[文章编号] 1007－7405(2020)05－0385－07 DOI:10.19715/j.jmuzr.2020.05.11

股票信息挖掘与LSTM预测

陈伟斌[1]，林奕真[2]，王宗跃[2]

（1. 集美大学信息化中心，福建 厦门 361021；2. 集美大学计算机工程学院，福建 厦门 361021）

[摘要] 由于受到经济环境、政治政策、市场新闻等多种因素的影响，使得预测股票动态变得极具挑战性。研究了5种常用的预测股价变动的预测方法，通过逐步增加模型的输入维度进行预测分析。首先，建立5种优化的预测模型——基于时间序列的自回归平均模型（ARMA）、灰色预测模型（GM（1，1））、BP神经网络模型（BPNN）、基于改进网格寻优算法的支持向量回归（SVR）模型、基于Tensorflow的长短时记忆神经网络模型（LSTM），研究单一维度的模型输入，即，将各股票的收盘价作为这5种模型的输入。通过实验验证，发现基于LSTM的效果明显优于其他传统机器学习算法。然后，增加模型的输入维度进行研究，即，将影响股价的13个指标作为LSTM模型的输入来预测股价，所得的模型在训练集上的均方误差为0.1438。最后，进一步增加模型的输入维度，即，通过新闻数据挖掘提取14个特征，再结合13个股价指标，以这27个维度特征作为LSTM模型的输入来预测股价，所得的模型在训练集上的均方误差为0.1045。通过实验验证得出，所采用的输入27个维度的方法，比输入13个维度在预测问题上表现得更稳健。

[关键词] 股票预测；长短时记忆神经网络（LSTM）；回归分析

[中图分类号] TP 183

Stock Information Mining and LSTM Prediction

CHEN Weibin[1], LIN Yizhen[2], WANG Zongyue[2]

(1. Informatization Center, Jimei University, Xiamen 361021, China;
2. School of Computer Engineering, Jimei University, Xiamen 361021, China)

Abstract: Predicting the dynamic stock price is a challenging task under the influence of economic environment, political policies, market news and other factors. This paper focuses on three ways to predict the stock market, and gradually increases the input dimension of the model for analysis. First, It was considered that single dimension input(closing price) and adopt five optimized forecasting models, autoregressive moving average (ARMA) based on time series, grey prediction (GM(1,1)), back-propagation neural network(BPNN), support vector regression (SVR) based on improved grid search optimization algorithm, and long short-time memory (LSTM) based on Tensorflow. It is found that the Tensorflow based LSTM is significantly better than other traditional machine learning algorithms. Then, the input dimension of the model is added to study, namely, 13 indexes affecting stock price are used as input of LSTM model to predict stock price, and the mean square error of the model on the training set is 0.1438. Finally, the input dimension of the model is further increased. 14 features extracted by news mining and 13 stock price indexes are combined (27 dimensions) as input of LSTM

[收稿日期] 2019－09－23
[基金项目] 国家自然科学基金项目（61701191）；厦门市科技计划项目（3502Z20183032，3502Z20203057）
[作者简介] 陈伟斌（1972—），男，硕士，实验师，主要从事大数据挖掘、股票预测等方面研究。通信作者：王宗跃（1979—），男，博士，副教授，硕导，主要从事计算机视觉、大数据分析等方面研究。E-mail:wangzongyue@jmu.edu.cn

model to predict stock price, and the mean square error of the model on the training set is 0.1045. It is concluded that the 27-dimensions model is more robust than that of 13 dimensions.

