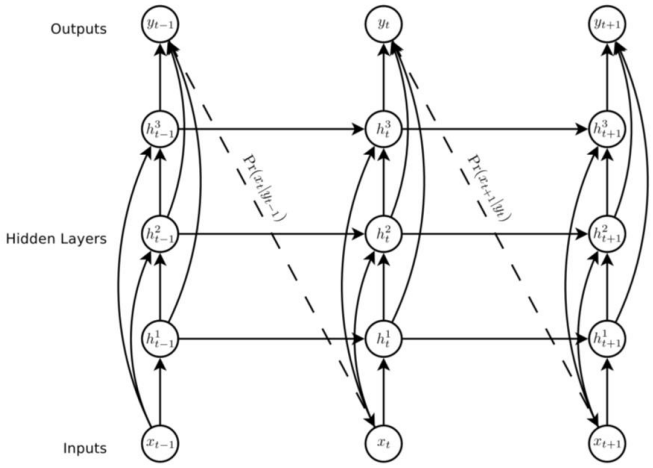

图 1 RNN 结构

2.2 LSTM模型

由于梯度的消失导致的无法长期依赖问题,目前已经开发了许多 RNN 的变体来解决这个问题,例如由 Hochreiter 和 Schmidhuber 提出的 LSTM 算法[6].它定义了一种可以通基于数据的重要性来阻止和传递信息的门控单元,并且通过反向传播的学习过程来估计允许存储或删除单元中的数据的权重.

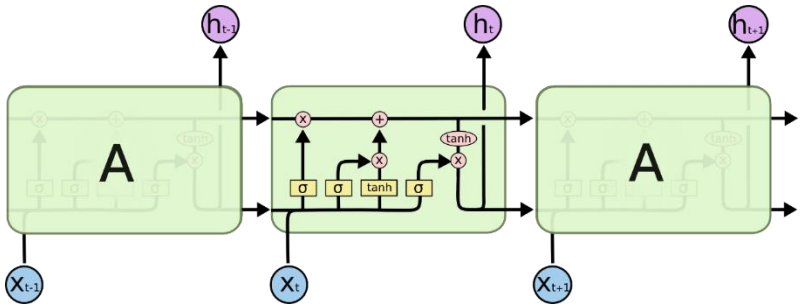

图 2 LSTM 基本结构

图 2 展示了 LSTM 的基本结构, 其中最为核心的是细胞的设置（图中最上面的平行流）, 其贯穿整个训练和预测过程,细胞状态相当于是一种数据的传送带,上面只有少量的线性交互, 数据信息在上面基本保持不变,而为了删除和添加一些信息,加入了一些门结构,其中包括三个主要的门,分别为输入门、遗忘门和输出门.

在训练过程中,LSTM 网络会在细胞流中选择一些信息丢弃,这个过程就在遗忘门中完成, 如下式：

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{1}$$

其中 x_t 表示当前轮神经元的输入, h_{t-1} 表示上一次神经元的输出, W_f 和 b_f 分别代表该层的权重和偏移, σ 为 sigmoid 函数,其公式为：

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{2}$$

遗忘门读取 h_{t-1} 和 x_t ,通过 sigmoid 函数映射到一个在区间[0,1]范围上的数值 f_t ,其中 0 代表全部信息舍弃,1 代表全部信息保留.得到 f_t 后与前一个细胞中的 C_{t-1} 进行相乘,从而丢弃一些我们确定要丢弃的信息

通过遗忘门后,需要决定该让多少新的信息加入感到细胞中,这里就设立了输入门来完成此功能,公式如下：

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{3}$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{4}$$

$$\tanh(x) = \frac{\sinh x}{\cosh x} = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{5}$$

其中通过 sigmoid 函数决定需要更新的信息i_t,然后另外一边通过 tanh 层生成出一个向量$\tilde{C}_t$,其为备选更新信息,然后将它们相乘后加入到刚刚经过遗忘门后的信息,对细胞状态进行一个更新,将其数值更新为C_t,如（6）所示.

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{6}$$

最后,需要确定输出的值,这个输出将会代表本轮训练最终的细胞状态,这里就需要将细胞中哪些信息输出,如下所示：

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{7}$$

$$h_t = o_t * \tanh(C_t) \tag{8}$$

上一轮输出h_{t-1}和本轮输入x_t一起通过 sigmoid 函数得到o_t,然后将本轮细胞中C_t进行 tanh 函数变换后与o_t相乘,最后得到本轮的输出h_t.

3 多单元预测模型

3.1 引言

越来越多的人类社会活动被数字化记录,形成了数字化社会.基于时间系列的数据是其中主要的数据类别.而通过深度学习等人工智能手段对已经发生过的历史数据进行学习,进而对即将发生的未来的数据进行准确预测,是对数字化社会进行科学管理的基础.因此如何对时间系列数据进行有效地学习和预测,特别是能够预测连续多个时间单元的数据,例如预测连续多天的数据,是其中重要的技术手段.

目前,对于某类时间序列的预测,一般都是通过历史数据输入到人工智能的神经网络之中,然后得到模型,而预测的时候会将不反馈误差的测试数据一个个输入到模型之中,得出结果,但是值得注意的是,这些模型本质上都是一个输入对应一个预测值.比如,在经典的多维度输入 LSTM 股票预测网络中,训练好的模型会根据输入的一系列股票测试数据,给出相对应的预测结果,该模型如果需要预测未来几天的价格,是必须要知道未来几天的输入数据的,但是实际生活中,每天只会更新一条股票数据,而并不能实现一个输入对应未来多天的数值.如果要将多维输入都去一个个预测,然后再输入到网络之中去预测后面的结果的话,会导致误差相当大,因为要想多个维度都预测准确是一件很难的事情.所以本文针对上面这一问题提出了一种新的训练预测方法,该方法在理论上可以根据一个单元的数据未来 N 个连续单元.

3.2 原理介绍

本人针上面所述的问题,提出了预测未来连续 N 个单元数据的方法,该方法由相同、但独

立被训练的智能算法组成.这里的智能算法不单指本文的 LSTM 算法,其适用于任何有关于时间序列预测的算法.方法中每个智能算法单独训练,预测未来某一个单元数据后,一起组成预测未来 N 个单元的数据.

