计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190900128

基于 LSTM-GA 的股票价格涨跌预测模型

包振山[1]　郭俊南[1]　谢　源[2]　张文博[1]
1 北京工业大学信息学部　北京 100124
2 康曼德资本管理有限公司　北京 100600
(baozhenshan@bjut.edu.cn)

摘　要　如何准确地进行股票预测一直是量化金融领域的重要问题。长短期记忆细胞神经网络(LSTM)的出现较好地解决了股票预测这类的复杂序列化数据学习的问题。然而前期研究结果表明单一使用该方法仍存在预测不平衡、陷入局部极值导致能力不佳的问题。基于上述问题,文中利用将遗传算法(GA)解决调参问题来保证模型预测的平衡性,由此构建了新型股票预测模型。该模型分为三部分,首先利用 LSTM 网络进行收盘价的预测,再利用基于遗传算法的判别机制,最终获取下一刻股票的涨跌信号。这一模型不同于先前的研究,主要针对 LSTM 模型的输出模块进行了改进。文中使用了中证 500 的日内分钟数据进行测试验证。实验得出,改进模型的各方面指标均优于单独的 LSTM 模型。

关键词:长短期记忆神经网络;遗传算法;机器学习;股票预测
中图法分类号　TP183

Model for Stock Price Trend Prediction Based on LSTM and GA

BAO Zhen-shan[1],GUO Jun-nan[1],XIE Yuan[2] and ZHANG Wen-bo[1]
1 Faulty of Information Technology,Beijing University of Technology,Beijing 100124,China
2 Commando Capital Company,Beijing 100600,China

Abstract　How to make an accurate financial time series prediction is one of the important quantitative financial problems. Long and short term memory neural network (LSTM) has solved the complex serialized data learning problems such as stock prediction much better. However,the results of previous studies show that there are still some problems such as unbalanced prediction and local minimum value,which lead to poor prediction ability. Based on the above problems,the genetic algorithm (GA) is used to solve the parameter adjustment problem to ensure the balance of model prediction,and a new stock prediction model is constructed. First,LSTM neural network is used to predict closing price. Then,the prediction results are calculated to the judgment method based on genetic algorithm. Finally,the predicted stock's ups and downs signals are gained as the output. This model is different from the previous state-of-the-art and is mainly improved for the output module of the LSTM model. High-frequency trading data of Index China are used for verification. The results show that the improved model is better than the LSTM model.

Keywords　Long short-term memory,Genetic Algorithm,Machine learning,Stock prediction

1　引言

股票市场的预测一直备受关注,然而由于其固有的噪声环境和相对于市场趋势的较大波动性,其掣肘因素特别多,因此该过程非常复杂。以往股票市场预测技术大致可分为基于预测的技术和基于聚类的技术两大类[1]。本文主要讨论基于预测技术的股票价格预测问题。目前针对股票价格预测的研究有人工神经网络(ANN)、支持向量机(SVM)、时间序列模型、基于模糊的技术和随机性来优化定价模型等方法[1]。

在前人工作中,人工神经网络(ANN)已被证实善于处理复杂关系问题,但是神经网络的测试和训练速度较慢。此外,过度拟合、陷入局部极小值和黑盒技术是神经网络的缺点[2]。SVM 的特征选择方法无法指出所需的最优特征数量,严重影响了系统精度,因此该研究不易拓展[3]。Krauss[4]将随机森林方法应用到了股票预测问题上,取得了较好成绩。Fischer 等[5]评价,随机森林可以作为任何创新机器学习模型的强大基准。GARCH 是经典的时间序列模型,广泛应用于时间序列的预测,但它们都假设时间序列的值是一个线性生成过程。然而,市场特征是非线性的,与政治和经济条件及其经营者的期望相互作用,使得 GARCH 假设不适用于许多金融时间序列应用[6]。长短期记忆网络(LSTM)是一种时间递归神经网络,适合处理和预测时间序列中间隔和延迟相对较长的重要事件。这一技术特征与股票预测问题有着很高的契合度,LSTM 弥补了 GARCH 受限于线性模型的问题,极有可能为股票预测问题提供一个解决方案。Fischer 等基于 S&P500 数据通过 LSTM 模型进行收益率的预测。实验表明,收益率预测准确率在 51%～54%左右,要优于随机森林和 DNN 模型[5]。这样的预测结果并不令人满意,导致准确率不理想的

本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家重点研发计划项目(2017YFC0803300)
This work was supported by the National Key R&D Program of China (2017YFC0803300).
通信作者:张文博(zhangwenbo@bjut.edu.cn)

原因有很多，比如模型结构不佳、特征样本不适合、神经网络通病——陷入局部极值。在该问题的讨论中，有研究者调整了模型参数，结合多维度数据处理特征样本等。

延续以往的研究，本文基于中国中证 500 股票数据提出了一种基于 LSTM 网络的机器学习方法来预测未来价格，而后又在 LSTM 模型基础上改进为股票涨跌的预测，该方法采用遗传算法改进调参模型，从而提升了预测效果，弥补了单一应用 LSTM 网络的不足。

2 相关工作

2.1 LSTM 原理与应用

LSTM 最早是由 Hochreiter 等提出的[7]，2000 年 Schmidhuber 等改进了 LSTM 网络，提出遗忘门的方法，适用于连续性的预测[8]。之后 Grave 的书中又对 LSTM 改进和推广[9]。在许多问题上，LSTM 取得了相当大的成功，并得到了广泛的应用。