Keywords: stock prediction; long short-term memory neural network(LSTM) ; regression analysis

0 引言

由于受到经济环境、政治政策、市场新闻等多种因素的影响，使得预测股票动态变得极具挑战性。1900 年，法国数学家 Louis Bachelier 率先研究股价行为的随机特征，并激发起了股价研究的兴趣。Fan 等[1]在 2003 年提出基于统计的方法进行预测的模型 ARMA（自回归平均模型）。然而，ARMA 模型假设滞后变量之间存在线性关系，因此对现实世界复杂系统能线性近似，但未能预测非线性和非平稳过程的演化。随着机器学习的发展，类似人类自学习的非线性人工神经网络具有良好的性能，可以用来预测金融时间序列，包括反向传播神经网络（BPNN）[2]、径向基函数神经网络[3]等。Qiu 等[4]使用遗传算法优化人工神经网络模型，预测日本股市指数价格走向。Kumar 等[5]通过离散小波变换将金融时间序列分解成 BPNN 的输入变量，以预测未来的股票价格。文献［6］提出了一种基于长短时记忆神经网络（long short-term memory neural network,LSTM）的股票收益预测系统，并在中国股市进行了测试。文献［7］用 LSTM 进行股票价格预测，通过股票的历史价格数据进行时序学习，实验结果证实了 LSTM 在时间序列预测上具有良好的性能，但并未考虑将不同类型的维度数据输入 LSTM 网络，使得模型过于简单，稳健性不足。因此本文通过网络爬虫获取 13 个影响股价的指标数据，再通过新闻挖掘提取 14 个特征，将这 27 个维度特征输入 LSTM 模型对股票进行动态预测，以期获得更好的预测效果。

1 LSTM 参数分析

由 Google 开发的开源人工智能库 Tensorflow 提供了一个表达机器学习算法的接口和执行这些算法的应用程序。本系统利用 LSTM 结合 TensorFlow 深度学习，对传统技术分析中的算法加以改进。本研究发现，当改变参数时，会显著影响神经网络的性能。因此，必须仔细设置参数，以确保 TensorFlow 的收敛性。

1）优化器

用同样大小的数据量[8]对目前深度学习训练中常用的优化器性能进行研究，研究结果见表 1。在实证研究中，Adam 优化器不仅在速度和熵损失方面都展示出了不错的性能，而且在预测模型算法的训练中也表现出了最好的性能。表 1 还表明，无论选择哪种优化器，只使用一个 CPU 进行 TensorFlow 训练大约需要 0.3 s（不包括初始化时间和数据加载时间），而使用 GPU 将进一步提高速度。Adam 表示自适应力矩估计，可代替经典随机梯度下降过程，应用于多层神经网络上的逻辑回归算法。由于 Adam 计算效率高，适合处理大数据集问题，超参数很少需要调整，且内存较小，因此适合用 Adam 优化器来更新基于训练数据迭代的网络权重。由此得出结论：对于大型模型和数据集，Adam 可以有效地解决实际的深度学习问题。

表 1 不同优化器的时间消耗和平均成本

Tab. 1 Time consumption and average cost of different optimizers

优化器 Optimizer	每次迭代训练平均时间 The average training time of each iteration/s	平均交叉熵损失 Average cross entropy
Adam	0.321	0.910
Adadelta	0.336	1.187
RMSProp	0.331	0.945
Momentum	0.316	1.010

2）学习率

在本研究中，对不同的学习率进行研究。较大的学习率导致梯度反弹和爆炸，较小的学习率使成本函数不能收敛到最小值。此外，一个小的学习率可能会导致梯度停留在局部最小值。表 2 显示了 TensorFlow 采用 Adam 优化器在不同学习率下的交叉熵损失[8]。可以看到学习速度在 0.001 到 0.005 之间通常会产生最好的结果且训练成本最低。因此，选择学习率为 0.001，因为它所产生的训练成本最低。

3）LSTM 叠置层数

通过叠加多层 LSTM，提取底层 LSTM 的隐藏状态信息，得到实例训练算法，取得了良好的效果。实验结果表明，单层 LSTM 模型的采样准确率为 0.66，2 层 LSTM 模型的样本准确率高达 0.78[9]。但是，由于受 tan h 的限制，LSTM 模型在超过 5 层之后需要消耗更多的计算资源。如果堆栈层更高，模型所消耗的计算资源就会增加。为了获得更准确的预测结果，有必要在可接受的时间范围内结合计算资源，选择具有适当堆叠层数的 LSTM 模型。因此，本文采用 2 层的叠置 LSTM 模型。

