

DOI:10.13232/j.cnki.jnju.2021.01.008

基于多视角股票特征的股票预测研究

李金轩, 杜军平*, 薛 哲

(智能通信软件与多媒体北京市重点实验室, 北京邮电大学计算机学院, 北京, 100876)

摘 要: 股票价格预测是金融行业中的一个重要研究内容, 能够更准确地分析股票价格走势对于投资机构至关重要。目前, 关于自动化预测股票价格发展的研究工作相对较少, 还有许多问题需要解决。针对传统股票预测方法中视角单一、无法充分考虑数据的各特征重要度的问题, 提出一种基于多视角股票特征的股票预测方法, 通过计算股票数据的 Ma, Macd, Kdj, Boll 特征指标, 训练每个指标下的弱学习器, 并进行多个弱学习器的集成学习, 最终用于预测股票价格走势。使用美国股票新闻数据集进行验证。结果表明, 基于多视角股票特征的股票预测方法预测得到的股票价格与实际价格之间的平均误差与均方误差分别为 1.9321 和 0.0581, 优于传统的基于单一指标的股票预测结果。

关键词: 股票预测, 多视角, 弱学习器, 集成学习

中图分类号: TP391

文献标识码: A

Research on stock forecasting based on multi-view stock features

Li Jinxuan, Du Junping*, Xue Zhe

(Beijing Key Laboratory of Intelligent Telecommunication Software and Multimedia, School of Computer Science,
Beijing University of Posts and Telecommunications, 100876, Beijing, China)

Abstract: Stock price forecasting is an important research content in financial industry, and it is very important for investment institutions to analyze the stock price trend more accurately. At present, there are relatively few researches on the automatic forecasting of stock price development, and many problems need to be solved. Among them, the perspective of traditional stock forecasting methods is single so that we can't fully consider the importance of each feature and the main features of the data. In this paper, we propose a stock forecasting method based on multi-view stock features. By calculating the Ma, Macd, Kdj, Boll feature indexes of stock data, we train the weak learners under each index, and carry out the integrated learning of multiple weak learners. Finally, the stock price trend is predicted. The stock news data set of the United States is used for verification. The experimental results show that the average error and the mean square error between the stock price and the actual price predicted by the stock forecast method based on multi-view stock features are 1.9321 and 0.0581, respectively. It is superior to traditional stock forecast results based on a single index.

Key words: stock forecast, multi-view, weak learner, ensemble learning

人们对一个事物的全面认识往往需要从多个视角获取, 通过多个视角的观察可以形成事物的多视角数据, 而每个视角的数据又包含多个维度的特征, 最终形成多视角特征数据。在很多实际

场景中, 多视角特征数据的应用十分广泛。例如, 机械工业生产中不仅要考虑生产时的物理数据, 也要考虑市场行情等数据; 而对于股票预测, 在长期的投资过程中, 人们对股票分析有一些经验, 形

基金项目: 国家自然科学基金(61902037, 61772083, 61802028), 广西科技重大专项(桂科 AA18118054)

收稿日期: 2020-09-16

* 通讯联系人, E-mail: junpingdu@126.com

成了一些如Ma(平均移动线指数)、Macd(异同平均移动线指数)、Kdj(随机指标)、Boll(布林线指标)等数据指标,很多股票软件将这些数据指标可视化,人们根据可视化指标可以更好地指导股票投资。因此,基于多个视角的特征数据可实现基于多视角特征数据的股票预测研究,而且多视角特征数据可以较全面地表示股票数据的相关信息,从而通过多视角特征数据进行股票预测的深入研究。

股票预测研究是金融大数据领域的一个重要研究方向。股票预测的精度受到如市场信心、法规政策等外界因素^[1]以及股票历史数据等内在因素的影响,同时股票预测可为股民的股票投资活动提供建议,促进经济的发展和社会的进步。因此,股票预测问题有很强的研究与现实意义^[2]。

