Computer Science and Application 计算机科学与应用, 2019, 9(10), 1882-1891
Published Online October 2019 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2019.910211

Hans汉斯

Multi-Category Prediction of Stock Price Based on LSTM

Fengxian Chen[1], Jianqun Zhao[2]

[1]College of Electronics and Information Engineering, Tongji University, Shanghai
[2]The College of Economics and Business Administration, Guangdong Polytechnic of Science and Trade, Guangzhou Guangdong
Email: 978753958@qq.com

Received: Sep. 24th, 2019; accepted: Oct. 9th, 2019; published: Oct. 16th, 2019

Abstract

In order to predict the rise and fall of stock prices, a method based on long-term and short-term memory network (LSTM) is proposed. According to the stock price increase and decrease, through the quantitative classification of the ups and downs, it is transformed into a multi-classification problem. The basic transaction information of the stock is used as the feature input, and it is trained by the neural network. Finally, the stock's ups and downs are classified and predicted. The data set is divided into three parts: the whole category set of the Shanghai and Shenzhen 300 constituent stocks, the bank set and the securities. The experimental results show that the model has a better forecasting effect in the case of multi-classification of the ups and downs; at the same time, in a certain kind of stocks when making predictions, the model trained with the historical trading information of such stocks is better than the model trained with the overall stock trading information.

Keywords

Stock Forecasting, LSTM, Shanghai and Shenzhen 300, Multiple Classification

基于LSTM的股票价格的多分类预测

陈奉贤[1]，赵建群[2]

[1]同济大学，电子与信息工程学院，上海
[2]广东科贸职业学院，经济管理学院，广东 广州
Email: 978753958@qq.com

收稿日期：2019年9月24日；录用日期：2019年10月9日；发布日期：2019年10月16日

文章引用: 陈奉贤, 赵建群. 基于 LSTM 的股票价格的多分类预测[J]. 计算机科学与应用, 2019, 9(10): 1882-1891.
DOI: 10.12677/csa.2019.910211

摘　要

为对股票价格的涨跌幅度进行预测，提出一种基于长短期记忆网络(LSTM)的方法。根据股票涨跌幅问题，通过对涨跌幅度做多值量化分类，将其转化成一个多分类问题。将股票的基本交易信息作为特征输入，利用神经网络对其训练，最后对股票的涨跌幅度做分类预测。数据集分沪深300成分股整体类、银行类和证券类三种股票集，实验结果表明该模型在涨跌幅多分类情况下，有比较好的预测效果；同时，在对某一类股票进行预测时，用该类股票的历史交易信息训练的模型要比以整体股票交易信息训练的模型效果好。

关键词

股票预测，LSTM，沪深300，多分类

1. 引言

随着我国经济的快速发展，政府、投资机构以及投资者们对股票预测的需求也越来越多[1]。因此，对股票价格走势的分析成为越来越多研究者关注的课题。但股票价格高度的波动性与不确定性，使其成为计算机领域和金融领域的一大难题[2]。

股票投资通常会选择某一类或某一只股票来作为投资对象，对这一类或一只股票进行预测，既可以将整体的股票交易信息作为训练数据，也可以只选择该类或该只股票的交易信息作为训练数据。模型有决策树[3]、LR [4]、支持向量机[5]等传统机器学习的方法，也有深度学习的方法[6]。欧阳金亮和陆黎明等人[7]通过引入动态调整学习率的方法和附加量，对传统 BP 算法进行改进，然后用改进的算法对单只股票青岛海尔进行验证。毛景慧[8]通过研究基于 LSTM 的股票价格序列影响因素，使用沪深 300 的 290 只股票数据进行整体的建模预测。

本文根据以往研究中不同训练集的训练，同时考虑股票数据的时序性，选择用对时序序列有较好性能的 LSTM 网络分别对其训练，将训练好的模型用于预测 6 分类和 3 分类情况下的次日收盘价的涨跌幅，并对结果做对比分析。

