基于先验知识和深度学习的股票价格预测

2272078 李硕

摘 要 股票价格预测的问题一直都是数十年来学者们致力于研究的问题。随着股票市场变得日益庞大，传统时间序列预测模型如 ARIMA 模型等等对股票价格进行预测已经不能取得非常好的效果。神经网络技术的出现带来了新的预测方法，其中的 LSTM 深度学习模型加入了三个门控，对于长序的时间序列数据有着更好的预测效果。注意力机制的引入也允许模型关注时序数据的更多特征。但股票价格往往与股评，新闻事件等存在一定的相关关系，而基于深度学习的方法并未很好研究影响股价数据的先验信息，故本文对股价数据先验数据的引入与编码做出研究，利用股评的金融情绪分析结果作为模型的先验信息，并在传统 LSTM 的基础上引入 CBAM 注意力模块，在爬取的一年的数据上进行训练与预测，实验证明模型预测效果优于传统基于时序数据的方法。

1 引言

股票的交易数据是经典的时间序列，在神经网络方法流行之前，很多研究者都使用了线性回归的方法比如 GARCH、ARIMA 等等对股票价格进行预测，并取得了不错的成果。但是这些方法仅适于预测平稳的时间序列数据，而股票数据的平稳性往往较差。其他的方法针对股票数据的非线性特性采取了数据差分的方式进行预处理，但是这些操作本身会造成原始数据的损失，传统线性模型在股票价格预测上仍有较大的局限性。

近些年，人工智能技术开始快速发展，机器学习方法逐渐变得流行，支持向量机，感知器模型等机器学习的方法被广泛的应用于股票价格预测。如今，随着深度神经网络的发展，深度学习模型逐渐取代传统机器学习方法：RNN 循环神经网络是第一个应用于时序预测的神经网络模型，但其存在记忆选择和梯度爆炸的问题。LSTM 神经网络设计了输入输出遗忘的门控单元解决了长序列数据被模型全部记忆的问题，以及循环神经网络出现的梯度消失和爆炸等等问题。与传统的时间序列模型相比，LSTM 神经网络由于其内部的非线性激活函数，在分析非线性相关数据方面具有一定的优势。在时序数据特别是股价数据中，往往在剧烈变化处反应了较大的信息量，故注意力机制的引入有利于模型给予平稳序列更少的关注而集中注意力于重要区域。

神经网络模型往往善于学习抽象的高维特征，而显式的人类先验知识往往对模式识别等过程具有强烈的暗示。股价数据作为人类生产生活过程的一种反映，天然与生产生活活动存在一定的关联。例如股评数据与股价数据往往存在较大的正相关关系，天灾人祸，金融丑闻等与股价数据则存在较大的负相关关系。金融情绪分析作为一种自然语言处理技术，允许针对文本数据进行其对金融数据的乐观或悲观程度的评价。在传统股价时序数据的基础上增加金融情绪分析结果，将有利于对股价进行更加精准的预测。

2 实验方法

2.1 金融情绪分析

2.1.1 金融股评与创业板指数据爬取

如今国内有大量包含股票评论信息的金融类的网站。股评信息潜移默化的影响着单只股票或者大盘整体的趋势。一方面股评有数据量庞大，无关数据多，文本大多比较短的特点。另一方面因为股评和投资人的利益紧密相关，所以大多数评论情感倾向明显，例如积极的情感评论有“明天肯定要上涨，一定要加仓”，而负面的有“感觉到最高点了，快跑啊，不要当接盘侠”等。

本文以东方财富网股评作为爬取对象。与个股相比，股指是由不同行业中的不同股票组成，所以股指一般会比其他股票的数据更加平稳，从而更能反映经济的整体势头和总体状况。因此选择选取创业板指数自 2021 年以来一年的数据作为样本数据，并加入投资者情绪作为新的特征参与神经网络的训练，考察股评先验对模型预测的结果的影响。本文爬取创业板指股吧 110000 条左右的帖子，帖子的时间范围为 2021 年 10 月 15 日至 2022 年 10 月 19 日，内容包括标题、正文以及时间数据。每天的帖子数量从几十条到几百条不等。