表 2　学习率及其相关成本比较

Tab. 2　Comparison of learning efficiency and cost

学习率 Learning rate	平均交叉熵损失 Average cross entropy
0.000 01	0.99
0.000 05	0.91
0.000 1	0.89
0.000 5	0.89
0.001	1.05
0.005	1.63
0.01	2.05
0.05	2.21

2　股票信息获取

2.1　影响股价的 13 个指标数据

通过编写爬虫程序，爬取 2015 年 1 月至 2019 年 5 月的各支股票每日的成交量、成交额、成交笔数、收盘价、市盈率、市净率、总股本、自由流通股本、总市值、流通市值、自由流通市值、自由流通市值_PIT、换手率等 13 个影响股价的指标数据，如表 3 所示。

表 3　影响股价的 13 个指标数据(000001)

Tab. 3　13 indicators of affecting stock price(000001)

日期 Date	成交量/股 Volume/Share	成交额/元 Turnover/Yuan	成交笔数/笔 NST	收盘价/元 CP/Yuan	市盈率 PE (TTM)	市净率 PB (LF)	总股本/股 Capitalization equity/Share
2019-05-17	96 500 085	1 208 569 909.00	48 837	12.44	8.321 3	0.939 9	17 170 411 366
2019-05-16	63 490 143	816 740 932.00	32 403	12.85	8.595 6	0.970 9	17 170 411 366
2019-05-15	110 398 851	1 417 050 112.00	48 666	12.92	8.642 4	0.976 2	17 170 411 366
2019-05-14	118 259 819	1 477 008 362.00	58 647	12.49	8.354 8	0.943 7	17 170 411 366
2019-05-13	74 191 779	916 213 719.00	40 192	12.30	8.227 7	0.929 3	17 170 411 366
2019-05-10	119 239 908	1 489 975 624.00	63 150	12.68	8.481 9	0.958 0	17 170 411 366
2019-05-09	155 715 161	1 904 905 488.00	70 453	12.16	8.134 0	0.918 7	17 170 411 366
2019-05-08	97 545 081	1 236 655 982.00	51 207	12.60	8.428 3	0.952 0	17 170 411 366
2019-05-07	107 265 841	1 386 247 078.00	50 312	12.95	8.662 5	0.978 4	17 170 411 366
2019-05-06	210 866 784	2 742 409 411.00	90 280	12.87	8.609 0	1.046 4	17 170 411 366
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2015-01-07	170 012 067	2 634 796 408.80	72 692	10.18	9.197 4	1.395 5	11 424 894 787
2015-01-06	216 642 140	3 453 446 167.70	80 109	10.38	9.375 7	1.422 5	11 424 894 787
2015-01-05	286 043 643	4 565 387 846.40	92 327	10.54	9.518 3	1.444 2	11 424 894 787