2. 模型与实现

2.1. 模型理论基础

人工神经网络是 1980 年代兴起的人工智能领域的一类研究热点，它通过对人脑神经元网络进行抽象，构建人工神经单元，再通过不同的连接方式搭建不同的网络结构[9]。

神经网络分为多种，BP 神经网络是一种按照误差逆向传播算法训练的多层前馈神经网络，也是目前应用最广泛的神经网络[10]；卷积神经网络(CNN)则是通过构造卷积层来提取输入特征，再利用前馈连接来完成特征的输出，它是深度学习的代表算法之一[11]；循环神经网络(RNN)适用于输入是序列的数据，它是一种在序列的演进方向进行递归，循环单元按照链式连接的一种神经网络[12]。由于其具有参数共享

的特点，因此适合于序列的非线性特征学习。

长短期记忆网络(LSTM)则是对 RNN 的一种改进[13]，它通过引入门机制构建特殊的记忆神经单元，从而解决 RNN 不能实现信息的长期依赖问题。LSTM 结构如图 1 所示，其包括输入门 i_t、输出门 o_t、遗忘门 f_t 等门结构，这些门结构通过以下的递归方程来更新细胞状态 C_t，同时激活从输入到输出的映射。

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

$$i_t = \sigma\left(W_i \cdot \left[h_{t-1}, X_t\right] + b_i\right) \tag{2}$$

$$o_t = \sigma\left(W_o \cdot \left[h_{t-1}, X_t\right] + b_o\right) \tag{3}$$

其中，f_t、i_t、o_t 分别表示 t 时刻的遗忘门、输入门和输出门。σ 表示激活函数，W 和 b 分别表示权重矩阵和偏置，下标 f, i, o 代表遗忘门、输入门和输出门。h_{t-1} 表示 $t-1$ 时刻 LSTM 细胞的输出，x_t 代表 t 时刻的输入。

遗忘门 f_t 决定神经单元遗弃哪些信息，该门层通过读取 h_{t-1} 和 x_t 的状态，输出一个 0 到 1 之间的值，0 代表完全舍弃，1 代表完全保留。输入门 i_t 决定神经单元要更新的值，它从遗忘门筛选完的信息中，通过 tanh 函数更新神经单元状态。最终神经单元输出的状态由输出门 o_t 决定，其先用 sigmoid 层决定要输出的神经单元状态，然后将这些状态用 tanh 函数压缩在−1 到 1 之间。

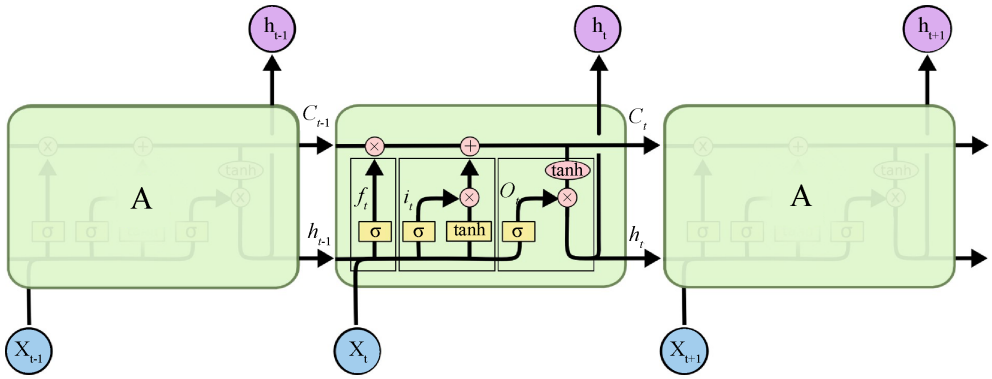

Figure 1. LSTM unit structure [14]
图 1. LSTM 单元结构[14]

2.2. 股票预测问题描述

本文主要针对预测股票涨跌幅度的目标，将其转换为一个多分类任务来进行处理。

影响股票涨跌的因素有很多，与股票本身信息相关的有其基本交易数据如开盘价、收盘价、最高价、最低价、交易量、涨跌幅等 6 类，还有交易数据衍生出的一些统计技术指标，如换手率等[10]。