LSTM 神经网络的前身是循环神经网络。循环神经网络(RNN)是通过内部循环学习序列模式的神经网络。RNN 网络中有很多循环回路，它可以将信息持续传递下去。权重的学习调整采用的是链式法则反向传播。当数值反向传播到活化函数中，如 sigmoid 和 Tanh 函数，斜率会变得极小(或极大)，出现梯度消失(或梯度爆炸)的问题。LSTM 模型是为了避免这些问题而开发的。Hochreiter 等提出了内存单元和门[7]，这样的结构可以长时间存储信息，同时忘记不必要的信息。

LSTM 网络采用记忆单元来代替神经元。图 1 是 LSTM 记忆单元示意图，一个 LSTM 单元由 1 个记忆细胞(C_t)和 3 个门结构组成，包括输入门(i_t)、遗忘门(f_t)、输出门(o_t)。在 t 时刻，x_t 代表输入数据，h_t 代表隐藏层状态。符号×代表向量外积，符号＋代表叠加运算。LSTM 的运算公式如式(1)－式(6)所示。

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

$$i_t=\sigma(U_i x_t+W_i h_{t-1}+b_i) \tag{2}$$

$$u_t=\tanh(U_u x_t+W_u h_{t-1}+b_u) \tag{3}$$

$$c_t=f_t * c_{t-1}+i_t * u_t \tag{4}$$

$$o_t=\sigma(U_o x_t+W_o h_{t-1}+b_o) \tag{5}$$

$$h_t=o_t * \tanh(c_t) \tag{6}$$

其中，U、W 代表矩阵权重，b 代表偏移量，σ 是 sigmoid 函数，符号 * 代表向量外积。

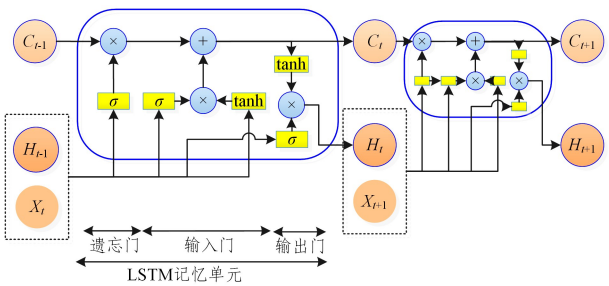

图 1 LSTM 记忆单元

Fig. 1 LSTM memory unit

遗忘门计算 x_t，h_{t-1}，b_f 的加权和，通过 sigmoid 函数得到 f_t($f_t\in(0,1)$)，如式(1)所示。f_t 代表上一个记忆细胞(C_{t-1})中需要遗忘的信息权重。换而言之，遗忘门是用来控制上一记忆细胞中保留的信息量，在式(4)有所显现。输入门，决定了要接收多少新信息到记忆细胞(C_t))(见式(2))。C_t 是记忆细胞的储存权重(见式(3))。原有信息和新增信息分别由遗忘门和输入门控制，得到了当前的记忆细胞(C_t)(见式(4))。最后通过式(5)输出门来过滤记忆细胞(C_t)。更新后的记忆细胞通过式(6)获得当前的隐藏层状态 h_t。最后进行反向传播，得到由这些存储块组成的 LSTM 模型。

LSTM 具有时间序列的概念，所以在处理序列数据、上下文相关的实际问题方面表现格外出色。如今最新的研究也围绕这一优势展开。Qin 等将 LSTM 用到程序 bug 反馈报告分类上，根据文本的内容分辨是否为真实的 bug[10]。Huang 等同样抓住了 LSTM 善于捕捉序列信息的特点，他们提出了基于主题信息的双向 LSTM(BiLSTM)情感分类模型，用来将文字的情绪进行分类[11]。另一方面 LSTM 受到金融预测领域的青睐。Nelson 等利用 LSTM 算法预测股票市场的价格趋势，对其结果进行了分析，结果表明，当预测某只股票的价格在不久的将来是否会上涨时，其准确性平均达到 55.9%[12]。

2.2 LSTM 在股票预测领域的应用

Lin 等从模型结构角度展开研究。他们将 LSTM 应用到金融预测上，配合遗传算法进行初始值调优[13]，相比传统 LSTM 其能更快地在满足预测能力的情况下完成训练。另一方面，人们发现特征样本对一个模型的性能有很大的影响。史建楠等提出一种基于动态模态分解-长短期记忆神经网络(DMD-LSTM)的股票价格时间序列预测方法。他们认为股票市场是整体相关的复杂系统，并利用动态模态分解算法对股票原始数据进行特征提取。实验发现相比传统的预测方法，在特定的市场背景下其能实现更高的价格预测精度[14]。陈佳等采取线性组合对原始数据去噪降维，之后进行应用系统聚类对特征参数再次精简，用这样的特征去训练 LSTM 模型。实验结果表明该方法计算量小，预测结果在速度和准确度两方面均得到显著提升[15]。