日期 Date	自由流通股本/股 FFE/Share	总市值/元 MC/Yuan	流通市值/元 CMV	自由流通市值/元 FCMV	自由流通市值_PIT FCMV_PIT/元	换手率 TR
2019-05-17	7 220 554 960	213 599 917 393.04	213 599 917 393.04	89 823 703 702.40	89 823 703 702.40	0.562 0
2019-05-16	7 220 554 960	220 639 786 053.10	220 639 786 053.10	92 784 131 236.00	92 784 131 236.00	0.369 8
2019-05-15	7 220 554 960	221 841 714 848.72	221 841 714 848.72	93 289 570 083.20	93 289 570 083.20	0.643 0
2019-05-14	7 220 554 960	214 458 437 961.34	214 458 437 961.34	90 184 731 450.40	90 184 731 450.40	0.688 7
2019-05-13	7 220 554 960	211 196 059 801.80	211 196 059 801.80	88 812 826 008.00	88 812 826 008.00	0.432 1
2019-05-10	7 220 554 960	217 720 816 120.88	217 720 816 120.88	91 556 636 892.80	91 556 636 892.80	0.694 5
2019-05-09	7 220 554 960	208 792 202 210.56	208 792 202 210.56	87 801 948 313.60	87 801 948 313.60	0.906 9
2019-05-08	7 220 554 960	216 347 183 211.60	216 347 183 211.60	90 978 992 496.00	90 978 992 496.00	0.568 1
2019-05-07	7 220 554 960	222 356 827 189.70	222 356 827 189.70	93 506 186 732.00	93 506 186 732.00	0.624 7
2019-05-06	7 220 554 960	220 983 194 280.42	220 983 194 280.42	92 928 542 335.20	92 928 542 335.20	1.228 1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
2015-01-07	4 961 088 120	176 857 371 302.76	176 857 371 302.76	76 797 644 097.60	72 607 982 500.08	1.728 3
2015-01-06	4 961 088 120	180 284 839 738.86	180 284 839 738.86	78 285 970 533.60	74 015 113 943.88	2.202 4
2015-01-05	4 961 088 120	183 026 814 487.74	183 026 814 487.74	79 476 631 682.40	75 140 819 098.92	2.907 9

2.2 新闻数据

通过编写爬虫程序，从各大门户网站上爬取股票新闻，利用 BeautifulSoup 解析 html 标签，将新闻分为日期、url、新闻标题、新闻正文、股票代码等数据结构进行存储，并用 countchn 正则表达式提取新闻中的中文，用 chardet 解决中文乱码问题，同时去掉网页中杂乱的“ < ”和“ > ”，将其替换成引号，即转换成字符串格式进行存储。以股票代码 000001 为例，爬取该只股票的新闻，时间范围从 2015 年 1 月到 2019 年 5 月，数据处理步骤主要有新闻分词、去停用词处理和特征提取两部分。

1）新闻分词与去停用词

根据股票代码解析日期、标题与正文，分类出利好与利空，以及中性消息。当新闻中出现“受益”“提升”“改善”“稳健”“看好”“有望”“收购”“利好”“优势”“增持”“回升”“利润增长”“提高”“拉升”等词语时，股票表现为利好，即有看涨趋势，保存为利好词库；反之，当新闻中出现“下滑”“低于”“下降”“拖累”“跌”“降”“亏损”“违规”“处罚”“利空”“减持”“调查”“质疑”“裁员”“远低”“或受影响”“悬念陡生”等词语时，股票表现为利空，可能会下跌，则保存成利空词库。利用 jieba. cut_for_search 将大段新闻进行分词处理，同时去除中文中常用的 1888 个停用词（如的、的话、等、等等、地）以减少无关信息的妨碍，提高分词的检索准确率与检索比对速度。

2）新闻特征提取

首先，将新闻按天分组，并计算当天的新闻数、当天新闻标题及正文的利好与利空数量，同时获取新闻最大最小时间；接着，分别进行标题的利好利空分组统计和正文的利好利空分组统计。经过上述一系列操作后，可提取出以下 14 个特征：

’day_sum’：当天新闻数；’1_x’：标题利好数；’2_x’：标题利空数；’3_x’：标题中性数；’1_y’：正文利好数；’2_y’：正文利空数；’3_y’：正文中性数；’week’：周一至周日；’good’：这支股票当天标题与正文利好新闻总数；’bad’：这支股票当天标题与正文利空新闻总数；’middle’：这支股票当天标题与正文中性新闻总数；’good_rate’：利好比例；’title_good_bad’：标题利好比利空；’article_good_bad’：正文利好比利空。

3 实验结果与分析

3.1 LSTM 与其他时间序列方法比较

在确定了每个预测模型最优参数的情况下，建立 ARMA(1,1)、GM(1,1)、BPNN、SVR、LSTM 模型。将所爬取的 2015 年 1 月至 2019 年 5 月的数据进行序列化、平稳化、数据过滤与清洗后，采用八二分的原则划分训练集和测试集，将单一维度的收盘价输入各个预测模型，模型输出为所预测的收盘价。通过实验验证，各模型预测结果与评价指标如表 4 所示。