首先对爬取数据进行预处理，包括将空数据删

除，合并正文与标题，统一时间格式。最终剩下约106000 条数据。

	comment	time
0	是什么拉起了指数是什么拉起了指数	2021-10-15
1	又要大跌了，今天继续宁国指数大涨，个股大跌，下周一将大跌	2021-10-15
2	就我是？就我是？	2021-10-15
3	我们机构缴纳下半年去做价值投资去了我们几个都清仓创业板的垃圾股票，我们去做PE10倍的华泰证...	2021-10-15
4	如果不是相信国家我会坚持价值投资？！如果不是相信国家我会坚持价值投资？！	2021-10-15
...	...	...

图 2.1 股评爬取内容

2.1.2 LSTM

长短期记忆神经网络 (Long Short Term Memory, LSTM) 于 1997 年提出，它相对于 RNN 神经网络最大的改变是引入了门的概念，分别构造了三个门控单位：即输入门、遗忘门以及输出门，它们作为 LSTM 的记忆单元并对记忆单元进行控制。经过这三个门控的信息筛选，LSTM 神经网络就能够及时对信息进行保留或是清除。新信息从输入门输入并对记忆单元的状态进行更新，遗忘门负责选择对记忆单元的状态信息进行筛选，将无用的信息遗忘，将有用的信息保留，输出门负责输出此时得到的输出数据。遗忘门是决定对隐藏状态信息进行保留或遗

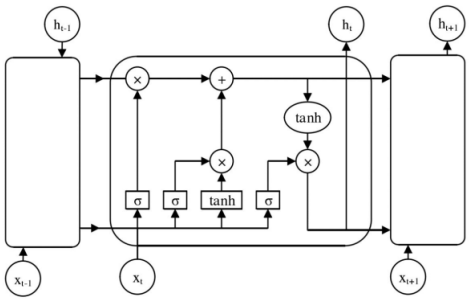

图 2.2 LSTM 结构图

忘的门控单元。遗忘门会将上一个时刻的隐藏状态与这一时刻的输入同时传给 sigmoid 函数，sigmoid 函数会产生 0~1 之间的输出，若是这个输出接近 0，那么就代表应该舍弃，若是输出接近 1，那么就代表应该保留。sigmoid 函数的输出特性使得它很适合用于 LSTM 神经网络并决定信息的保留与删除。它的计算公式如下所示：

$$f(t) = \sigma(W_f h_{t-1} + U_f x_t + b_f)$$

其中 W_f、U_f 以及 b_f 分别为遗忘门的权重、输入数据的权重以及偏置。输入门会同时接收隐藏状态信息以及此时刻输入的信息，它会同时进行两个计算，一个是用 sigmoid 函数对输入的向量信息进行更新，作用与遗忘门相似；临一个是用 tanh 函数输出新的候选向量，并再用 sigmoid 函数保留在 [0,1] 之间的候选向量。它们的计算公式如下所示：

$$i_t = \sigma(W_i h_{t-1} + U_i x_t + b_i)$$

$$\widetilde{c_t} = tanh(W_c h_{t-1} + U_c x_t + b_c)$$

其中 W_i、U_i、b_i、W_c、U_c、b_c 和遗忘门一样是权重系数以及偏置。

在进行遗忘门以及输入门的计算之后，便要根据这两个门的输出结果计算更新的单元状态。更新的公式为：

$$c_t = c_{t-1} \odot + f_t + i_t \odot \widetilde{c_t}$$

其中 $\odot$ 是 Hadamard 向量元素乘积，经过运算之后，便根据前一时刻的单元状态以及这一时刻的候选向量等等数据进行更新，得到了这一时刻的单元状态信息。