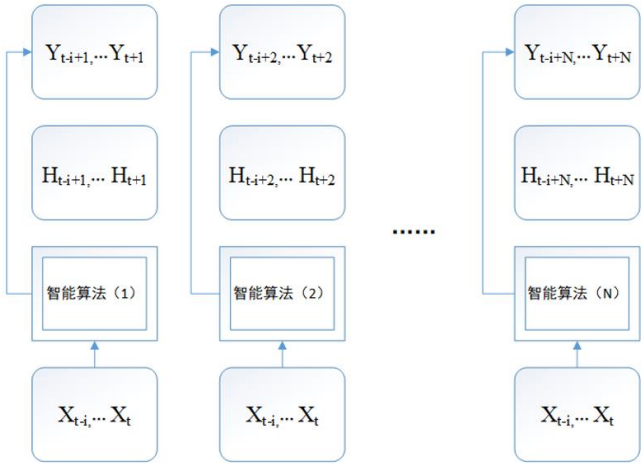

图 3 训练方法示意图

本方法的创新点在于用于训练和预测的时间序列的组织方式.对于预测未来第 1 个单元数据的算法,其输出标签是对应输入数据后 1 个单元的数据,即输入数据→标签：$x_t \rightarrow y_{t+1}$.对于预测未来第 2 个单元数据的算法,其输出标签是对应输入数据后 2 个单元的数据,即输入数据→标签：$x_t \rightarrow y_{t+2}$.对于预测未来第 N 单元数据的算法,其输出标签是对应输入数据后 N 个单元数据,即输入数据→标签：$x_t \rightarrow y_{t+N}$.

训练过程,如图 3 所示,其中$x_{t-i}, \ldots, x_t$为一维的输入数据（图中的数据可以拓展成多维数组,只需将每个输入变成多维的向量）,i 代表该输入数据的长度,t 代表最后一个时间数据单元的下标.H 代表标签值（实际数据）,当需要预测未来 1 天的时候,将输入数据对应的标签值设定为$H_{t-i+1}, \ldots, H_{t+1}$,预测未来第 N 天的时候,就将标签值设定为$H_{t-i+N}, \ldots, H_{t+N}$.Y 则代表输出的预测值.训练的时候,分别将$x_{t-i}, \ldots, x_t$输入到 N 个独立的智能算法中,预测出 Y,并且通过 Y 和 H 的误差计算后,修改算法中的模型的参数,最终可以得到 N 个训练好的模型.

最终通过训练,可以训练出 N 个智能模型,那么通过向这 N 个智能模型输入预测数据,可以分别预测出 N 个数据,这 N 个数据就组成了未来 N 天的预测结果.

3.3 预测方法应用

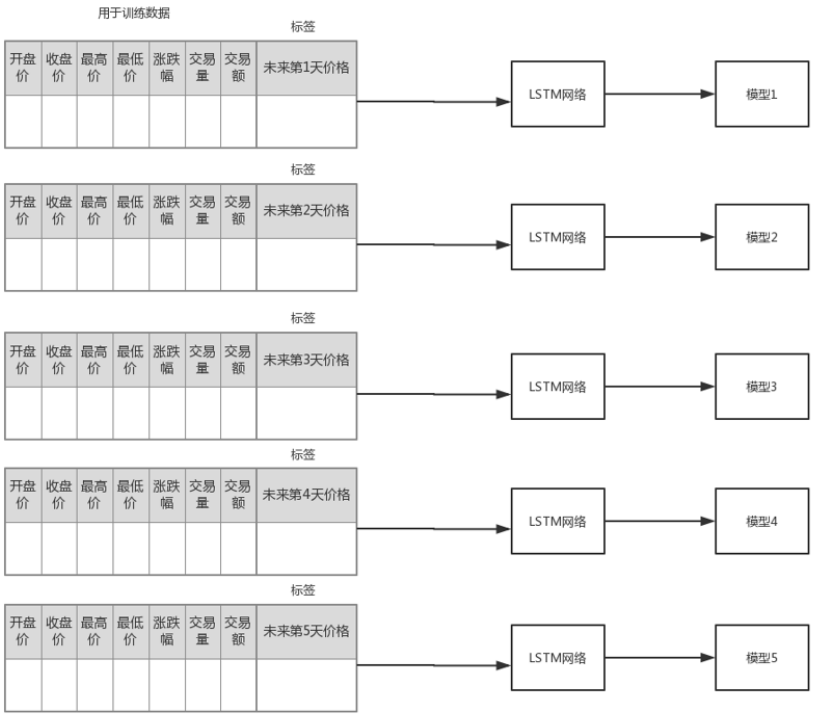

图 4 LSTM 预测未来 5 天价格的训练示意图

本文选择了预测未来股票 5 天的最高价和最低价来应用该预测方法.如图 4 所示,首先已知数据为 8 维的数据,8 个维度分别为开盘价、收盘价、最低价、最高价、涨跌幅、交易量、交易额、标签(未来第 1 天的价格),然后可以通过标签的调整,整理出 5 套数据,分别使用 LSTM 网络去训练数据,可以获得 5 个模型.并且这里每一天的预测都是一个新的模型,这样可以尽可能用更多的历史数据去预测结果.比如要预测 1 月 2 日的价格,那么 1 月 1 日之前的所有历史数据作为训练集,1 月 1 日的数据作为测试集.