目前关于股票预测的研究还存在以下问题:数据相关特征的一次性整体描述,目前股票预测的研究是将股票数据的所有特征整体作为输入变量,接着用机器学习方法来进行训练得到相应的模型,并得到最终预测结果,股票数据的特征没有被分层或分类看待,不便于理解,也无法满足股民的有时只想关注某个或某些特征的需求;目前股票预测的研究大多是利用股票的特征数据训练模型后直接给出预测的结果,无法给出数据的各特征重要度以及主要特征的参考意见,也就无法指导股民投资应关注的主要方向。

针对以上问题,本文提出一种基于多视角特征数据的股票预测方法,通过多个独立的长短期记忆力网络(Long and Short Term Memory network, LSTM)计算各个视角的股票数据,并通过集成学习自适应不同视角对预测结果的贡献,最终得到多视角股票数据的预测结果。该方法可以计算不同视角对最终预测结果的贡献程度,提高了最终的预测精度。

1 相关工作

近些年来,国内外学者使用各种各样的方法进行股票的预测研究,也提出了多种股票预测模型。目前,国外针对股票预测的研究,从数据内容来看主要分两个方面:一是根据不同时间股票价格数据或者与其相关的指标数据建立预测模型,

再进行拟合预测的工作;二是根据股票舆情数据并通过自然语言处理与挖掘的方法进行建模并实现预测分析,舆情数据包括经济新闻、财经新闻、股票评论、社交媒体等数据。从采用的预测方法或技术来看:一是通过神经网络、支持向量机(Support Vector Machine, SVM)等单一机器学习方法或多种机器学习方法混合模型;二是使用自然语言处理和文本数据挖掘进行处理。具体研究情况如下。

Seker et al^[3]关注土耳其的主流媒体发布的经济新闻数据和股票市场收盘价之间的关系。首先通过多种特征提取方法提取股票市场的价格演变数据与经济新闻的文本数据特征,并通过SVM算法、K近邻(K-Nearest Neighbors, KNN)算法与决策树三种分类器进行分类,最后通过各个分类器进行投票得到最终结果。Schumaker and Chen^[4]利用KNN、决策树和SVM等方法对股票的价格数据与相关行业部门分类的财经新闻数据进行研究以进行股票价格的预测,使用该系统进行投资的回报率高于股票交易专家。Boyacioglu and Avci^[5]利用基于模糊推理系统的自适应网络进行股票市场收益预测研究,输入六个经济变量和三个指数,通过基于模糊推理系统的自适应网络进行相应的建模与预测。Park and Shin^[6]认为股票价格的变化受多种因素的影响,如市场政策、汇率、银行存贷款利率、全球市场股价以及国际经济发展情况,这些因素之间存在关联性,可通过半监督学习构建基于这些因素的结点和它们之间的相似性的边缘网络,并对全球经济指数和200家上市公司的2007年1月至2008年8月的股票价格进行了预测。Guresen et al^[7]评估了动态和静态的神经网络模型在股票市场预测中的有效性,通过使用纳斯达克证券交易所的股票数据来比较多层感知器、动态人工神经网络和结合广义自回归条件异方差的混合神经网络这三个模型的预测结果。

国内相关的股票预测的研究主要是把开盘价、收盘价等通过其产生的数据整体作为研究对象,再利用KNN、神经网络、SVM和它们的优化方法来进行股票预测研究。王波和程福云^[8]通过KNN算法,基于前 n 个交易日的股票历史收盘价

预测 $n+1$ 日的股票价格. 黄秋萍等^[9]通过 BP 神经网络与小波神经网络, 利用开盘价、最高价、收盘价等股票数据, 建立对应的股票价格预测模型, 实现根据前一交易日的股票数据预测当天交易日的收盘价. 任治国^[10]、孙秋韵^[11]、杜晓^[12]分别对灰色模型、遗传算法与 BP 神经网络进行改进, 同样通过股票历史数据训练模型来预测当日的股票收盘价. 文献[13—15]利用 LSTM 网络对股票的时序信息进行建模, 获取股票变化的潜在规律, 然后对股票价格的变化进行预测.