2.3 对比组相关应用

Ho[16]提出了随机决策森林算法。随机森林由许多不同训练数据引导样本构建的深度不相关决策树组成。随机森林算法采用随机特征选择不相关决策树和 Bagging 两种关键技术来建立引导样本数据。假设定义有 B 棵决策树的森林从训练数据中抽取一个引导样本，构建决策树。每个决策中，只有 p 个特性的一个子集 m 可用来作为潜在的分割条件。一旦决策树达到最大深度 J，就停止构建，即一棵决策树构建完成。最终得到一个含有 B 棵树的森林。使用森林进行分类，最终结果由每棵树的投票结果决定。

我们将树的数量 B 设置为 1000，并将最大深度 J 设置为 20，这样就可以实现更高阶的交互。随机特征选择的默认值为 $m=\sqrt{p}$。这种配置下的随机森林是 Krauss 等[4]的最佳单一技术。随机森林需要配置的超参数较少，模型性能相对稳定，因此本实验以此算法作为对照组。

3 股票预测模型框架设计

3.1 总体框架设计

股票预测模型分为两部分，由 LSTM 网络和涨跌决策模块组成，如图 2 所示。我们设计了 LSTM 网络，用来学习股票历史数据规律，预测出下一时刻的股票收盘价。涨跌决策模块用来将收盘价数据转换成涨跌信号。我们根据具体实验设计了两套算法，一是传统的阈值方式来定义涨跌；另一种是

结合股票历史数据来自适应地调整阈值大小定义涨跌。

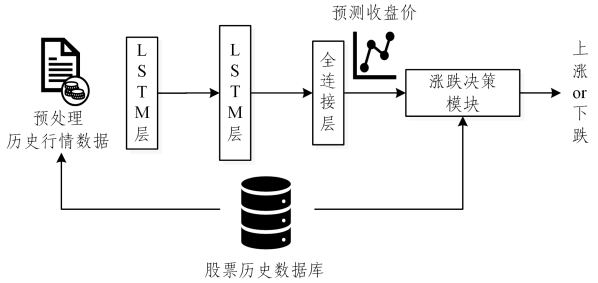

图 2 预测模型总体设计框图

Fig. 2 Overall design of prediction model

3.2 LSTM 网络实现与配置

本文的实验在 python 3.6 上配合一系列依赖库完成的。LSTM 网络是基于 Keras[17] 开发的，在 Google TensorFlow 之上运行。TensorFlow 是一个辅助构建和部署 ml 模型的平台。Keras 是一个高级神经网络 API，用 Python 编写，能够在 TensorFlow，Cntk 或 Theano 平台上运行。它的优势在于，能够以最快的速度将实验设想实现[18]。表 1 展示了本文实验的软件。

表 1 软件环境

Table 1 Software Configuration

编译环境	机器学习环境	依赖库
Python3.6	Keras＋TensorFlow	Pandas＋NumPy Matplotlib

下面具体描述 LSTM 网络的配置。

1）使用 RMSprop 算法作为优化器。选择 RMSprop 是因为在 2016 年 Chollet 提出：RMSprop 通常对于循环神经网络来说是一个好的选择[17]。

2）在每个 LSTM 层后面加入 dropout 层。由于过高的 dropout 参数会导致学习效果的下降，所以我们设置 dropout 比率为 0.1[1]。

3）加入“早停”（early stopping）机制来降低过拟合的几率。我们设置最大训练周期为 1 000，早停的容纳程度为 20。根据 Granger 的研究，有 20％的样本在训练中用于回测[19]。所以我们将训练样本随机分为两部分，80％用于训练学习，20％用于回测确认。

结合上述经验和我们实验中的探索，得出了本文的网络结构，如图 3 所示。

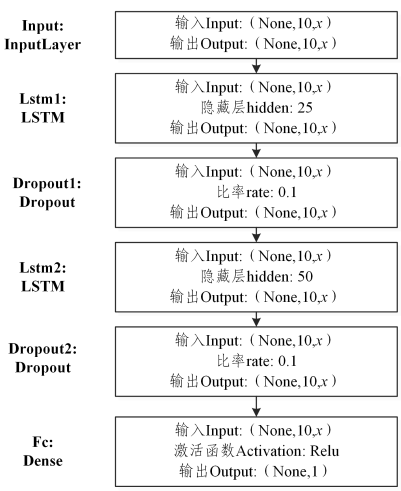

图 3 LSTM 网络结构

Fig. 3 Construction of LSTM Net

1）神经网络第一部分是单特征向量的输入层。每个向量由 10 个时序数据组成。

2）双层 LSTM 层，第一层 LSTM 设有 25 个隐藏层后接一个比例参数为 0.1 的 dropout 层。第二层 LSTM 设有 50 个隐藏层后接一个参数为 0.1 的 dropout 层。这样对每个参数进行大量的观测，可以在遇到干扰较大的训练数据时进行更可靠的估计，并减少过度拟合的风险。

3）神经网络的输出层采用 ReLU 激活函数来输出最终的结果。

4 实验验证

4.1 数据集的构造

本次实验是基于中证 500 指数（000905.SH）从 2018 年 1 月至 2018 年 12 月的每日分钟序列数据开展的。中证 500 指数是中证指数有限公司所开发的指数中的一种，其样本空间内股票是由全部 A 股中剔除沪深 300 指数成份股及总市值排名前 300 名的股票后，总市值排名靠前的 500 只股票组成，综合反映中国 A 股市场中一批中小市值公司的股票价格表现。考虑到尽量使影响因素单一化，我们选择了分钟序列的高频数据。这样的数据主要与股票的交易买卖有关，其他因素影响较少，我们选择了每分钟的收盘价格作为反映股票变化的特征量。我们从 wind 金融量化接口获取了 2018 年中证 500 分钟序列的收盘价数据，原始数据存储格式如表 2 所列。