输出门会先用 sigmoid 函数处理当前的输入信息，接着与现在的单元状态相结合共同得到新的隐藏状态。计算的公式为：

$$O_t = \sigma(W_o h_{t-1} + U_o x_t + b_o)$$

$$h_t = O_t \odot tanh(C_t)$$

最后输出门会计算并输出此时刻的输出值：

$$\widehat{y_t} = \sigma(V h_t + c)$$

经过上述三个门的计算输出，能得到此时刻的单元状态以及隐藏状态并传递到下一时刻进行计算。

2.1.3 金融情绪分析

本次实验的金融情绪分析使用百度开源的中文文本情感分析模型 Senta 进行实现。文本情感分析旨在识别和提取文本中的倾向、立场、评价、观点等主观信息。具体的说，情感分析主要包括两类任务：情感倾向分类（简称情感分类）和观点抽取。情感分析是人工智能的重要研究方向，具有很高的学术价值。同时，情感分析在消费决策、舆情分析、个性化推荐等领域均有重要的应用，具有很高的商业价值。其中，情感分类是用于识别主观文本中的情感倾向的技术。本文的情感倾向包括积极、中性、消极三类。情感分类是情感分析的核心问题，一直以来是研究的重点，同时它在用户消费习惯分析、危机舆情监控等均有重要应用。

为达到较好的效果，本次实验采用基于 bi-LSTM 的 Senta 模型。总体而言，该模型包括三层，即单词语义层、句子语义层和输出层。单词语义层可将输入文本中的每一个词语转化成连续的语义向量嵌入；句子语义层可通过 bi-LSTM，将词语语义的序列转化成整个句子的语义表示；输出层可基于

句子语义计算情感倾向的概率。

使用该模型对获得的金融股评进行情感分析，再将每日的股评进行整合求均值，可获得一个可以代表当日股民对创业板指走向的态度，如图 2.3 所示。

time	pos_p	neg_p
2021-10-15 00:00:00	0.0954	0.9046
2021-10-16 00:00:00	0.1888	0.8112
2021-10-17 00:00:00	0.7233	0.2767
2021-10-18 00:00:00	0.1225	0.8775
2021-10-19 00:00:00	0.14345	0.85655
2021-10-20 00:00:00	0.1887	0.8113
2021-10-21 00:00:00	0.10115	0.89885

图 2.3　金融情绪分析得分

对该情绪数据进行分析，可以发现存在两个问题，一是评论普遍消极，每一天的平均情绪分数普遍低于 0.5，且区分度不高；二是评论中存在对走向看好，但被判断为消极的评论。

2.2　股票价格预测

2.2.1　CBAM

CBAM（Convolutional Block Attention Module）是一种用于前馈卷积神经网络的简单而有效的注意力模块，它融合了通道注意力 (channel Attention) 和空间注意力 (Spatial Attention)，同时该注意力模块非常轻量化，而且能够即插即用，可以用在现存的任何一个卷积神经网络中。其组成如图 2.4 所示，该模块包含两个完全独立的顺序的子模块：通道注意力 CA 和和空间注意力模块 SA，特征将顺序经过两个模块进行空间和时间的注意力加权。

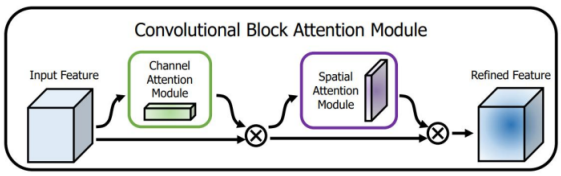

图 2.4　CBAM 结构

首先，CBAM 的输入是一个网络中间特征图 F，将特征图输入至通道注意力模块获取通道注意力 M_c，将注意力权重作用于中间特征图。然后，将施加了通道注意力的特征图输入至空间注意力模块获取空间注意力 M_s，再次将注意力权重作用到特征图上；最终，经过这两个注意力模块的串行操作，最初的特征图就经过了通道和空间两个注意力机制的处理，自适应细化特征。完整的注意力过程由如下公式所示：

$$F' = M_c(F) \bigotimes F$$

$$F'' = M_s(F') \bigotimes F'$$

(1) 通道注意力模块：

通道注意力模块主要利用特征图之间的通道间关系生成通道注意力图。它将特征图在空间维度进行全局平均池化和全局最大池化，将空间维度压缩为 1 而完全保留通道信息；然后将两个池化后的特征送入共享的多层感知机进行特征提取；最后将两特征相加，经过 sigmoid 激活得到最终的通道注意力权重矩阵 M_c，其计算方式如公下公式所示：