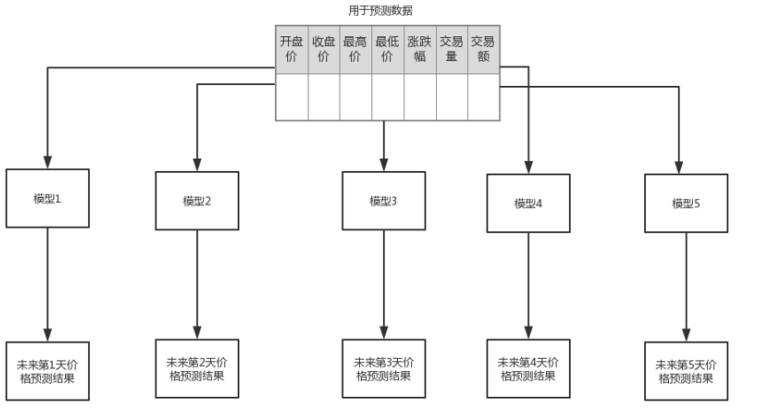

图 5 LSTM 预测未来 5 天价格的预测示意图

预测过程如图 5 所示,预测时将 7 维 X 的预测数据分别放入到训练出来的 5 个模型中,可以计算出 5 个结果,即未来第 1 天的价格至未来第 5 天的价格.这样就得到了关于一天的未来 5 日的结果.后续的所有实验都采用了这种方式来训练和预测.

4 数据

4.1 数据介绍

本文主要的研究对象为中国 A 股的股票,所以采用的数据都为中国 A 股股票和市场指数的历史数据.这些历史数据来源于网易财经和 Tushare 这两个网络数据平台,并且为了存储历史数据,构建了一个后台数据库,数据库每天在收盘后会自动更新当日股票和指数数据,图 3 说明了本文使用的股票数据的所有属性.而指数数据是一些股票的集合,所以其属性与股票数据的属性相同,比如指数的价格代表了该指数集合内所有股票价格的总和.

属性名称	介绍
日期	每只股票或指数一天只会有一条数据,所以日期就代表了数据的顺序.
股票代码	在 A 股市场中,每一支股票或指数都会有其独有的代码,一般为 6 位数.
股票名称	代表了上市公司或指数的名称,一般为 3 到 4 个汉字组成.
价格差	股票当前交易日与上个交易日的价格变动差,这里使用的是绝对值.
涨跌幅	股票的价格差除上上一个交易日的收盘价,这里使用的不是绝对值.
交易量	股票在一个交易日内的所有成交数量的总和.
交易额	股票当前交易日内的所有交易额度的总和,以人民币计算.
开盘价	股票在每个交易日上午 9.30 开盘时候的价格（集合竞价产生 ）.
收盘价	股票在当前交易日下午 15.00 最终收盘时候的价格.
最高价	股票在当前交易日之中产生的最高的交易价格.
最低价	股票在当前交易日之中产生的最低的交易价格.

表 1 股票数据属性介绍

在中国 A 股市场挂牌上市的公司共有 3000 多家,根据当前计算机的性能,是无法短时间内将所有上市公司的股票都去训练实验的,所以在有限的时间的前提下,我挑选了一些有代表性的股票和指数作为本论文中的实验数据,其中包括上证指数、东方财富、平安银行和贵州茅台.

上证指数,上海证券综合指数的简称,其包含了上海证券交易所全部上市的股票,最能反映中国经济状况的指数,其发布时间为 1991 年 7 月 15 日,所以目前的历史数据大概有 7000 多条.

东方财富,股票代码 300059,其为东方财富信息股份有限公司发行的股票,该公司是一家科技公司,于 2010 年 3 月 19 日在深圳证券交易所创业板挂牌上市,该股票价格波动较大,市盈率较高,目前的历史数据有 2000 条左右.

平安银行,股票代码 000001,其为平安银行股份有限公司发现的股票,该公司于 2012 年 1 月通过收购深圳发展银行的方式在深圳证券交易所正式挂牌.该股票属于金融类别的公司,市盈率较低,目前的历史数据有 1500 多条.

贵州茅台,股票代码 600519,其为贵州茅台酒股份有限公司发行的股票,该公司于 2001 年

8 月 27 日在上海证券交易所挂牌上市,属于白酒行业的公司,市盈率中等,市目前中国 A 股市场上价格最高的股票,目前的历史数据有 4000 多条.

本文通过分别使用上面这些股票或指数的历史数据进行实验研究,每支股票的板块、行业类别、市盈率特点都不同,可以更加全面的反映出模型的性能.

4.2 数据预处理

本文主要采用了两种数据处理的方式分别为：数据归一化处理和价格比例化.下面分别介绍两种处理细节.

4.2.1 数据归一化

数据的归一化（normalization）是指将数据按照一定的比例缩放,使其落入一个教小的特定区间之中.目前数据标准化方法有多种,归结起来可以分为直线型方法(如极值法、标准差法)、折线型方法(如三折线法)、曲线型方法(如半正态分布法).不同的方法,对系统的评价结果会产生不同的影响.而本文中所采用的为 min-max 归一化方法,式子如下：

$$X^* = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{9}$$

其中 x^* 代表归一化后的数据, x 代表原始数据, $X_{\min}$ 为原始数据中的最小值, $X_{\max}$ 为原始数据的最大值.这样操作可以将我们的股票数据映射到[0,1]区间中,使得数据精度提升.

4.2.2 价格比例化

在股票数据中有很多关于价格的数据,如最高价、最低价、收盘价、开盘价等,这些数据的范围一般比涨跌幅的数据范围大,并且该数据在一段时间内,在波动不大的情况下,数据的大小很可能是非常相似的,这可能会导致预测的时候会偏向给出与之前相似价格的结果,最终导致在股票大幅波动的时候无法准确预测.为了解决这个可能发生的问题,本文使用了一个原创方法,将价格变成比例后再输入到神经元当中,具体式子如下：

$$r_t = \frac{p_t - p_{t-1}}{p_{t-1}} \tag{10}$$

该式子中 p_t 表示第 t 天的股票价格（如最高价）, p_{t-1} 表示第 t-1 天的股票价格,最终算出 r_t 来替代 p_t 的位置,将其作为输入数据输入.这样的方法看似简单,但实际上可以起到很好的作用,使得预测的结果更为准确.