2 基于多视角特征的预测模型的提出

本节主要介绍基于多视角特征的股票预测模型: 2.1 给出模型所需的定义, 2.2 描述如何构建多视角股票数据弱学习器, 2.3 描述多视角集成预测模型的构建.

2.1 基本定义 人们往往通过多个视角来客观全面地认识事物, 这些视角包含的信息相互补充. 不同视角包含的信息可以通过不同特征来描

述, 从而形成针对事物的多视角特征数据 (Multi-View Feature Data, MVFD). 多视角特征数据就是描述事物的多个视角的多个特征. $V_i = (f_1^{V_i}, f_2^{V_i}, \dots, f_j^{V_i})$ 表示第 i 个视角的特征, $f_j^{V_i}$ 表示第 i 个视角中的第 j 个特征, 则多视角特征数据可以表示为 $X = (V_1, V_2, \dots, V_n)$.

为了研究影响不同的视角对股票预测结果的重要度, 将每个视角的特征数据本身作为一个整体输入学习器中, 再对得到的多个输出结果进行集成学习, 从而得到最终结果. 通过对训练样本集进行分类采样得到 N 个基于视角分类的样本采样机; 对这 N 个基于视角分类的样本集分别使用学习器进行训练, 得到 N 个弱学习器; 再通过结合策略将 N 个弱学习器进行整合, 得到最终的强学习器. 结合策略的选择往往根据解决的问题不同而有区别, 解决回归预测问题时结合策略应选择平均法, 而解决分类预测问题时结合策略应使用投票法. 基于多视角特征数据集成学习的 MVLSTM-IL 模型结果如图 1 所示.

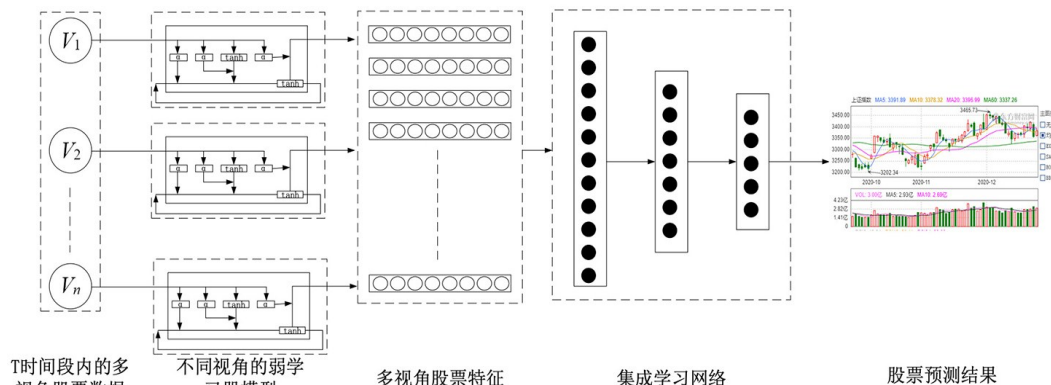


图 1 MVLSTM-IL 模型结构图

Fig. 1 The diagram of MVLSTM-IL model

2.2 多视角股票数据弱学习器的构建 由于股票数据是一种连续、非线性的数据, 股民投资时往往会参考过去一段时间这只股票的涨跌情况, 所以弱学习器的选择应该尽量覆盖一段时间内的股票数据, 并通过过往的数据来预测股票下一段的数据. 因此弱学习器的构建使用 LSTM 网络. 对于第 i 个视角的特征数据集, 初始化第一层 LSTM 网络的隐含层, 该层输入为 t 时刻前的一个视角的股票特征, 通过一次线性变换与 ReLU