除了交易数据，股市波动还通常和舆论、政策等因素相关。但这些特征信息不能直观、即时的反映到后续的股票价格中去，同时这些信息是否与股票的基本信息耦合也尚未论证。因此本文只对股票的 6 类基本交易数据作为输入特征。记某连续 T 个交易日的价格序列为：

$$P_1, P_2, \cdots, P_T$$

本文根据股票价格的前 T 天的历史交易信息，预测下一天股票收盘价较第 T 天的涨跌幅区间 Y_{T+1}。6 分类下的第 $T+1$ 天的收盘价涨跌幅区间定义如下：

$$Y_{T+1}=y\left(P_1,P_2,\cdots,P_T\right)=\begin{cases}5, & Y_{T+1}>2\%\\ 4, & 1\%<Y_{T+1}<2\%\\ 3, & 0<Y_{T+1}<1\%\\ 2, & -1\%<Y_{T+1}<0\\ 1, & -2\%<Y_{T+1}<-1\%\\ 0, & Y_{T+1}<-2\%\end{cases} \tag{4}$$

0~5 分别代表了 6 个涨跌幅区间，模型最终是预测第$T+1$天收盘价的涨跌幅属于哪个区间。

在股票预测时，预测股票是否出现大涨或大跌往往比只预测涨跌更具有实际意义，因此，本文还做了涨跌幅 3 分类的情况，既将第二天股票收盘价涨幅超过 1%的作为大涨，跌幅超过−1%的作为大跌，在这之间的认为是不涨不跌。3 分类下第$T+1$天的收盘价涨跌幅区间定义如式(4)所示：

$$Y_T=\begin{cases}0, & Y_T>1\%\\ 1, & -1\%<Y_T<1\%\\ 2, & Y_T<-1\%\end{cases} \tag{5}$$

2.3. 预测方法

本文利用能长久保持时序信息的 LSTM 作为网络框架，网络的输入包括 6 个基本交易数据特征：开盘价、最高价、最低价、收盘价、涨跌幅和交易量。

考虑到单层的网络对特征学习效果不佳，构建一个三层的 LSTM 网络，每层神经元节点数为 128，层与层之间全连接。神经网络的输出层接一个 Softmax 层，将输出转换为每个涨跌幅区间的概率，并以取得最大概率的区间为预测区间。在 6 分类下，最后的输出是 6 维，既 6 种涨跌幅区间的概率值。在 3 分类下，最后的输出是 3 维，既 3 种涨跌幅区间的概率值。在两种不同的分类下，都将概率取得最大的涨跌幅区间作为模型最终的预测区间。