$$M_c(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F)))$$

(2) 空间注意力模块：

空间注意力模块主要利用特征间的空间关系生成空间注意图。其主要过程包括将经过通道注意力计算后的特征图在通道维度分别进行最大池化和平均池化，将通道维度压缩为 1 而保留空间信息，然后将池化特征在通道维度拼接，经过卷积层提取特征，同时将通道维度降至 1 之后经过 sigmoid 函数激活，得到空间注意力权重矩阵 M_s，其计算方式如下公式所示。其中 $f^{7\times7}$ 表示一个卷积核尺寸为 7× 的卷积操作。

$$M_s(F) = \sigma(f^{7\times7}([AvgPool(F); MaxPool(F)]))$$

2.2.2　模型介绍

(1) SE 模型：

SE 模型全称 Squeeze-and-Excitation Networks，是一种对 CBAM 通道注意力机制的改进。具体来说，给定一个输入 x，其特征通道数为 C’，通过一系列卷积等一般变换后得到一个特征通道数为 C 的特征。接下来首先进行 Squeeze 操作，顺着空间维度将每个二维的特征通道变成一个实数，这个实数某种程度上具有全局的感受野，并且输出的维度和输入的特征通道数相匹配。它表征着在特征通道上响应的全局分布，而且使得靠近输入的层也可以获得全局的感受野。其次是 Excitation 操作，它是一个类似于循环神经网络中门的机制。通过参数来为每个特征通道生成权重，其中这个参数被学习用来显式地建模特征通道间的相关性。最后是一个 Scale 的操作，将 Excitation 的输出的权重看做是进过特征选择后的每个特征通道的重要性，然后通过乘法逐通道加权到先前的特征上，完成在通道维度上的对原始特征的重标定。

(2) ECA 模型：

ECA 模型全称 Efficient Channel Attention-Net 。同样是一种对 CBAM 通道注意力机制的改进。ECA

是一种不降维的局部跨信道交互策略，该策略通过一维卷积实现,克服了性能和复杂性之间的矛盾,只通过增加少量的参数，就能获得明显的性能增益。ECA 剖析了 SE 模块，证明避免降维和适当的跨通道交互对于学习高性能和高效率的通道注意力是重要的。

(3) HW 模型：

HW 模型指代的是 CBAM 模型中的空间注意力机制。具体见 2.2.1。

(4) 整体模型：

本文采用的整体模型以 LSTM 为基础，分别加入 CNN 和注意力机制，构建多种模型进行对比。

图 2.5 整体模型结构

包括 CNN+LSTM 模型、CNN+LSTM+SE 模型、CNN+LSTM+ECA 模型、CNN+LSTM+HW 模型、以及 CNN+LSTM+CBAM 模型。经过后续实验最终得出 CNN+LSTM+CBAM 模型表现最好，模型整体结构如上图所示。实验中分别用 SE，ECA 与 HW 模型替换 CBAM。其中 CNN 为预处理的 STEM 网络。

3 实验方法

3.1 数据集与评价指标

本文的数据是创业板块自 2021 年 10 月 18 日至 2022 年 10 月 19 日共 245 个交易日的数据。其中开盘价、最高价、最低价、收盘价和成交量为股票的原始数据，情绪指数平均值、情绪指数中位数为金融情绪分析部分得到的数据。选择开盘价、最高价、最低价、收盘价、成交量以及情绪指数六个特征作为神经网络的数据输入，使用 5 天的交易数据预测下一天的收盘指数。

(1) 数据预处理

在找到数据之后首先对数据进行标准化处理。因为指数数据的差值很大，会引发数值的问题。并且不同指标如收盘价以及成交量等等之间的数量级相差较大，如果不进行标准化处理，那么数量级较高的数据将被神经网络更加注意，而数量级较小的数据会被淡化其影响。为了让所有数据都能公平的加入神经网络的训练，在建立模型并分析之前有必要对数据进行标准化处理。