5 交易策略

5.1 基于1日预测的全仓交易策略

为了模拟收益情况,这里基于 LSTM 模型未来一日的预测结果,设置了一个全仓的交易策略.图 6 展示了该策略的具体流程,在交易过程中,一次迭代代表一天的操作,开盘前先判断当前是否持全仓,如果未持仓就尝试以预测第一日的最低价格全仓买入股票,如果预测第一天最低价的价格在实际当天出现过,则买入成功,否则无法买入,然后进入新一轮迭代,时间往后一天.如果一开始已持有全仓,就去判断预测第一天的最高价是否大于当初持仓买入的价格,如果大于则尝试以预测的最高价全仓卖出,卖出成功与否也取决于实际当天是否出现过这个成交价.最后进入一轮迭代,如此往复模拟.

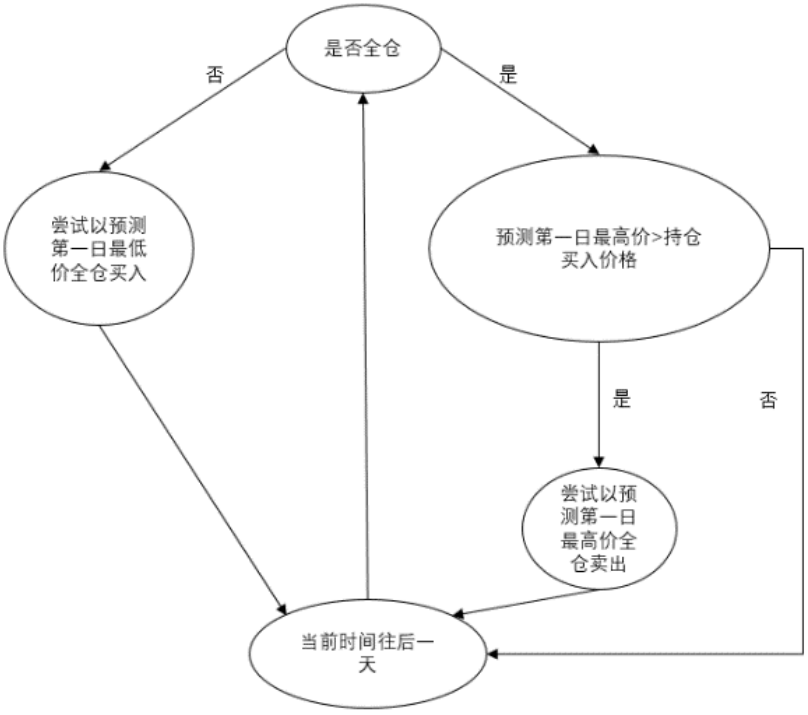

图 6 基于 1 日预测的全仓交易策略流程图

5.2 基于1日预测的T+0交易策略

在中国的股市中,当天买入的股票是无法当天卖出的,只能在第二天卖出,这是所谓的T+1规定,而一些投资者为了可以当天买入卖出,设计了 T+0 的交易策略.T+0 交易策略流程如图 7 所示,每笔交易都是按照半仓的量来操作的.首先迭代开始,判断当前是否持有全仓,如果是,则以预测的最高价尝试卖出半仓股票,注意这里交易成功与否与上面一个策略一样,都是取决于当天实际是否出现此交易价格.如果未持有全仓,则判断是否有半仓,如果没有半仓,就以预测的最低价尝试买入半仓股票.如果判断有半仓,分别尝试在预测最高价卖出半仓和在预测最低价买入半仓这两个操作.最后时间往后一天,进行下一轮迭代.

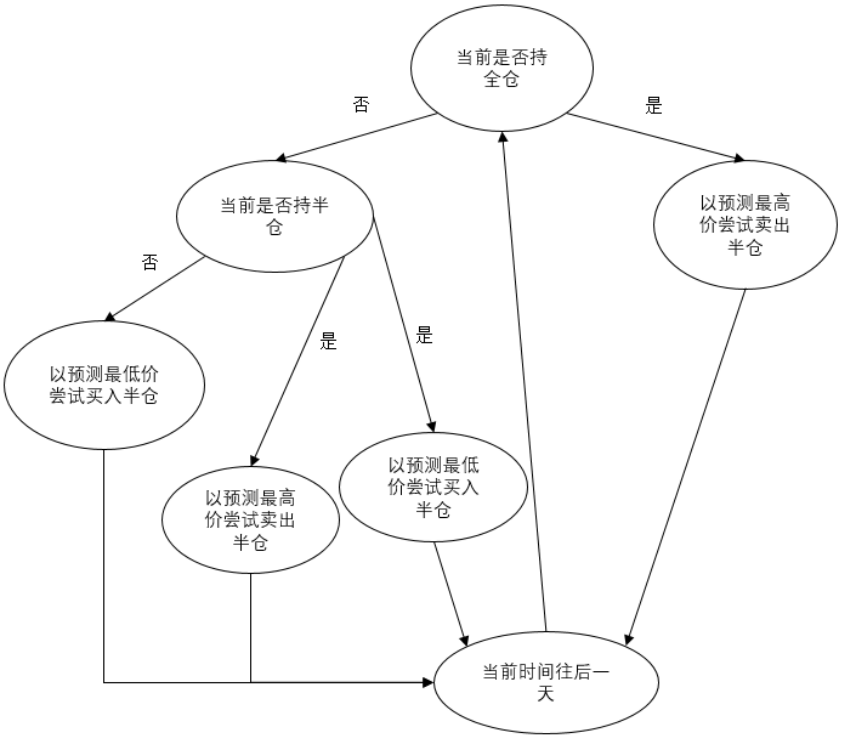

图 7 T+0 交易策略流程图

6 实验

6.1 实验环境

在第 4 节数据部分中提到我建立了中国 A 股股票数据库,从中选取了了上证指数、平安银行、东方财富和贵州茅台这四个数据作为实验数据,其中上证指数是最重要的基准数据,其是在上海挂牌上市的公司所有股价的总和,基本反应了中国这几十年的经济情况.对于编程环境,我使用了谷歌 Tensorflow 深度学习框架,并使用 Python3.6 版本实现了本文所描述的所有关于股票价格预测的方法.实验中采用的单天训练轮数为 2000 次,学习率为 0.0006,隐层数量为 10,这个参数配置是多次实验固定下来的较优配置,由于本论文不考虑调参数的对比,只研究不同方法对于结果的影响,所以所有实验的参数都统一为该配置.