表 4 各模型预测结果对比

Tab. 4 Comparison of forecast results of each model

股票代码 Stock code	ARMA_RMS	GM_GOF	BP_RMS	SVR_RMS	LSTM_RMS
000001	0.117	0.600	0.794	0.342	0.014
000002	0.729	0.666	0.822	1.227	0.023
000004	1.906	0.866	0.796	1.753	0.028
000006	0.070	0.800	0.843	0.499	0.036
000007	0.075	0.633	0.819	0.748	0.019
000008	0.138	0.433	0.883	0.626	0.027
000011	0.167	0.900	0.800	0.676	0.063
000012	0.054	0.666	0.877	0.392	0.021
000014	0.073	0.800	0.849	1.065	0.037
000016	0.070	0.533	0.817	0.341	0.022

由表 4 的结果分析可知：基于 Tensorflow 的 LSTM 模型在预测中的表现都明显优于其他传统的机器学习模型和算法。

3.2　基于 LSTM 的指标预测

1）数据集的准备

将所爬取的数据分成两部分：一部分是从 2015 年 1 月至 2019 年 5 月按时间顺序的前 80% 的每支股票的指标数据，将其作为训练集用于训练模型，并取训练集后 175 个数据作为验证集以调整 LSTM 模型的超参数（如迭代次数 epochs、LSTM 隐藏层神经元个数等）；另一部分是后 20% 的指标数据，将其作为测试集用于预测。

2）数据平稳处理

将爬取的原始数据经过一阶差分处理使得序列平稳。因为只有在序列平稳的前提下，才可进行后续研究。图 1 为数据进行平稳处理的结果图。

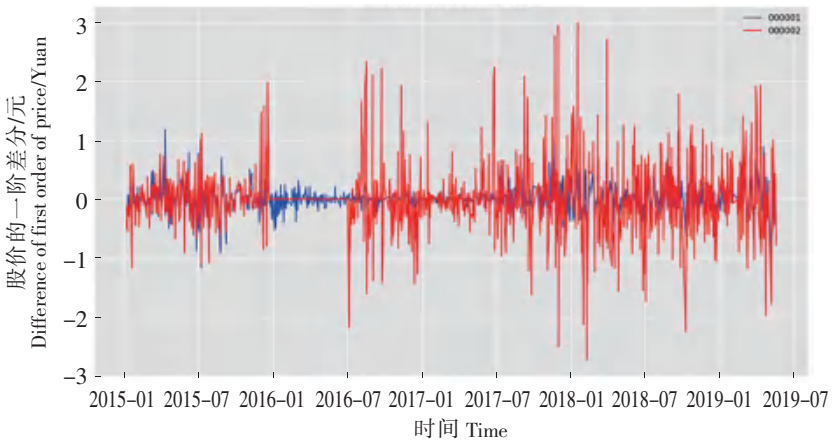

图 1　数据平稳处理结果

Fig.1　The result of data stabilization

3）模型训练过程

将训练集中经平稳处理并归一化后的各支股票的 13 个指标输入 LSTM 模型中。在 LSTM 模型参数中，设置时间步长 time_step = 5，批处理 batch_size = 25，输入维度 data_dim = 13。训练集与验证集的损失函数（均方误差）如图 2 所示。

4）评估模型

以均方误差作为所建立的 LSTM 模型的评价指标。实验结果（见图 3）表明，在训练集上预测的均方误差为 0.1438，股价的预测值和真实值走势较吻合，且评价指标数值较接近于 0。由此可见，本研究所建立的 LSTM 模型性能较好。