LSTM 预测模型如图 2 所示。模型在 Softmax 后是两种不同区间分类的网络输出。

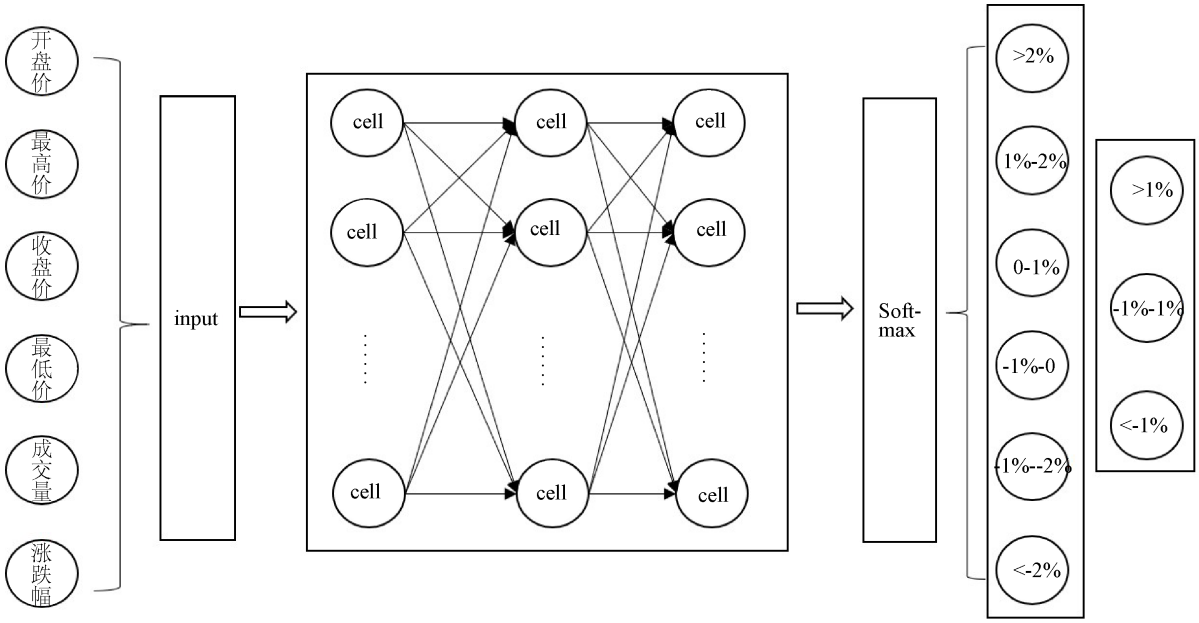

Figure 2. LSTM prediction model
图 2. LSTM 预测模型

本文使用有监督的方式对 LSTM 网络进行训练，既以 T 天的数据集作为输入，将式(3)和式(4)所示下一天较第 T 天的涨跌幅区间用 Onehot 编码处理后，作为标签输入到网络中。

为了优化网络权重，通常需要使用损失函数来估计预测值和标签的差异。本文使用交叉熵损失函数，定义如下：

$$L=-\frac{1}{n}\sum_{j=1}^{n}\sum_{k=0}^{c}p_j(k)\log\hat{p}_j(k) \tag{6}$$

其中，n 表示训练样本数，$\hat{p}_j(k)$ 表示模型的预测样本 j 属于类别 k 的概率，$p_j(k)$ 表示样本 j 的实际概率，如果样本 j 属于 k，则取值为 1，否则为 0。

预测值和标签值越接近，则损失函数的值就越小。因此需要通过反向传播来不断更新模型的权重值，使损失函数不断减小，让模型不断地学习数据特征。

在实际参数优化上，本文选用收敛速度较快、适应性较好的 Adam 算法[15]对损失函数进行优化。

3. 实验与结果

3.1. 数据集描述

本文所采用的数据集分为两个部分，一部分是从沪深 300 股票集中选取在 2013 年到 2017 年之间有连续交易数据的 160 只股票作为整体股票的数据集 1，选取上市的在 16 只银行股票作为类别股的数集 2，上市的 14 只证券类股票作为类别股的数集 3。

数据集的基本信息如表 1 所示。

数据集包含了从 2013 年到 2017 年的股票基本交易信息，由于有的股票会在正常交易日出现停盘的情况，因此需要对停盘日期的数据填充，本文使用的方式是用停盘交易日前一天的数据来填充该日的交易信息。在处理后的股票数据集中，每个交易日都包含了开盘价、收盘价、涨跌幅等信息。表 2 是部分数据集信息。

Table 1. Data set overview
表 1. 数据集概况

数据集	交易天数(天)	股票数量(只)
数据集 1	1215	160
数据集 2	1215	16
数据集 3	1215	14

由表 2 可知，成交量和其他几个特征在数值上有很大的差异，因此在数据集送入到 LSTM 网络之前，还需要对其做归一化处理，以降低不同特征之间的数值差异，减免奇异样本在优化时对梯度方向的影响。

Table 2. Partial data set information
表 2. 部分数据集信息

日期	开盘价	最高价	最低价	收盘价	涨跌幅	成交量
2013/1/4	16.32	16.45	15.92	15.99	−0.19	443851
2013/1/7	15.98	16.35	15.88	16.3	1.94	357169
2013/1/8	16.3	16.37	15.86	16	−1.84	312479
2013/1/9	15.96	16.02	15.8	15.86	−0.88	251329