表 2 2018 年 3 月 1 日部分交易时间、收盘价数据

Table 2 Part of closing price data in 2018.3.1

交易时间	收盘价
2018/3/1 9:30	5975.920
2018/3/1 9:31	5992.372
2018/3/1 9:32	6000.012
…	…
2018/3/1 14:57	6088.414
2018/3/1 14:58	6088.591
2018/3/1 14:59	6089.798

4.1.1 数据预处理

排查原始数据，将空缺数据进行补齐或剔除，方便进行模型训练与测试。这里主要进行了两部分操作：1）由于一些特殊原因某些间断时刻收盘价数据是空缺的，考虑到数据集是分钟序列数据，前后数据变化不会很大，使用上一分钟的数据来弥补不会对模型产生过大的影响，这里默认使用上一分钟的数据进行代替。2）中国股票市场规定，周六日以及重大节日时是全天休市的。故一些日期内的所有分钟数据均是无效的，这里将这些时间节点的分钟数据全部剔除，只保留交易日的数据。考虑到我们将预测股价的数值，符合正态分布的数据集更利于模型的学习和预测输出，而收盘价数据不能很好地满足这一条件，需要在将原始数据输入网络之前对其进行转化。这里我们采用了标准化（Z-SCORE）对数据进行规范化：

$$x_{norm}=\frac{x-\bar{x}}{\sigma_{\bar{x}}} \tag{7}$$

其中，x 是需要归一化的值的向量，$\bar{x}$是数据的平均值，$\sigma_{\bar{x}}$ 是数据的标准差。我们不是简单地将所有的值都归一化，而是考虑每日的价格变化存在差异，为了更好地捕捉每日股票变化的特征知识，对每个输入序列使用单独的正则化器，对每样本序列使用不同的平均值和标准差。

4.1.2 训练集与测试集划分

借鉴 Krauss 的样本处理方式[14]，定义学习周期的概念。学习周期分为学习样本组(2 个月，按天计算约 40 样本序列)和测试样本组(1 个月，按天计算约 20 样本序列)。如图 4 所示，我们将整个数据集划分为 4 个学习周期。每次输入模型学习的样本序列按天划分，即一次输入为日内分序列 240 个数据(9 点 30 分至 11 点 30 分，13 点至 15 点)。每一个样本由 10 个序列数据(已知输入值)和序列的下一个数据(预测输出值)组成。

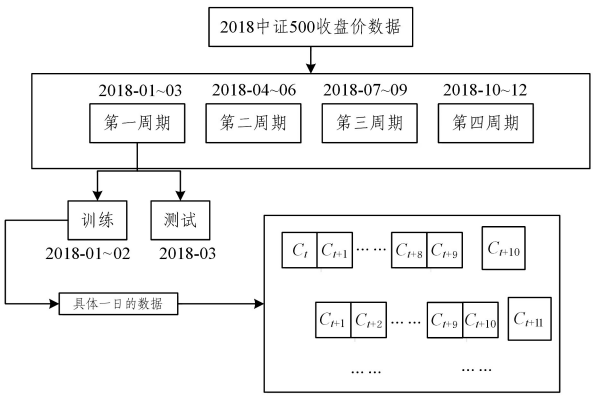

图 4 样本划分结构与格式(C_t 表示 t 时刻的收盘价)

Fig. 4 Sampling structure and format(C_t is closing price at t moment)

4.2 结果分析

将第 3 节设计的模型应用到中证 500 分钟序列收盘价数据集上，得到预测结果。首先计算了预测值与真实值之间的误差，得到了相关数据(见表 3)。综合平均值、方差发现，平均误差不大，误差大部分分布在±5 范围内。我们随机选取了部分测试数据绘制图 5 的对比曲线，可以更直观地看出预测的效果。

表 3 误差分析

Table 3 Errors analyzing

数据	误差分析		
	平均值	方差	范围
第一周期	0.00563	3.41717	[−14.284, 12.867]
第二周期	−0.05000	6.04668	[−16.761, 22.290]
第三周期	−0.03527	5.43115	[−13.025, 18.512]
第四周期	0.09761	4.65242	[−12.187, 15.115]

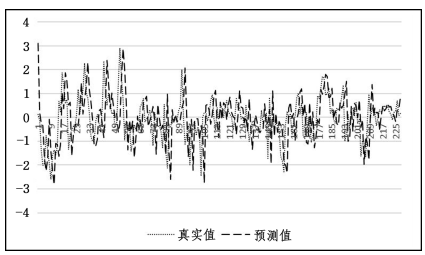

图 5 2018 年 3 月 8 日预测结果对比(标准化数据)

Fig. 5 Prediction comparison 2018.3.8(standardization)