6.2 评估标准介绍

本论文实验主要关注预测未来股票或指数价格,使用多维时间序列预测未来 N 天的方法去训练,这里为了减少计算量,一般我将 N 设为小于等于 5,即预测未来 5 天内的最高价和最低价.每次实验都为仿真实验,在预测历史上某天股价的时候,后面的数据是未曾加入到训练集的,并且在预测一天价格后,训练集会加入该天的真实数据,然后重新建立模型,不会再读取之前训练的权重.比如当前需要预测 3 月 1 日后股价未来 5 日的股价,这里会用 3 月 1 日之前的数据作为训练集去训练,然后将 3 月 1 日的当天数据作为测试集去预测,然后清零权重,训练集加入 3 月 1 日的数据,然后预测任务向后挪一天,清空之前训练的权重重新训练,将 3 月 2 日数据作为测试集,预测 3 月 2 日后未来 5 日的股价.然后使用这种方法记录每天任务的结果,最后给

出计算误差和评估.

实验用了多种评估指标：相对误差（MRE）、绝对误差（MAE）、均分根误差（RMSE）、判断系数（R^2）,其中 MAE 和 RMSE 越接近 0,说明预测效果越好,而判断系数更接近 1,表面预测结果更准确,判断系数只有超过 0.5 的情况下,才代表拟合有意义.这些指标的数学定义如下：

$$MRE = \frac{1}{n}\sum_{t=0}^{n-1}\frac{\left|y_t - \hat{y}_t\right|}{y_t} \tag{11}$$

$$MAE = \frac{1}{n}\sum_{t=0}^{n-1}\left|y_t - \hat{y}_t\right| \tag{12}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=0}^{n-1}(y_t - \hat{y}_t)^2} \tag{13}$$

$$R^2 = 1 - \frac{\sum_{t=0}^{n-1}(y_t - \hat{y}_t)^2}{\sum_{t=0}^{n-1}(y_t - \bar{y})^2} \tag{14}$$

其中y_t代表实际值,$\hat{y}_t$代表预测值,t的范围为[0,n),$\bar{y}$为实际值的均值.后续的实验中都采用了这些评估标准去综合比较.

6.3 数据预处理实验

6.3.1 数据归一化对比实验

预测	归一化	MRE	MAE	RMSE	R^2
第一天最高价	是	**0.039**	**0.511**	**0.047**	**0.828**
第一天最高价	否	0.043	0.605	0.056	0.769
第一天最低价	是	0.048	0.752	0.080	0.695
第一天最低价	否	0.053	0.847	0.132	0.541

表 2 东方财富预测结果

这里使用了东方财富的历史数据作为输入数据,在 LSTM 模型上进行了对比实验,如表 2 所示,其中可以看出有加入归一化处理的模型整体效果较好,其中预测未来第一天最高价的模型拟合效果最好,判断系数达到了 82.8%,在其他评估指标上也都表现较好.而未加入归一化处理的模型效果相对较差,预测第一天最低价格的模型预测效果最差,判断系数只达到了 54.1%,只比 50%高一点,参考价值不大.

所以通过对比得知,对于较大范围的数据,对数据预先进行归一化处理后再训练,可以在一定程度上提升模型预测的精度.在后续的实验中,都加入了数据归一化的处理.

6.3.2 价格比例化对比实验

我进行了价格比例化和非比例化的对比实验,这两者差别在于数据处理部分上.我使用了东方财富和上证指数的历史数据作为输入数据,将它们的最高价、最低价、开盘价和收盘价分别做了比例化,而非比例化的模型就直接输入的为实际价格.实验结果如表 3 和表 4 所示,从实验数据对比可以看出,在同一股票或者股指下,将价格数据比例化后再输入的数据总体误差较小,MAE、MRE、RMSE 都较小,判断系数更接近 1,说明将输入价格比例化后可以帮助股票和指数的预测更为准确.在后面的对比实验中,都加入了价格比例化作为数据的预处理操作.

预测	比例化	MRE	MAE	RMSE	R^2
第一天最高价	是	**0.023**	**0.323**	**0.033**	**0.940**
第一天最高价	否	0.039	0.511	0.047	0.828
第一天最低价	是	0.024	0.329	0.037	0.901
第一天最低价	否	0.048	0.752	0.080	0.695

表 3 东方财富预测结果

预测	比例化	MRE	MAE	RMSE	R^2
第一天最高价	是	**0.010**	**28.80**	**0.013**	**0.968**
第一天最高价	否	0.013	35.14	0.017	0.940
第一天最低价	是	0.011	31.85	0.015	0.956
第一天最低价	否	0.012	34.06	0.016	0.938

表 4 上证指数预测结果

6.4 股票未来5日价格预测

预测	MRE	MAE	RMSE	R^2
第一天最高价	**0.023**	**0.323**	**0.033**	**0.940**
第二天最高价	0.031	0.445	0.046	0.888
第三天最高价	0.039	0.574	0.057	0.836
第四天最高价	0.048	0.704	0.070	0.769
第五天最高价	0.053	0.781	0.078	0.721
第一天最低价	**0.023**	**0.329**	**0.031**	**0.941**
第二天最低价	0.029	0.300	0.044	0.878
第三天最低价	0.038	0.535	0.058	0.793
第四天最低价	0.047	0.647	0.067	0.755
第五天最低价	0.054	0.759	0.078	0.675

表 5 东方财富股价预测结果

上面的小节中的实验主要研究的是预处理方法,没有重点观察预测未来多天数据的结果,这里采用了多单元预测模型,用东方财富的历史数据作为输入数据,在 LSTM 网络上预测了

未来 5 日的最高价和最低价.其结果如表 5 所示,从结果可以看出,不管是最高价还是最低价,都是在第一天预测最为准确,往后预测的天数,误差逐渐越大,拟合效果也逐渐变差,这主要是因为越往后天数的价格与当前数据的相关性较小,而股票市场一个瞬息万变的市场,分秒间的突发信息都可能影响价格的走势,所以这里导致后面预测价格的误差较大,这属于意料之内的结果,尽管越往后效果越差,但是值得注意的是,预测第二天第三天的结果的 MRE 都在 0.04 以内,说明到第三天的价格还是较为准确的.往后两天的价格越来越差,最差的拟合的判断系数分别为 0.721 和 0.675,都大于 0.5,说明该拟合模型还是有一定的预测意义的.