为了进一步了解数据特征，本文对三个股票数据集根据其涨跌幅做了样本统计。

图 3、图 4 和图 5 分别是沪深 300 整体类、证券类和银行类在 2013~2015 年期间的股票集价格涨跌幅分布图。由图可知，三种类型的股票价格涨跌幅分布都近似于高斯分布，其中整体类和证券类的股票都存在部分涨跌停的股票，而银行类则出现的较少，这说明整体类和证券类的股票出现。

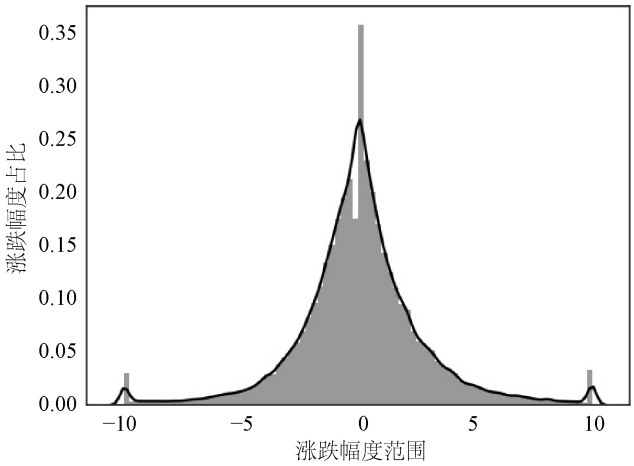

Figure 3. Shanghai and Shenzhen 300 overall stocks up and down distribution map
图 3. 沪深 300 整体类股票涨跌幅分布图

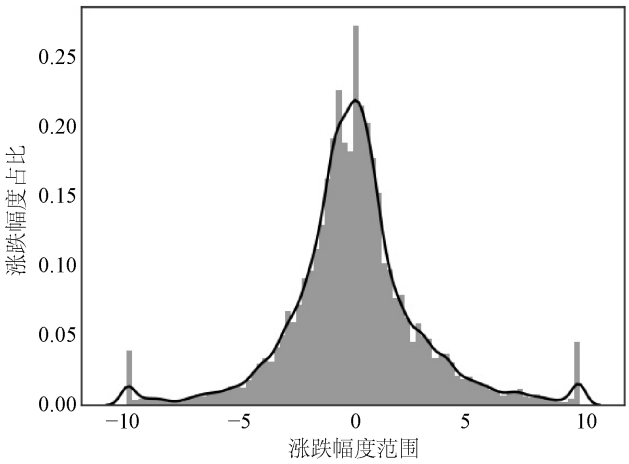

Figure 4. Securities stocks up and down distribution map
图 4. 证券类股票涨跌幅分布图

动荡的样本数较多，其价格涨跌幅变化更为复杂，预测难度也更大。除此之外，三类股票集的涨跌幅众数都向零点右侧偏移，这说明在这五年期间，沪深 300 股票市场整体体量有所增长。

在涨跌幅 3 分类下的数据占比直方图如图 6 所示，由图可知 3 种分类情况下数据分布基本均衡，不涨不跌类别的数据所占比例稍多于其他两类。

3.2. 实验结果

为了检验不同数据集训练的模型对股票的预测效果，本文用数据集 1、2 和 3 分别训练对应模型，为了评测训练效果，将数据集划分为训练集、验证集和测试集。对于每只股票，以整体数据的前 80% (2013

年~2016 年)作为训练集，剩余 20%作为测试集。训练集用于训练调节 LSTM 网络参数，测试集用于评测效果。在训练集里，又以 8 比 2 的比例分割训练集和验证集。验证集数据用于判断模型收敛时的迭代次数，避免过拟合。本文使用分类准确率作为判断模型优劣的标准。