我们将 4 个周期的数据分别进行训练和预测测试，将真实数据和对应的预测结果进行对比。在 Fischer 等[5]的研究中，用收益率数据来训练模型，之后由模型预测股票涨跌情况。他们的实验表明这样的模型正确率在 51%～54%左右，比随机森林(RAF)、DNN 模型要高。但显然，这样的预测模型不能令人满意。因此，我们对比下述两种预测股票涨跌的方案。

方案 1：将收盘价数据根据式(8)计算出每一时刻的收益率。用 4 个周期的收益率数据分别训练 4 个模型。

$$R_t=\frac{C_t}{C_{t-1}}-1 \tag{8}$$

方案 2 用收盘价来训练模型，预测出的收盘价再转换成收益率。

为了评估衡量网络性能，这里采用 4 个指标来衡量，即准确率(Accuracy)、精准率(Precision)、召回率(Recall)和 F 度量值(F-measure)，F 度量值是精度和召回之间的调和平均值。这些指标是根据真正性-TP、真负性-TN、假正性-FP、假负性-FN 计算的。

$$A=\frac{TP+TN}{TP+FP+TN+FN} \tag{9}$$

$$P=\frac{TP}{TP+FP} \tag{10}$$

$$R=\frac{TP}{TP+FN} \tag{11}$$

$$F1=2\frac{P\times R}{P+R} \tag{12}$$

我们分别在中证 500 数据集上测试了上述两种方案，表 4 是 4 个周期测试结果的平均值。可以看出，无论是方案 1 还是方案 2 在上涨类别的预测判断上都表现欠佳，在下跌判断上方案 2 明显优于方案 1。

表 4 模型指标的对比

Table 4 Comparison of indicators

(单位：%)

指标	类别	方案 1	方案 2
准确率	—	49.0	56.7
精准率	UP	29.1	27.1
	DOWN	68.2	68.9
召回率	UP	48.6	25.6
	DOWN	49.3	70.0
F1	UP	36.4	26.3
	DOWN	57.5	69.1

截取一段预测数据进行分析，图 6 是预测数据与真实值的对比，图 7 是将数据处理为涨跌信号后的对比结果(1 代表上涨，−1 代表下跌)。可以得到如下规律：当价格变动出现大的转折时预测误差较大，价格变动缓和时预测相对准确。从图 6、图 7 中看出，存在预测滞后真实值的情况。对此结果，我们做了理论分析，即出现这样的效果是训练不佳导致的。神经网络的瓶颈问题是陷入局部的极值，再进入的学习样本很难调整神经网络中的参数。而本文网络预测的结果受到最近的输入值影响较大，即预测出来的数值更接近当前输入的最后一个数值。而当价格出现大幅度波动时，自然会造成预测滞后于真实的现象。

图 6 2018 年 3 月 8 日模型预测结果部分数据对比(标准化数据)

Fig. 6 Partof model prediction results comparison 2018.3.8 (standardized)

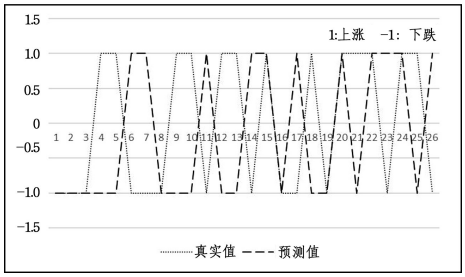

图 7 涨跌信号对比

Fig. 7 Stock signal comparison

4.3 改进方案

陷入局部极值这一问题属于神经网络的通病，还没有彻底解决的方法。针对股票预测的实际应用场景，可以通过调整一些预测参数来缓解这一问题。下面将具体描述这些参数在应用场景的含义。

4.3.1 场景描述

经过 LSTM 网络学习和预测，我们已经得到预测下一分钟的收盘价格，为了将预测的结果转化为涨跌信号，首先给出一些定义。记 X_{trend} 为涨跌信号，当 X_{trend} 为 1 时表示上涨，当 X_{trend} 为 0 时代表平稳，当 X_{trend} 为 −1 时代表下跌。T 时刻收盘价数据为 C_T。下面给出 X_{trend} 判断规则，在 t 时刻收盘价 C_t 与在 $t-\Delta t$ 时刻收盘价 $C_{t-\Delta t}$ 进行比较来决定 X_{trend} 的值，公式如下：

$$X_{\text{trend}}=\begin{cases}1, & C_t-C_{t-\Delta t}>\gamma\\ 0, & -\gamma\leqslant C_t-C_{t-\Delta t}\leqslant\gamma\\ -1, & C_t-C_{t-\Delta t}<-\gamma\end{cases}\tag{13}$$

其中，γ 是判断涨跌的阈值。

我们注意到：Δt，γ 是影响判断涨跌准确性的关键因子。因此可以把问题简化为优化问题，在合理的范畴内，寻找合适的 Δt、γ 使 X_{trend} 的判断尽可能准确。这样的操作可以忽略收盘价的小幅波动，能缓解滞后的问题。

4.3.2 参数搜索

Δt，γ 参数的选定对模型预测的时效性和参考价值有很大的影响。若 γ 阈值设置过大会降低模型对涨跌判断的敏感程度，即大部分被判断为平稳；Δt 设置过大，即预测的时间间距被拉长，可能错过交易时机，模型的参考价值降低。因此需要合理有效地搜索出参数，这里选用经典的搜索算法——遗传算法。