有许多标准化的方法，常用的有标准归一化以及最大最小归一化等等。本文采取的是最大最小值归一化方法,每个值 x 在归一化后会缩放为一个 0~1 之间的值 x^*，$x^* = \frac{x-x_{min}}{x_{max}-x_{min}}$ 。其中 x_{min} 是此指标的历史最小值，x_{max} 是此指标的历史最大值。经过此变换之后，每个指标的数据都会被缩放至统一的区间大小上，这样便有利于不同指标之间的公平竞争，便于模型进行选择重要特征，也有利于加快模型训练时梯度下降的速度，便于模型快速收敛。

(2) 评价指标

本文选择计算 MSE、RMSE、MAE 以及 R2 分数四个指标作为模型的评估标准，通过比较不同模型中四个指标的大小比较模型之间的优劣区别。其中 MAE 以及 R2 的值仅供额外的参考，本文主要通过比较 RMSE 以及 MSE 的值来对模型进行评价。下面将对四个评价指标进行介绍：

1）均方误差（MSE）

均方误差（MSE）描述了预测值与真实值之间偏离的程度大小。它的公式如下：

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

2）均方根误差（RMSE）

均方根误差（RMSE）主要是在 MSE 的基础上进行了开方运算得到的值，其公式如下：

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

3）平均绝对误差（MAE）

平均绝对误差（MAE）衡量了预测值与真实值的相差平均值，其公式如下：

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

4）决定系数（R2）

决定系数（R2）表现了真实值的离散程度，可以比较真实曲线与拟合曲线的接近程度。其公式如下：

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

3.2 实验结果

本文在数据集上分别使用卷积层加上长短期记忆网络以及以卷积层加长短期记忆网络为基础分别添加空间注意力和通道注意力机制的网络进行了对比，使用均方误差、均方根误差、平均绝对值误差以及 R2 分数作为指标进行对比测试。其中，均方误差、均方根误差、平均绝对误差都是误差值越小模型效果越好，R2 分数取值为 0-1，且数值约接近于 1 说明拟合曲线与原始曲线更接近。下表为各模

模型	MSE	RMSE	MAE	R2
CNN+LSTM	5649.04	75.16	60.42	0.35
CNN+LSTM+SE	4720.86	68.71	53.69	0.46
CNN+LSTM+ECA	7019.28	83.78	72.05	0.19
CNN+LSTM+HW	4699.23	68.55	57.66	0.46
CNN+LSTM+CBAM	2958.55	54.39	46.67	0.66
CNN+LSTM+CBAM	6541.8	80.9	71.91	0.25

型使用各个评价指标进行评价的实验结果展示。通过实验证实应用添加 CBAM 模块即同时添加了空间和通道的注意力的深度神经网络，同时添加金融情绪特征的模型在所有模型中预测效果最好。

具体来看，6 组实验分别探究了注意力机制与金融情绪先验对模型性能的影响。实验 1-5 表明对于 CBAM 注意力机制，仅使用空间注意力的 HW 模型与仅使用通道注意力及其改进 SE 与 ECA 模型效果均不如完整使用空间和通道注意力机制的 CBAM 模型，说明在股价时序数据中信息分布并不均匀，部分区域包含较强的特征信息。实验 5，6 表明对于特征维度而言，添加基于金融情绪分析的先验特征同样能提升模型的预测性能，说明金融情绪先验中包含了一定的股价时序特征。

图 3.1，3.2 分别为实验 5，6 预测曲线与真实曲线的对比，其中蓝色为真实曲线，红色为预测曲线，明显可以看到添加了金融先验的实验 6 预测曲线更接近于真实曲线。综上，合理的注意力机制与金融先验信息均对模型性能提升具有积极影响。

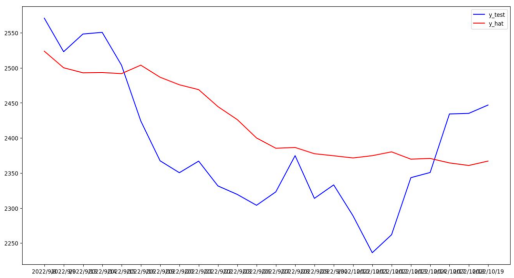

图 3.1　不加金融先验的实验结果

图 3.1　加金融先验的实验结果