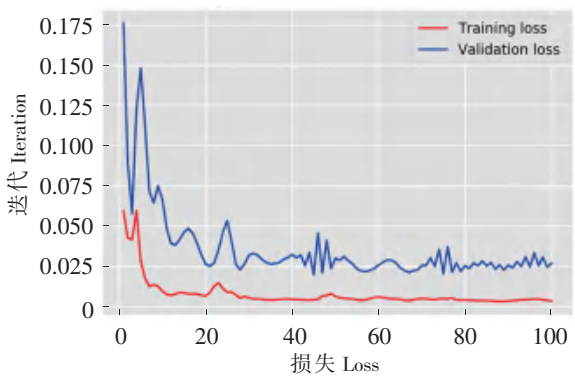

图 2　训练集和验证集的损失函数(13 维)

Fig.2　Loss function of training set and validation set(13 dimensions)

图 3　训练集股价预测值与真实值对比图(13 维)

Fig.3　Comparison of the stock price forecast value and real value of the training set(13 dimensions)

5）模型预测

训练好 LSTM 模型之后，将测试集用于股价的预测。预测实验结果（见图 4）表明，在测试集上预测的均方误差为 0. 3205。由此可见，本研究所建立的 LSTM 模型对于输入多维度影响股价指标来预测股价时效果较好。

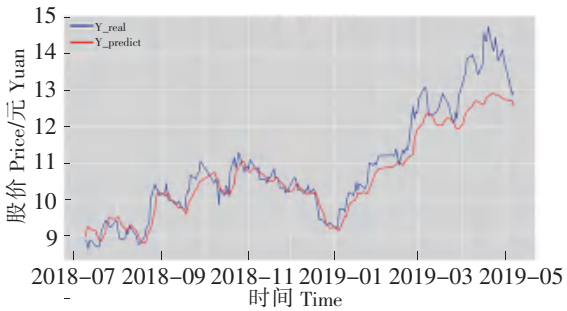

图 4 测试集股价预测值与真实值对比图(13 维)

Fig.4 Comparison of the stock price forecast value and real value of the test set(13 dimensions)

3. 3 基于新闻挖掘的股市预测

1）LSTM 模型训练

在该实验中将 13 个股价指标和新闻数据中所提取的 14 个特征作为训练维度特征。训练集与测试集也是采用八二分原则，将 27 个维度特征输入 LSTM 模型进行训练。该方法在训练集上预测的均方误差为 0. 1045，训练集与验证集的损失函数（均方误差）如图 5 所示。通过实验验证，对比训练集股价预测值与真实值（见图 6），可见其预测值和真实值吻合度较高。

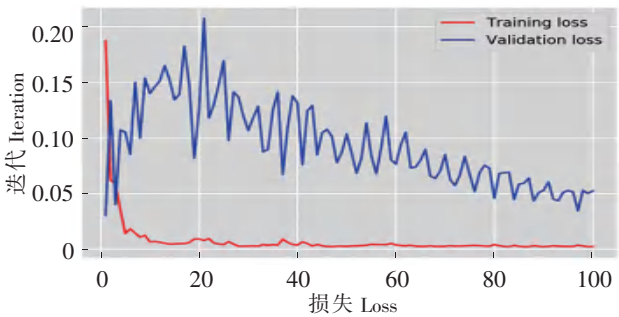

图 5 训练集和验证集的损失函数(27 维)

Fig.5 Loss function of training set and validation set(27 dimensions)

图 6 训练集股价预测值与真实值对比图(27 维)

Fig.6 Comparison of the stock price forecast value and real value of the training set(27 dimensions)

2）模型预测

经过新闻文本挖掘预测股价以验证和测试该模型，在测试集上预测的均方误差为 0. 2617，预测结果如图 7 所示。相比于 13 维度的 LSTM 模型输入，通过新闻挖掘提取特征增加模型的输入维度，使得模型在预测问题上表现得更稳健。