(1)遗传算法原理

遗传算法(GA)是一种常用的求解方法，通常用于优化工程、计算机科学、经济管理等领域的问题。与蚁群算法和模拟退火算法等进化算法不同，GA 通过选择、交叉和突变个体来选择最佳物种进行生物繁殖。其模拟生物的进化来解决优化问题，程序化流程如图 8 所示。GA 算法的核心在于将实际问题编码以便模拟生物的进化过程。为了模拟生态的进化过程，比较常见的方式是引入基因链的概念。经过编码把一个物种转换成一条基因链，两条基因链的部分基因进行交换重组，获得新的基因链，这样就完成一次物种进化。GA 算法成熟的编码方式有很多，如二进制编码、格雷码编码、排列编码等。我们选用二进制编码方式来描述实际问题，完成交叉变异过程。

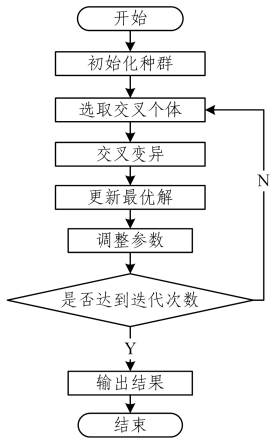

图 8 遗传算法示意图

Fig. 8 Sketch map of GA

(2)目标函数设计

为了更好地让 X_{trend} 给出涨跌信号，我们将目标函数定在精准率上。以真实的收盘价数据获得的涨跌信号作为标杆，与预测值给出的涨跌信号对比，计算精准率，其公式如下：

$$Precision=100\%\times\frac{\sum X_{\text{true\_trend}}\wedge X_{\text{predict\_trend}}}{X_{\text{predict\_trend}}}\tag{14}$$

其中，$X_{\text{true\_trend}}$ 表示真实情况的涨跌信号，$X_{\text{predict\_trend}}$ 表示模型预测出的涨跌信号，$Precision$ 表示精准率。

由于精准率由上涨判断、下跌判断、平稳判断 3 部分组成，笼统的计算会导致精准率在某一方面过低、过高或涨跌的样本分布异常。出现这种问题会导致给出的信号与真实情况相差很大，毫无意义。为此，我们在基础目标函数上做了一些特殊处理。

1)分开计算精准率。我们将上涨、下跌、平稳的信号分开，分别计算各自的精准率，再取三者的平均值作为目标函数。这种方法可以减小 3 种信号的精准率的差距，一定程度上解决涨跌样本分布异常的问题。

2)给出有效区间限定。当一种信号的精准率过低，说明我们的预测信号是不可信的，不希望出现在最终的方案当中。因此，我们做出有效区间限定，只有目标函数的值大于 60% 才会被记录，否则被直接放弃，即目标函数值为 0%。

3)Δt，γ 范围预处理。结合实际的应用情形和数据情况，我们可以给出一个更小的搜索范围，提高算法的效率。我们认为超出 10 min 的预测已经错过了最佳的交易时机，那么再进行研究就没有意义了，所以时间间隔，Δt 可以确定在 0～10 min 之间。阈值 γ 至少要在差值的最大最小值之间才有意义，由此阈值的初始值如下：

$$\gamma_0=\max_T|C_t-C_{t-\Delta t}|\tag{15}$$

其中，$T=t$ 时刻收盘价数据为 C_t。

综合上述的目标函数设计，我们缩小了遗传算法检索的范围，剔除了对于研究无意义的 Δt，γ 组合，能更准确地获得有效组合，以便给出准确的预测信号。目标函数如下：

$$Precision_{\text{stay}}(\Delta t,\gamma)=\frac{\sum X_{\text{stay\_true\_trend}}(\Delta t,\gamma)\wedge X_{\text{stay\_predict\_trend}}(\Delta t,\gamma)}{X_{\text{stay\_predict\_trend}}(\Delta t,\gamma)}\tag{16}$$

$$Precision_{\text{down}}(\Delta t,\gamma)=\frac{\sum X_{\text{down\_true\_trend}}(\Delta t,\gamma)\wedge X_{\text{down\_predict\_trend}}(\Delta t,\gamma)}{X_{\text{down\_predict\_trend}}(\Delta t,\gamma)}\tag{17}$$

$$Precision_{up}(\Delta t,\gamma)=\frac{\sum X_{up\_true\_trend}(\Delta t,\gamma)\wedge X_{up\_predict\_trend}(\Delta t,\gamma)}{X_{up\_predict\_trend}(\Delta t,\gamma)} \tag{18}$$

$$Precision(\Delta t,\gamma)=(Precision_{up}(\Delta t,\gamma)+Precision_{down}(\Delta t,\gamma)+Precision_{stay}(\Delta t,\gamma))/3 \tag{19}$$

$$Max\_f(\Delta t,\gamma)=\begin{cases}Precision(\Delta t,\gamma), & \begin{cases}Precision_{up}(\Delta t,\gamma)\geqslant 60\%\\ Precision_{down}(\Delta t,\gamma)\geqslant 60\%\\ Precision_{stay}(\Delta t,\gamma)\geqslant 60\%\end{cases}\\ 0, & \begin{cases}Precision_{up}(\Delta t,\gamma)<60\%\\ Precision_{down}(\Delta t,\gamma)<60\%\\ Precision_{stay}(\Delta t,\gamma)<60\%\end{cases}\end{cases}$$

$$\begin{cases}0<\Delta t<10, & \Delta t\in N\\ 0<\gamma<\max\limits_{T}|C_t-C_{t-\Delta t}|, & \gamma\in R\end{cases} \tag{20}$$