函数激活, 将输入的股票特征数据投影为维度为 d 的隐含层, 计算方法如式(1)所示:

$$h_0 = \text{ReLU}(W_0 \times V_0 + b_0) \quad (1)$$

其中, W_0 与 b_0 是待学习的线性变换的参数. 式(1)得到的结果即为初始化后的隐含层, 之后每次迭代的 LSTM 弱学习器模型输入包括两部分, 分别是 V_i 和 h_i , h_i 表示 $t-1$ 时刻的 LSTM 网络的隐含层, V_i 表示 t 时刻该视角下的股票数据. 将二者

连接组合输入到 t 时刻的LSTM网络中,经过预设的次数迭代后,得到最终的隐含层 h_t .具体计算方法如式(2)所示:

$$h_t = \text{LSTM}(h_{t-1} + V_i) \quad (2)$$

由于LSTM网络每次迭代后的结果都会影响最终的模型输出,故将每次迭代生成的中间结果进行合并计算损失,选择均方差作为损失函数.损失函数的形式如式(3)所示:

$$\text{loss} = \frac{1}{n} \sum_{i=1}^n \left(\frac{\sum_{i=1}^n h_i}{t} - \text{out} \right)^2 \quad (3)$$

其中, h_i 表示LSTM网络每次迭代产生的中间结果, t 表示迭代次数, n 表示每次投入训练的batch大小.通过训练完成后的多视角股票数据弱学习器得到不同视角下的弱学习器生成的结果,组合成一个 $n \times d$ 维矩阵用于接下来的集成预测, n 表示参与预测过程的视角数量, d 表示每个视角下结果特征的维度.

2.3 多视角集成预测模型的构建 经过以上多个弱学习器模型后,本文算法已经针对各个视角给出了该视角下的股票指数预测.传统方法中将多个学习器的结果进行集成时,对于分类问题主要通过投票法,即将多个学习器给出的分类结果进行投票,选择票数最高的分类作为最终的分类结果;而对于回归问题,往往采用加权平均法,将多个学习器得到的结果进行加权平均得到最终的结果.由于股票的多视角数据的特性,直接进行加权平均得到的预测结果并不合适.本研究在多视角股票数据弱学习器得到的结果的基础上使用一个深度神经网络模型拟合多视角股票特征与股票价格之间的映射关系,最终得到具体的股票数据变化预测.第一层为输入层,将多视角股票数据弱学习器模型生成的矩阵输入网络;最后一层为线性输出层,将深度神经网络的结果输出成股票价格预测值;中间为隐含层.隐含层节点若数量过多会使网络的复杂度增高,甚至出现过拟合,而若节点数过少则会使结果难以收敛.本文采用经验公式得到隐含层的节点数,如式(4)所示:

$$N = \sqrt{m+n} + \alpha \quad (4)$$

其中, m 为输入层节点数, n 为输出层节点数, α 为 $[1, 10]$ 之间的常数.

对于集成预测模型,输入样本为四个指标的预测值 x_i ,对应的隐含层和输出层分别为 z_i 和 y_i ; V_{ij} 是输入层和隐含层之间的权重, V_{jk} 是隐含层和输出层之间的权重, θ_i 和 γ_i 分别为隐含层和输出层的阈值.本文使用的激励函数如式(5)所示,输出层结果如式(6)所示:

$$f = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (5)$$

$$y_i = f\left(\sum_{j=1}^N V_{jk} \times z_i + \gamma_i\right) \quad (6)$$

本文的股票预测过程如算法1所示.

算法1 基于多视角股票特征的股票预测过程

输入: 美国股票新闻数据集

输出: 每只股票均价

Step 1. 对于数据集中每只股票;

Step 2. 根据股票原始数据计算 Ma, Macd, Kdj, Boll 指标作为各视角原始数据;

Step 3. 将各个视角输入各自的弱学习器模型, 根据时间顺序进行迭代训练, 得到各视角的预测结果;

Step 4. 通过均方差损失函数计算损失, 重复 Step 3 和 Step 4 直到完成弱学习器的预训练;

Step 5. 将各视角的弱学习器模型的输出结果组合为多视角股票特征矩阵;

Step 6. 将特征矩阵输入多视角集成预测模型中, 得到最终的预测结果;

Step 7. 通过均方差损失函数计算损失 loss , 反馈调整参数;

Step 8. 重复 Step 5 至 Step 8, 直到预测值与实际值的误差小于设定的阈值.