在现实生活中,越往后的股票价格肯定是越难预测的,但是如果知道一个未来趋势和走向,是股票交易的一个很好的参考.

6.5 股票未来1日价格预测

通过上面的分析看出,对于未来第一天预测的各项评估指数都是非常优秀的,图 8 和图 9 展示了东方财富预测未来第一天最高价和最低价与实际价格的情况对比,横坐标为天数,总共 140 天.对比来看,蓝色的预测价格整体上与红色的实际价格较为吻合,并且在一些大幅上涨和下跌的时候也能做出大致趋势预测,预测结果非常优秀.

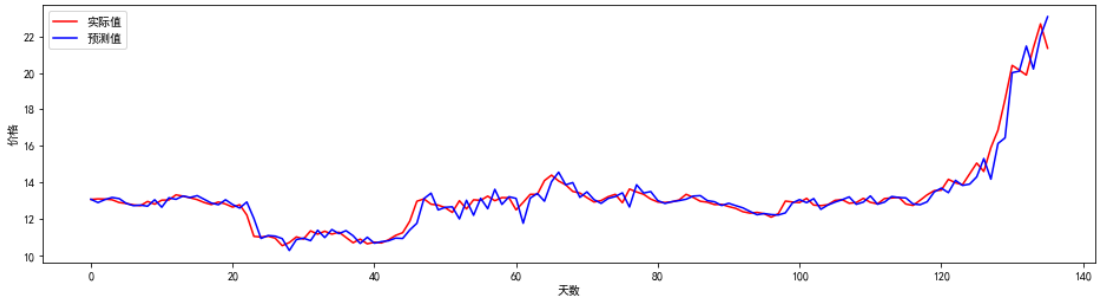

图 8 东方财富预测第一天最高价对比图

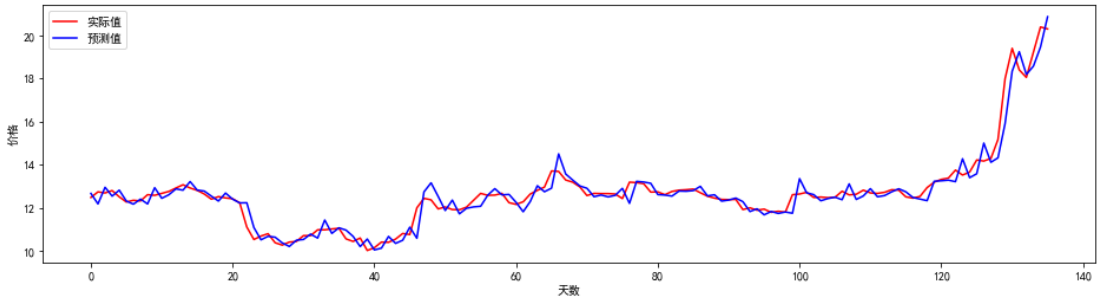

图 9 东方财富预测第一天最低价对比图

LSTM 在东方财富下的预测第一天的结果非常不错,接下来看看能否在其他股票上拥有同样的好结果.这里加入了平安银行和贵州茅台作为对比股票,表 6 给出了 LSTM 在这些股票数据下的预测结果,其中可以看出,股票走势较为稳定的贵州茅台拟合效果最好,最高和最低价的判断系数都达到了 97.5%左右的水平,平安银行的预测结果也不逊色,MRE 值都在 1.5%以下,说明整体误差较小.然而,上面分析过的东方财富的预测结果的R^2在三只股票里最小,这可能是因为东方财富相对其他两只股票而言,股价波动较大、整体资金较小,LSTM 可能在这类别股票上的预测结果没有其他稳定一些的股票的效果好.但是尽管如此,东方财富的预测结

果还是拥有一定价值的,因为其整体 MRE 控制在 2.5%以下,完全可以利用一些的交易策略,从中获得稳定的收益.

预测	MRE	MAE	RMSE	R^2
东方财富第一天最高价	0.023	0.323	0.033	0.940
东方财富第一天最低价	0.025	0.329	0.037	0.901
平安银行第一天最高价	0.015	0.175	0.020	0.954
平安银行第一天最低价	**0.014**	0.159	**0.019**	0.957
贵州茅台第一天最高价	0.015	10.52	0.022	0.975
贵州茅台第一天最低价	0.016	10.92	0.022	**0.974**

表 6 三支股票股价预测结果

6.6 股票交易策略研究

6.6.1 基于 1 日预测的全仓交易策略结果分析

在 5.1 中,本文提出了一种基于未来 1 日价格的全仓交易策略,这里结合之前 LSTM 预测的多支股票的结果进行了分别进行了模拟交易,其中包括上证指数、平安银行、东方财富和贵州茅台.这里每天的交易完全是模拟真实环境操作的,模拟中每笔交易都有计算 0.0001 的交易税和卖出股票时额外的 0.001 印花税,所以这里策略的收益率是算上了交易成本之后的结果.图 10 给出策略在各个股票上的收益情况,蓝色曲线代表策略收益率,红色曲线代表股票实际收益率,横坐标为天数.通过对比可以看出,经过 100 多天的交易之后,该策略最终在每只股票上都是盈利的,但是只有在平安银行和东方财富上的收益超出实际的股票收益,说明该策略可能在波动较大、整体上扬的股票下的收益会更比实际高.对于上阵指数和贵州茅台来说,收益率一直都处于实际收益以下,但是整体结果还是非常好的,内有出现较高的亏损，并且最终都是盈利的，也都基本接近了实际一半的收益率,并且在一些关键上涨点上决策较好.对于风险来控制来说,该策略容易在股票出现下跌的时候容易被套住,从东方财富和平安银行这两只股票中期的走势就可以看出,当股票出现下跌,收益率也几乎同一时间下跌,说明该策略的风险较高.