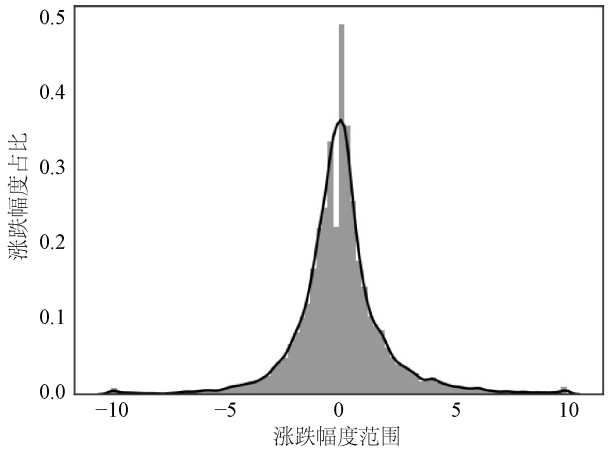

Figure 5. Bank stocks up and down distribution map
图 5. 银行类股票涨跌幅分布图

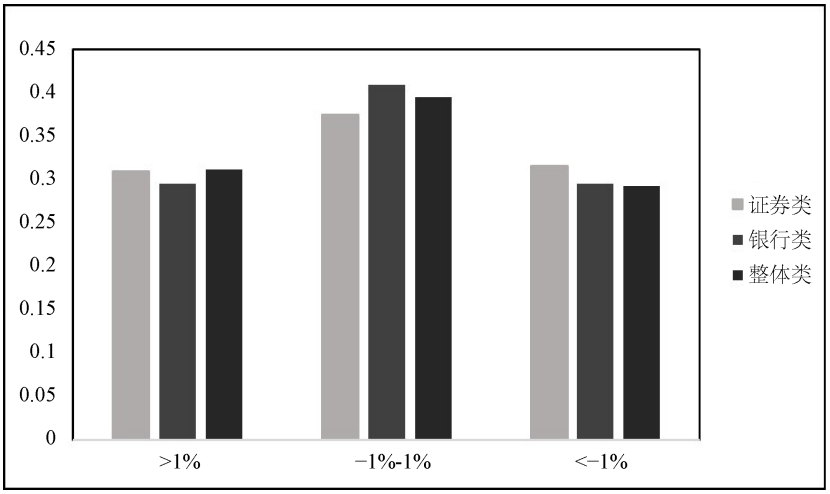

Figure 6. 3 kinds of classification data distribution map
图 6. 3 种分类数据分布图

本文开始设定的隐藏层节点数为 128，学习率是 0.01，训练迭代次数为 150 次。使用沪深 300 的 160 只股票训练的模型记为 M1，用银行类的 16 只股票训练的模型记为 M2，用证券类的 14 只股票训练的模型记为 M3。在训练过程中，通过验证集的损失函数值变换情况来判断模型是否收敛。图 7、图 8 和图 9 分别表示的是 M1、M2 和 M3 模型迭代 150 次时训练集和验证集损失函数值随迭代次数的变化情况。

由图 7 和图 9 可知，M1 和 M3 模型在迭代了 100 次左右时就已收敛，由图 8 可知，M2 模型在迭代了 80 次左右收敛，在此之后的迭代训练过程中，三种模型虽然训练集的损失函数值仍在下降，但验证集的损失函数值开始振荡，说明模型出现了过拟合。