3 实验与结果

本文使用的数据集为美国股票新闻数据集, 该数据集涵盖了自2008年6月8日到2016年7月1日的道琼斯工业指数股票数据, 包含600只股票, 记录每只股票的日期、开盘价、收盘价、当日最高与最低价, 平均每只股票记录天数为2613 d, 故最终的数据集包括1567800条数据. 经过清洗与筛选, 本文将该数据集分为训练集和测试集, 分别包含600000条和300000条数据. 根据股票领域一些专业指数的定义与计算方法, 将原始数据信息进行计算, 得到四种股票指数, 分别为 Ma, Macd, Kdj, Boll 并以此作为四种股票视角.

实验环境为运行于 Ubuntu 下的 tensorflow 深度学习框架,配置 NVIDIA CUDA 9.0, cuDNN 7.5 深度学习库加速 GPU 计算;Python 版本为 Python 3.5.

现有的预测算法的评测标准主要为计算预测结果与真实数据之间的误差和与均方误差,在通常标准上,本文额外给出不同指数的预测值与真实值的误差.

3.1 主要参数设置 在进行实验之前,将数据集包含的开盘价、收盘价、当日最高与最低价以及基于以上数据计算得到 Ma, Macd, Kdj, Boll 指数进行归一化,将不同范围的数据映射到(0,1)之间. Ma, Macd, Kdj, Boll 指数分别包括 ma10, ma20, ma60; macd. dea, macd. diff, macd. macd; kdj. k, kdj. d, kdj. j; boll. lower, boll. mid, boll. upper 各三种特征. 因此各视角的输入维度为 1×3 维.

为了最后合并计算损失,在训练过程中将弱学习器中隐含层节点的数目设置为 3. 对于集成预测模型,经过实验验证,输入层神经元数量设置为 12,隐含层层数设置为 1,隐含层节点数量设置为 6.

3.2 实验和结果 在训练弱学习器时采用 Adam 作为优化器,学习率设置为 0.01,权重衰减设置为 0.001,随机从数据集中选择了 60 只股票进行训练,30 只进行测试. 测试结果如图 2 所示. 各个视角的弱学习器模型在四种指标下的预测结果如图 3 所示. 可以看出, MVLSTM-IL 算法在 Ma, Macd, Boll 三种指标的预测结果与原始数据比较接近,并且通过 MVLSTM-IL 算法,使用多视角股票特征数据得到的七日内股价预测与真实

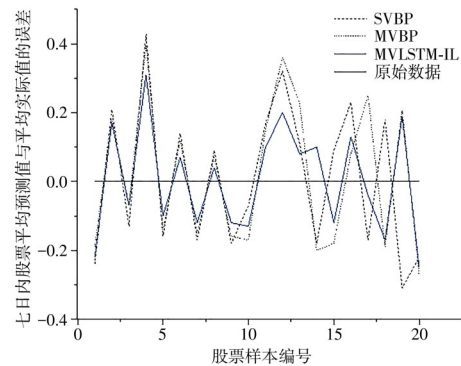


图 2 三种方法下股票的七日预测平均值与真实值的误差
Fig. 2 The error between the seven days forecast average value and the real value of stocks by three methods