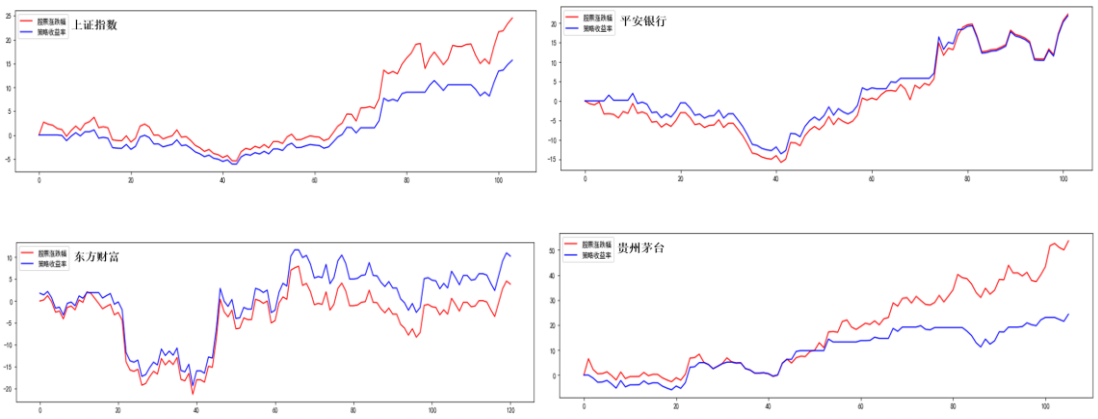

图 10 基于 1 日预测的全仓交易结果

6.6.2 基于 1 日预测的 T+0 交易策略结果分析

在 5.2 中介绍了基于 1 日预测的 T+0 交易策略,这里也在 LSTM 的结果上进行了实验,算上交易成本之后的各支股票的结果如图 11 所示.结果可以看出东方财富、平安银行、上证指数的收益率都比全仓交易策略的收益要高,只有贵州茅台的收益较低,这可能是因为 T+0 规则对于一直上扬的股票收益不是特别高,频繁的交易会导致上涨机会的措施.

从规避风险的效率来看,T+0 策略明显有很好的风险控制能力,四支股票最低亏损率都比全仓交易的要低的多,尤其是在东方财富和平安银行这两只股票上表现最好,它们中期都有超过 15%的亏损,而对于策略收益率,在东方财富上最高的亏损在 5%左右,而平安银行甚至只有 2-3%的最高亏损,整体策略的风险控制是非常好的.这说明 T+0 策略相比全仓策略来说,在几乎没有减少收益率的情况下,有效的控制了亏损风险,整体表现较好,是更加适合 LSTM 预测结果的交易策略.

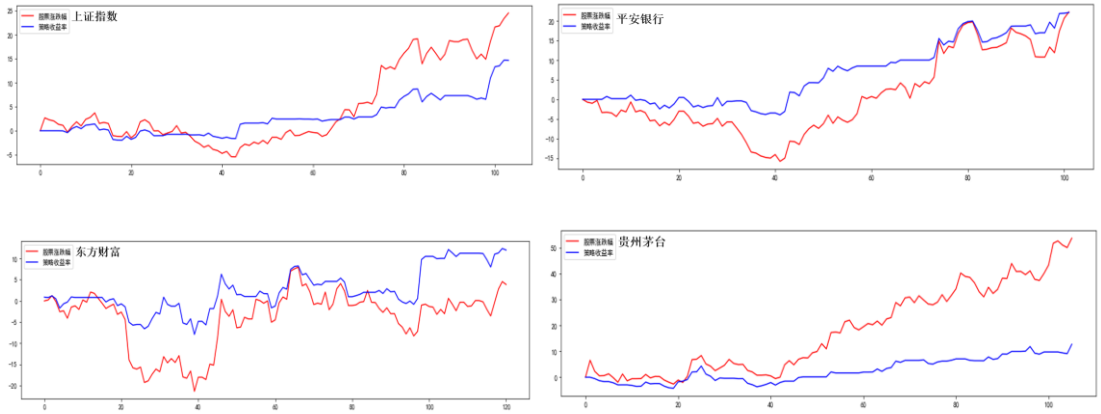

图 11 T+0 交易策略交易结果

7 总结与展望

本文主要概述了股票预测在实际生活中的商业价值,简单介绍了量化投资的背景和时间序列分析和预测的一些方法,同时也介绍了 LSTM 网络在股票预测领域中的国内外发展状况.与此同时,重点描述了 LSTM 网络的原理和核心结构,给出了一些相关的函数方程.重点介绍了本文中使用的股票数据的来源、属性和公司概括分析.本文提出了一种关于预测未来多个单元的时间序列训练方法,并给出了股票预测未来 5 日价格的应用案例.通过谷歌的深度学习框架 TensorFlow 和 Python 语言实现了 LSTM 网络,并利用该网络分别做了有关数据归一化、价格比例化、预测未来 5 日价格和预测未来 1 日价格的对比实验,根据一些统计指标分析表明,数据归一化和价格比例化都能在一定程度上提高预测的精度,预测未来 5 日的结果表面,越往后的天数的预测价格准确率越低,模型拟合程度越差,但是判断系数都超过了 50%,说明仍染有一定的使用价值.在预测未来 1 日的对比实验中发现,市盈率较高的东方财富拟合效果比平安银行和贵州茅台都要差,但是三者的判断系数均在 90%以上,其中贵州茅台效果最好,判断系数达到了 97.5%的水平.最后,本文根据各股票的未来第一天预测结果,设定了两种模拟交易策略,分别为全仓交易策略和 T+0 交易策略,从结果的收益率来看,两个策略的收益率都较高,拥有很好的商业价值,同时 T+0 交易策略的风险控制能力较低,在实际股票大幅度亏损的情况下,策略的亏损情况较少,尤其在平安银行和东方财富这两只股票上表现风险控制能力最好.