其中，$X_{up\_true\_trend}$，$X_{down\_true\_trend}$，$X_{stay\_true\_trend}$ 分别表示真实情况的上涨、下跌、平稳的涨跌信号，如 $X_{up\_true\_trend}$ 上涨记为 1，其余情况记为 0。同理，$X_{up\_true\_trend}$，$X_{down\_true\_trend}$，$X_{stay\_true\_trend}$ 表示模型预测出的上涨、下跌、平稳的涨跌信号。$Max\_f(\Delta t,\gamma)$ 表示以 $\Delta t,\gamma$ 组合为自变量的目标函数。

4.4 对比实验

将基于 GA 算法阈值优化的 LSTM 模型应用到中证 500 数据集上进行测试。将真实的涨跌信号和预测的信号进行对比，图 9 中 1 代表上涨，0 代表平稳，−1 代表下跌。可以看到参数调整后的涨跌信号波动频率降低了，随之预测信号的准确性也有所提升，之前 LSTM 模型存在的滞后问题虽然仍存在但有了明显改善。

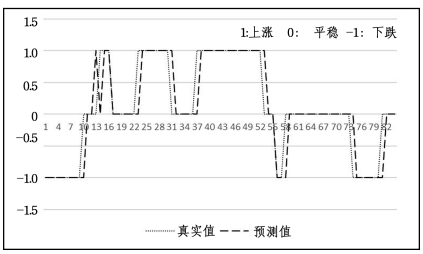

图 9 改进涨跌信号对比

Fig. 9 Improvement of stock signal comparison

对基于 GA 算法阈值优化的 LSTM 算法、LSTM 算法、随机森林 RAF 算法分别计算衡量指标并记录 4 个周期的平均值，结果如表 5 所列。

表 5 改进模型的指标对比

Table 5 Comparison of indicators of improvement model

（单位：%）

指标	类别	LSTM+GA	LSTM	RAF
准确率	—	89.4	56.7	67.4
精准率	上涨	93.5	27.1	65.3
	下跌	93.1	68.9	69.4
召回率	上涨	89.4	25.6	67.5
	下跌	89.2	70.0	67.5
F1	上涨	91.4	26.3	66.4
	下跌	91.1	69.1	68.4

RAF 算法输入数据以及样例格式与 LSTM 算法完全一致，RAF 配置参照了文献[4]。基于 GA 算法阈值优化的 LSTM 模型的指标要远高于 LSTM 和 RAF 算法。不过需要注意的是，由于基于 GA 算法阈值优化的 LSTM 模型参数的调整，预测值的含义有了变化。相比 LSTM 和 RAF 模型预测结果的参考价值有所下降。

对比 RAF 模型和 LSTM 模型，发现 LSTM 的回归预测偏向于下跌情况，各项指标下跌偏高，上涨偏低。而 RAF 更为均衡。基于 GA 算法阈值优化的 LSTM 模型有效改善了 LSTM 模型的不足。

结束语 金融市场预测问题一直是金融领域的核心问题。随着机器学习技术的进步，研究人员开始尝试将机器学习模型应用到金融领域来解决实际问题。借鉴之前的研究，我们从 LSTM 神经网络入手，构建了用于预测股票收盘价格的神经网络模型。结合实际应用场景，结合 LSTM 模型和 GA 算法来实现股票价格的涨跌预测。通过实验分析发现，LSTM 的预测在小幅变化的价格变动中表现良好，但在大幅价格变化时表现出滞后的现象。结合 LSTM 和 GA 的涨跌预测模型能较为准确地预测市场的涨跌情况，对投资有一定的参考价值。

未来的工作可以分为两部分。在 LSTM 预测中出现的滞后现象是机器学习技术的普遍问题。追溯根源是用于模型学习的样本不能适应训练模型。本文只采用了收盘价数据样本进行了训练，影响相对单一化，若加入成交量、最高价、最低价等数据，理论上会带来更好的效果，值得之后进行研究探索。另一方面，本文中的涨跌预测还有很大的局限性，仅预测了下一刻涨跌的情况，对于投资的参考价值有限。作为投资者，更希望预测出接下来一段时间市场的走势如何，对市场的变动进行更深入的探究。

致谢 感谢深圳市康曼德资本管理公司在本文完成过程中对数据、策略及评估等方面工作给予的指导和帮助！

参 考 文 献

[1] GANDHMAL D P, KUMAR K. Systematic analysis and review of stock market prediction techni-ques[J]. Computer Science Review, 2019, 34: 100-190.

[2] CHAKRAVARTY S, DASH P K. A PSO based integrated functional link net and interval type-2 fuzzy logic system for predicting stock market indices[J]. Applied Soft Computing, 2012, 12(2): 931-941.

[3] NI L, NI Z, GAO Y, et al. Stock trend prediction based on fractal feature selection and support vector machine[J]. Expert Systems with Applications, 2011, 38(5): 5569-5576.

[4] KRAUSS C, DO X A, HUCK N, et al. Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500[J]. European Journal of Operational Research, 2017, 259(2): 689-702.