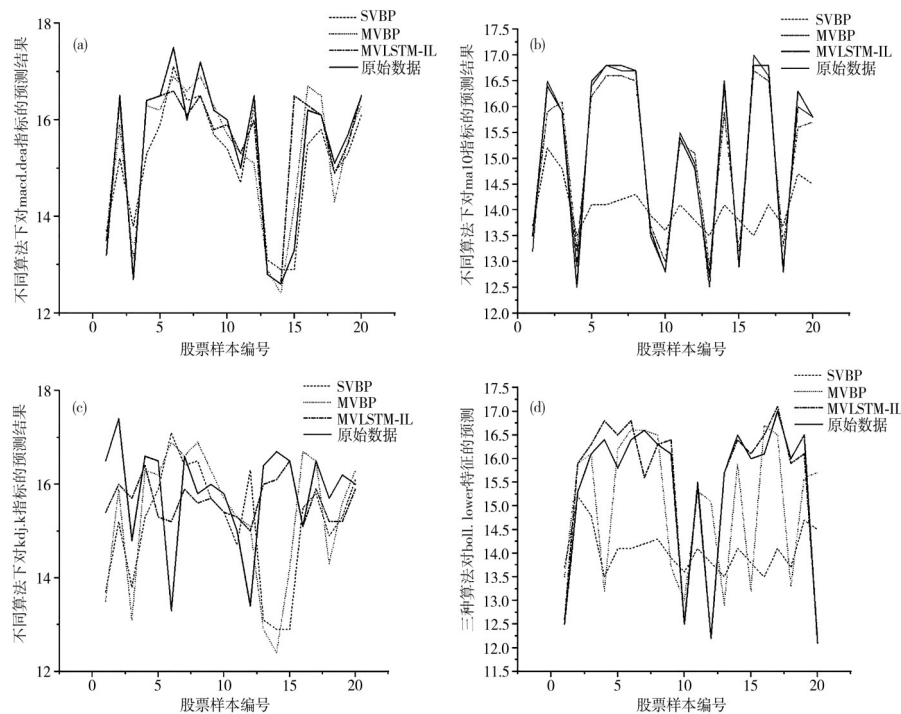


图 3 三种方法下股票的四指标预测平均值与真实值的误差

Fig. 3 The error between the predicted average value and the real value of four indexes of stock by three methods

数据的平均误差和直接使用股价这一单一视角或者使用BP神经网络作为弱学习器的预测模型相比,得到的平均误差更小,更接近真实的股价。几

种算法在股价的误差和、均方误差以及各指标平均误差的对比结果如表1所示。

表1 不同算法在美国股票新闻数据集下的评价指标

Table 1 Evaluation indexes of different algorithms in American stock news dataset

算法	平均误差和	均方误差	Ma 平均误差	Macd 平均误差	Kdj 平均误差	Boll 平均误差
SVBP	6.56	0.2399	10.2858	24.764	54.3147	14.2712
MVBP	2.5488	0.142	8.1762	19.803	47.2732	13.7837
Multinomial	5.57	0.1957	9.1331	21.533	51.4721	14.5742
MVLSTM-IL	1.9321	0.0581	7.8876	16.2733	46.8872	13.2231

4 总 结

本文在现有研究的基础上设计了基于多视角股票特征的股票价格预测模型。针对传统方法中数据相关特征的一次性整体描述无法分清多个视角对预测结果的贡献的问题,将视角的概念引入到股票数据的特征表示中,通过对多个股票指标建立基于LSTM网络的弱学习模型得到预测结果并对股票的多视角特征进行集成学习,得到最终的股票价格预测结果。基于美国股票新闻数据集,自2008年6月8日到2016年7月1日的道琼斯工业指数股票数据进行实验,本文模型得到的最终股票预测结果与真实的股票价格间的七日平均误差和与均方误差比直接对一体化的历史股价进行预测与使用多视角历史数据直接预测股票价格得到的七日平均误差和与均方误差都要小,证明本研究提出的模型可以更好地预测股价的波动,还发现Ma,Macd,Kdj,Boll四种股票价格评价指标中,Ma,Macd,Kdj指标对于最终股价预测准确度的影响比Boll更大。

参考文献

- [1] 崔东东. 基于深度信念网络的股票价格预测研究. 硕士学位论文. 武汉: 华中科技大学, 2016. (Cui D D. A study on the prediction of stock price based deep belief networks. Master Dissertation. Wuhan: Huazhong University of Science and Technology, 2016.)
- [2] 李鹏. 中国股市影响因素趋势分析. 时代金融, 2014(8): 153—154. (Li P. Analysis on the trend of influ-

encing factors in China's stock market. Times Finance, 2014(8): 153—154.)