因此用测试集对模型测试时，选择各自损失函数最小时的检查点模型。三种模型在不同测试集数据上的结果如表 3 所示。

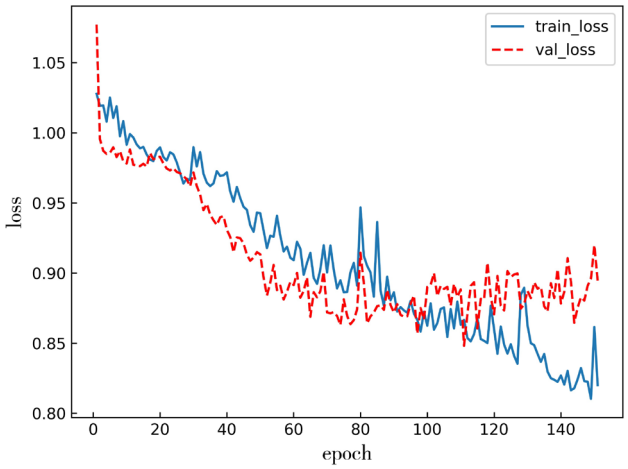

Figure 7. M1 model loss function
图 7. M1 模型损失函数

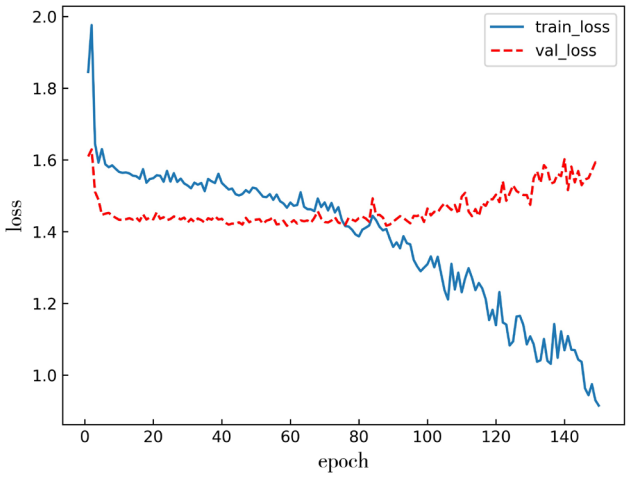

Figure 8. M2 model loss function
图 8. M2 模型损失函数

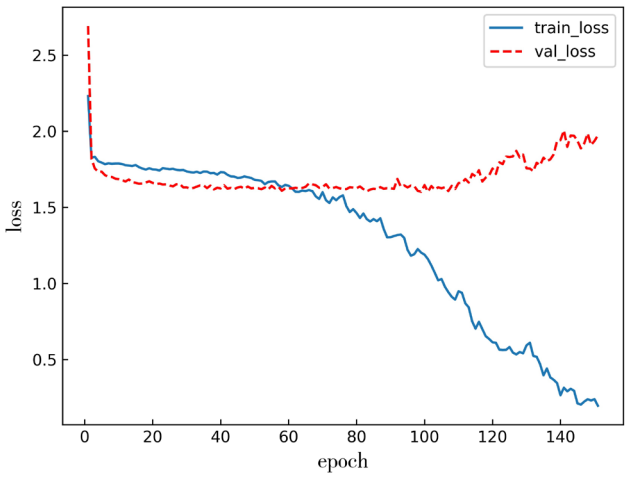

Figure 9. M2 model loss function
图 9. M3 模型损失函数

Table 3. 6 classification accuracy in different model
表 3. 不同模型 6 分类准确率

模型	测试集	准确率
M1	沪深 300	0.276
M1	银行类	0.311
M1	证券类	0.277
M2	银行类	0.334
M3	证券类	0.286

由表 3 可知，三种数据集的准确率都好于 6 分类随机预测的结果，这说明模型学习到了一定量的股票价格波动规律。其中，银行类的预测准确率高于其他两类，这是由于银行类的大盘股占比高，相对其他类型的股票，受非股票本身因素波动的影响较小。在同一数据集不同模型的表现效果上，M2 模型在银行类测试集上的表现优于 M1 模型，M3 模型在证券类测试集上的表现也优于 M1 模型。这说明对特定类别的股票预测时，选择该领域的股票数据进行训练，预测效果要好于用整体股票信息训练。

为了进一步评估模型的性能及更好的为投资者提供有用信息，本文还用了大涨、大跌和不涨不跌 3 分类的模型进行评估，表 4 是不同模型在三个数据集上 3 分类的准确率。