[5] FISCHER T, KRAUSS C. Deep learning with long short-term memory networks for financial market predictions[J]. European Journal of Operational Research, 2018, 270(2): 654-669

[6] HENRIQUE B M, SOBREIRO V A, KIMURA H, et al. Literature review: Machine learning techniques applied to financial market prediction[J]. Expert Systems with Applications, 2019, 124(JUN): 226-251.

[7] HOCHREITER S, SCHMIDHUBER J. Long short-term memo-

ry[J]. Neural Computation,1997,9(8):1735-1780.
[8] GERS F A,SCHMIDHUBER,JÜRGEN,et al. Learning to Forget:Continual Prediction with LSTM[J]. Neural Computation,2000,12(10):2451-2471.
[9] GRAVES A. Supervised Sequence Labelling with Recurrent Neural Networks[M]. Springer,2012.
[10] QIN H M,SUN X. Classifying Bug Reports into Bugs and Nonbugs Using LSTM[C]// The Tenth Asia-Pacifc Symposium on Internet-ware. 2018.
[11] HUANG Y M,JIANG Y,HASAN T,et al. A Topic BiLSTM Model for Sentiment Classification[C]// Innovation in Artificial Intelligence (ICIAI). 2018.
[12] NELSON D M Q , PEREIRA A C M , OLIVEIRA R A D. Stock market's price movement prediction with LSTM neural networks[C]// International Joint Conference on Neural Networks (IJCNN). 2017.
[13] LIN M,CHEN C X. Short-term prediction of stock market price based on GA optimization LSTM neurons[C]// International Conference on Deep Learning Technologies (ICDLT). 2018.
[14] SHIU J N,ZOU J Z,ZHANG J,et al. Research of stock price prediction based on dmd-lstm model [J/OL]. Application Research of Computers. https://doi. org/10. 19734/j. issn. 1001-3695. 2018. 08. 0657.
[15] CHEN J,LIU D X,WU D S. Stock index forecasting method based on feature selection and LSTM model[J]. Computer Engineering and Applications,2019,55(6):108-112.
[16] HO T K. Random decision forests [C]// International Conference on Document Analysis and Recognition. 1995:278-282.
[17] CHOLLET F. Keras [EB/OL]. https://github. com/fchollet/keras,2016.
[18] Keras Documentation[EB/OL]. https://keras. io.
[19] GRANGER C W J. Strategies for Modelling Nonlinear Time - Series Relationships[J]. Economic Record, 2010, 69(3): 233-238.

BAO Zhen-shan, born in 1965, is a member of China Computer Federation. His main research interests include machine learning and Financial technology.

ZHANG Wen-bo, born in 1980, Ph. D, lecturer, is a member of China Computer Federation. Her main research interests include heterogeneous computing and trust computing.

(上接第 443 页)

[2] 王肇刚. 基于网络拓扑约束的时序数据挖掘算法研究与应用[D]. 北京:北京邮电大学,2009.
[3] HAN J W,KAMBER M. 数据挖掘概念与技术(原书第 2 版)(计算机科学丛书)[M]. 北京:机械工业出版社,2008.
[4] AGRAWAI R. Mining association rules between sets of items in large databases[C]// Proceedings of the 1993 ACM SIGMOD Conference. Washington,D C,1993:207-216.
[5] HAN J,PEI J,YIN Y. Mining frequent patterns without candidate generation[C]// ACM SIGMOD International Conference on Management of Data. ACM,2000:1-12.
[6] HATONEN K. Knowledge discovery from telecommunication network alarm databases[C]// ICDE 96. New Orieans,1996:115-122.
[7] NING P,CUI Y,REEVES D S,et al. Techniques and tools for analyzing intrusion alerts[J]. ACM Transactions on Information and System Security(TISSEC),2004,7(2):274-318.
[8] 刘冬生,曾小荟,唐卫东,等. 一种新的告警关联聚类算法[J]. 计算机应用研究,2013,30(12):3786-3789,3793.
[9] 陈兴蜀,何涛,曾雪梅, 等. 基于告警属性聚类的攻击场景关联规则挖掘方法研究[J]. 工程科学与技术,2019,51(3):144-150.
[10] 樊迪,刘静,庄俊玺, 等. 基于因果知识发现的攻击场景重构研究[J]. 网络与信息安全学报,2017,3(4):58-68.
[11] 冯学伟,王东霞,黄敏桓, 等. 一种基于马尔可夫性质的因果知识挖掘方法[J]. 计算机研究与发展,2014,51(11):2493-2504.
[12] KHOSRAVI-FARMAD M,RAMAKI A A,BAFGHI A G. Risk-based Intrusion Response Management in IDS using Bayesian Decision Networks[C]// 2015 5th International Conference on Computer and Kknowledge Engineering(ICCKE). 2015:307-312.
[13] RAMAKI A A,RASOOLZADEGAN A ,BAFGHI A G. A Systematic Mapping Study on Intrusion Alert Analysis in Intrusion Detection Systems[J]. ACM Computing Surveys,2018,51(3):55.

DENG Tian-tian, Ph.D, senior engineer. Her research interests include big data analysis and open source ecology.

XIONG Yin-qiao, Ph.D. His research interests include privacy preserving, information security, and the Internet of Things.