- [3] Seker S E, Mert C, Al-Naami K, et al. Correlation between the economy news and stock market in turkey. International Journal of Business Intelligence Research, 2013, 4(4): 1—21.
- [4] Schumaker R P, Chen H. A quantitative stock prediction system based on financial news. Information Processing & Management, 2009, 45(5): 571—583.
- [5] Boyacioglu M A, Avci D. An Adaptive Network-Based Fuzzy Inference System (ANFIS) for the prediction of stock market return: the case of the istanbul stock exchange. Expert Systems with Applications, 2010, 37(12): 7908—7912.
- [6] Park K, Shin H. Stock price prediction based on a complex interrelation network of economic factors. Engineering Applications of Artificial Intelligence, 2013, 26(5—6): 1550—1561.
- [7] Guresen E, Kayakutlu G, Daim T U. Using artificial neural network models in stock market index prediction. Expert Systems with Applications, 2011, 38(8): 10389—10397.
- [8] 王波, 程福云. kNN算法在股票预测中的应用. 科技创业月刊, 2015, 28(16): 25—26. (Wang B, Cheng F Y. Application of KNN algorithm in stock forecast. Pioneering with Science & Technology Monthly, 2015, 28(16): 25—26.)
- [9] 黄秋萍, 周霞, 甘宇健等. SVM与神经网络模型在股票预测中的应用研究. 微型机与应用, 2015, 34(5): 88—90. (Huang Q P, Zhou X, Gan Y J, et al. Application of SVM and neural network model in the

- stock prediction research. Microcomputer & Its Applications, 2015, 34(5): 88—90.)
- [10] 任治国. 基于 GM(1,1) 动态等维残差修正: BP 神经网络组合预测模型. 硕士学位论文. 长沙: 湖南大学, 2015. (Ren Z G. The combination forecasting model based on the GM (1, 1) dynamic equal dimensional residual correction; BP neural network. Master Dissertation. Changsha: Hunan University, 2015.)
- [11] 孙秋韵. 基于改进 GA 优化参数的 SVR 股价预测研究. 硕士学位论文. 福州: 福建师范大学, 2015. (Sun Q Y. Research of stock price prediction based on SVR with parameters optimized by improved GA. Master Dissertation. Fuzhou: Fujian Normal University, 2015.)
- [12] 杜晓. 基于 MEA-BP 模型的股票预测研究. 硕士学位论文. 兰州: 兰州大学, 2015. (Du X. Stock prediction research based on MEA-BP model. Master Dissertation. Lanzhou: Lanzhou University, 2015.)
- [13] 宋刚, 张云峰, 包芳勋等. 基于粒子群优化 LSTM 的股票预测模型. 北京航空航天大学学报, 2019, 45(12): 2533—2542. (Song G, Zhang Y F, Bao F X, et al. Stock prediction model based on particle swarm optimization LSTM. Journal of Beijing University of Aeronautics and Astronautics, 2019, 45(12): 2533—2542.)
- [14] 宁昱博, 张玉军. LSTM-Adaboost 股价预测模型. 辽宁科技大学学报, 2019, 42(5): 383—388. (Ning Y B, Zhang Y J. LSTM-Adaboost's stock price forecasting model. Journal of University of Science and Technology Liaoning, 2019, 42(5): 383—388.)
- [15] 黄丽明, 陈维政, 闫宏飞等. 基于循环神经网络和深度学习的股票预测方法. 广西师范大学学报(自然科学版), 2019, 37(1): 13—22. (Huang L M, Chen W Z, Yan H F, et al. A stock prediction method based on recurrent neural network and deep learning. Journal of Guangxi Normal University (Natural Science Edition), 2019, 37(1): 13—22.)

(责任编辑 杨可盛)