Table 4. 3 classification accuracy in different model
表 4. 不同模型 3 分类准确率

模型	测试集	准确率
M1	沪深 300	0.517
M1	银行类	0.593
M1	证券类	0.548
M2	银行类	0.626
M3	证券类	0.586

由表 4 可以看出，3 分类的准确率要比 6 分类的准确率高，这是因为数据分布不均衡，股票价格的涨跌主要集中在了不涨不跌这一类上，所以在把涨跌幅较小的数据合为一类时，模型的准确率有大幅提升。不同数据集上，银行类数据的准确率高于其他两类。除此之外，在 3 分类模型上，同样存在某一领域的股票预测上，选用该领域的训练集训练出的模型效果要好于用整体股票信息训练出的模型。

4. 总结

在预测股票价格涨跌幅上，本文提出了一种基于 LSTM 的多分类预测模型。用沪深 300、银行类和证券类等多种数据集分别训练不同的模型，对比不同数据集上各模型在 6 分类和 3 分类情况下的预测效果。实验中发现，模型在 6 分类和 3 分类的股票涨跌幅预测上，均有比较好的效果。对特定类别的股票预测时，选择该领域的股票数据进行训练，预测效果要好于用整体股票信息的训练，最高时准确率可提升 2%。在对涨跌情况三分类的预测中，最高准确率可达 0.62。在今后的工作中，还可以根据预测任务的不同尝试使用不同的特征用于训练，探索股票的不同交易数据对于股票价格预测的影响。

参考文献

[1] 刘长虎, 陶建格, 崔衍秋. 股票价格指数的投资功能[J]. 市场论坛, 2004(6): 71-72.

[2] 黄丽明, 陈维政, 闫宏飞. 基于循环神经网络和深度学习的股票预测方法[J]. 广西师范大学学报(自然科学版), 2019, 137(1): 13-22.

[3] 沈金榕. 基于决策树的逐步回归算法及在股票预测上的应用[D]: [硕士学位论文]. 广州: 广东工业大学, 2017.

[4] Huo, J., Zheng, Y. and Chen, X. (2009) Implementation of Transaction Trend Prediction Model Based on Regression Analysis. *Journal of Baosha Teachers' College*, **117**, 19-23.

[5] 黄秋萍, 周霞, 甘宇健, 韦宇. SVM 与神经网络模型在股票预测中的应用研究[J]. 微型机与应用, 2015, 34(5): 88-90.

[6] 刘磊. 基于深度学习的股票价格趋势预测方法研究[D]: [硕士学位论文]. 昆明: 云南财经大学, 2017.

[7] 欧阳金亮, 陆黎明. 综合改进 BP 神经网络算法在股价预测中的应用[J]. 计算机与数字工程, 2011, 39(2): 57-59.

[8] 毛景慧. 基于 LSTM 深度神经网络的股市时间序列预测精度的影响因素研究[D]: [硕士学位论文]. 广州: 暨南大学, 2017.

[9] 阎平凡, 张长水. 人工神经网络与模拟进化计算[M]. 北京: 清华大学出版社, 1900.

[10] 闻新, 张兴旺, 朱亚萍, 李新. 智能故障诊断技术: MATLAB 应用[M]. 北京: 北京航空航天大学出版社, 2015.

[11] Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, L., Wang, G. and Cai, J. (2015) Recent Advances in Convolutional Neural Networks.

[12] Goodfellow, I., Bengio, Y. and Courville, A. (2016) Deep Learning (Vol. 1). MIT Press, Cambridge, 326-366.

[13] Graves, A. and Schmidhuber, J. (2005) Framewise Phoneme Classification with Bidirectional LSTM and Other Neural Network Architectures. *Neural Networks*, **18**, 602-610. https://doi.org/10.1016/j.neunet.2005.06.042

[14] KIMY (2014) Convolutional Neural Networks for Sentence Classification. In: *Proceedings of the* 2014 *Conference on Empirical Methods in Natural Language Processing*, ACL, Stroudsburg, 1746-1751. https://doi.org/10.3115/v1/D14-1181

[15] Kingma, D.P. and Bal, J. (2014) Adam: A Method for Stochastic Optimization. *Computer Science*. https://arXiv.org/abs/1412.6980