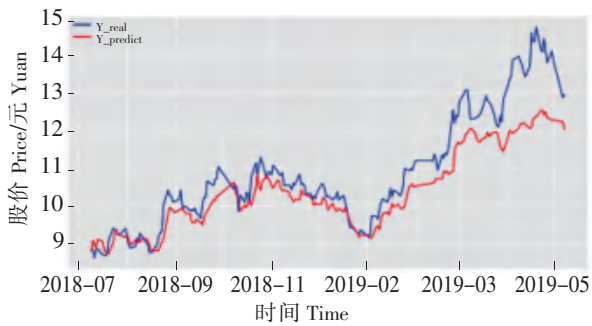

图 7 测试集股价预测值与真实值对比图(27 维)

Fig.7 Comparison of the stock price forecast value andreal value of the test set(27 dimensions)

4 结论

本文比较研究了 5 种模型及 3 种输入维度对股市预测的效果。首先，在单一维度的模型输入中的研究发现，由于 LSTM 是基于时序特征的长短时记忆模型，不同于传统的不含时序特征的 BPNN，因此基于深度学习的 LSTM 在处理连续时间序列的股价预测问题上的精度明显优于传统的机器学习算法。接着，经过实验发现，当成交量、自由流通股本等波动较大时，单一维度的收盘价模型对股价的波动不敏感，因此考虑增加模型的输入维度。于是将 13 个影响股价的指标输入 LSTM 模型，从而与股价建立关联并展开研究。然而，通过不同股票和不同交易日的研究分析，发现这 13 维指标只能反映出股票投资者对股价波动的影响，而当公司出现管理层组织架构变动或突发新闻事件时，股价会出

现不在股民控制范围内的不定性波动。因此，本研究再通过新闻挖掘的方式提取 14 个特征再输入 LSTM 模型，对股市进行进一步的预测。这使得本研究能预测突发状况对股市的影响，帮助股票投资者和公司提前做好应对准备，以降低经济损失。

由于股票实时趋势还受许多其他外在因素影响，因此，如何分析不同用户类型和用户情绪从而得出更加合理的股票投资组合模型，是本研究接下来的研究重点。虽然本文的动态预测模型包含历史序列中的金融危机等突发状况，但对研究金融风险的能力依然有限，还需继续研究新的理论方法来解决。例如，在未来工作中，可采用混合模型来应对具有线性和非线性变量的复杂股价时间序列等。

[参 考 文 献]

[1] FAN J, YAO Q. Nonlinear time series: nonparametric and parametric methods [M]. New York: Springer-Verlag, 2003.

[2] KARA Y, BOYACIOGLU M A, ÖMER KAAN BAYKAN. Predicting direction of stock price index movement using artificial neural networks and support vector machines: the sample of the istanbul stock exchange [J]. Expert Systems with Applications, 2011, 38(5): 5311-5319.

[3] SHEN W, GUO X, WU C, et al. Forecasting stock indices using radial basis function neural networks optimized by artificial fish swarm algorithm [J]. Knowledge-Based Systems, 2011, 24(3): 378-385.

[4] QIU M, SONG Y. Predicting the direction of stock market index movement using an optimized artificial neural network model [J]. PLOS ONE, 2016, 11(5): e0155133.

[5] KUMAR C S, SUMATHI M, SIVANANDAM S. Prediction of stock market price using hybrid of wavelet transform and artificial neural network [J]. Indian J Sci Technol, 2016, 9(8): 1-5.

[6] CHEN K, ZHOU Y, DAI F. A LSTM-based method for stock returns prediction: a case study of China stock market [C] //IEEE International Conference on Big Data (Big Data). Santa Clara: IEEE, 2015: 2823-2824.

[7] BAO W, YUE J, RAO Y L, 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.

[8] SANG C, PIERRO M D. Improving trading technical analysis with tensor flow long short-term memory (LSTM) neural network [J]. The Journal of Finance and Data Science, 2019, 5(1): 1-11.

[9] LIU S, LIAO GZ, DING Y F. Stock transaction prediction modeling and analysis based on LSTM [C] //13th IEEE Conference on Idustrial Electronicus and Applications (ICIEA). Wuhan: IEEE, 2018: 2787-2790.

(责任编辑　朱雪莲　英文审校　黄振坤)