通过本文的实验结果表面,加入数据归一化、价格比列化后的 LSTM 网络股票预测模型,在一定的股票交易策略下,是可以从股市中获取较为稳定的收益的,说明其拥有非常好的商业价值.对于未来工作,本文中的策略只结合了未来一日预测的策略,而后面预测天数的结果还没有加入到策略中去使用,所以策略研究这是未来值得研究的一个方向.同时,本文中的数据采用的为常见的股票数据,如果加入一些技术指标的数据,也有可能对模型的精度有一定的提升,这些都是未来的重点研究方向.

【参考文献】

[1] Markowitz H M. Portfolio Selection[J]. Journal of Finance, 1952,7:77-91. https://doi.org/10.1111/j.1540-6261.1952.tb01525.x

[2] 陈健，宋文达. 量化投资的特点、策略和发展研究[J]. 时代金融, 2016(29):245-247.

[3] Durbin M. All About High-Frequency Trading: The Easy Way To Get Started[M]. McGraw-Hill Press，2010.

[4] Ahmed K, El-Alfy E, Mohammed S. Evaluation of bidirectional LSTM for short-and long-term stock market prediction[C]// Conference: 2018 9th International Conference on Information and Communication Systems (ICICS), 151-156. 10.1109/IACS.2018.8355458.

[5] Thomas F, Christopher K. Deep learning with long short-term memory networks for financial market predictions[J]. European Journal of Operational Research, Elsevier, 2018,270(2):654-669. https://ideas.repec.org/a/eee/ejores/v270y2018i2p654-669.html

[6] Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators[J]. Neural Networks, 1989, 2(5):359-366.

[7] Hajizadeh E, Seifi A, Zarandi M H F, et al. A hybrid modeling approach for forecasting the volatility of S&P 500 index return[J]. Expert Systems with Applications, 2012, 39(1):431-436.

[8] Kristjanpoller W, Fadic A, Minutolo M C. Volatility forecast using hybrid Neural Network models[J]. Expert Systems with Applications, 2014, 41(5):2437-2442.

[9] Lendasse A, Bodt E D, Wertz V, et al. Non-linear financial time series forecasting - Application to the Bel 20 stock market index[J]. European Journal of Economics & Social Systems, 2001, 14(1):81-91.

[10] Perez L, Wang J. The Effectiveness of Data Augmentation in Image Classification using Deep Learning[J]. 2017.

[11] Wang J , Wang J . Forecasting stock market indexes using principle component analysis and stochastic time effective neural networks[J]. Neurocomputing, 2015, 156:68-78.

[12] Adhikari, Ratnadip. A mutual association based nonlinear ensemble mechanism for time series forecasting[J]. Applied Intelligence, 2015, 43(2):233-250.

[13] Rather A M, Agarwal A, Sastry V N. Recurrent neural network and a hybrid model for prediction of stock returns[J]. Expert Systems with Applications, 2015, 42(6):3234-3241.

[14] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997,9(8):1735-1780.

[15] Lipton Z C, Kale D C, Elkan C, et al. Learning to Diagnose with LSTM Recurrent Neural Networks[J]. Computer Science, 2015.

[16] Lecun Y, Bengio Y, Hinton G. Deep learning[J]. Nature, 2015, 521(7553):436.

[17] Nelson, D M, Pereira A C, Oliveira R A. Stock market's price move- ment prediction with LSTM neural networks[C]// In Proceedings of the international joint conference on neural networks (IJCNN) (2017:1419–1426). IEEE .

[18] BAO Wei, YUE Jun, RAO Yu-lei, et al. A deep learning framework for financial time series using stacked

autoencoders and long-short term memory[J]. PLOS ONE, 2017, 12(7):e0180944-.(in Chinese)

[19] LI Jia-hong, Bu Hui, Wu Jun-jie. 2017 International Conference on Service Systems and Service Management - Sentiment-aware stock market prediction: A deep learning method[C]// International Conference on Service Systems & Service Management. IEEE, 2017:1-6.(in Chinese)

[20] Akita R, Yoshihara A, Matsubara T, etal. (2016).Deep learningfor stock prediction using numerical and textual information[C]//In Proceedings of the IEEE/ACIS fifteenth international conference on computer and information science (ICIS) (pp. 1–6). IEEE.

致谢

在本篇论文完成之际，我要非常感谢我的毕业论文导师陈剑勇教授，和实验室项目导师林秋镇老师，感谢他们在大学期间给我了进入实验室的机会，让我可以利用当前最前沿的技术去研究我最感兴趣的领域，并且在这个过程中学习到了很多宝贵的知识，让我的科研能力上升了一个台阶. 同时我也非常感谢他们给我的研究和论文提出的每一个意见，如果没有这些意见，我可能无法写出本篇论文，我谨在此致以最崇高的敬意和最真挚的感谢！

感谢计算机与软件学院给了我学习计算机知识的机会，拓展了我的视野，让我成为了更有能力的人. 最后我也要感谢我的父母，以及支持我的同学、朋友，感谢你们四年的陪伴，谢谢！

Research on Stock Price Forecasting Methods Based on Long Short-Term Memory

【Abstract】

In the field of deep learning, Long Short-Term Memory (LSTM) networks are very suitable for time series analysis and prediction, while stock data belongs to very typical time series data. The purpose of this paper is to accurately predict through LSTM algorithm. The trend of stock prices, as a judgment of stock trading, helps investors achieve profitability. First, this paper uses stock historical data as the input of LSTM network. It is found through experiment that data normalization and price-proportional can improve the prediction accuracy of LSTM model. This paper invented the training method for predicting the future for many days. Through experiments, it is found that the method has a good value in predicting the short-term trend of stocks. Finally, this paper uses two stock trading strategies based on LSTM prediction results and finds that they all It is possible to obtain relatively stable returns from the stock market, especially the low risk on the T+0 strategy, indicating that it has very good commercial value.

【Key Words】

Stock forecast; LSTM; Deep learning; Quantitative investment