文章编号:1674-2974(2022)10-0111-08 DOI:10.16339/j.cnki.hdxbzkb.2022356

基于情感分析和GAN的股票价格预测方法

刘玉玲[1†],赵国龙[1],邹自然[2],吴升婷[1]
(1. 湖南大学 信息科学与工程学院,湖南 长沙 410082;2. 湖南大学 工商管理学院,湖南 长沙 410082)

摘 要:股票价格具有非平稳性和波动性特点,且投资者容易受自身情感影响,投资决策行为具有非理性特征,因此股票价格难以预测.针对预测股票价格的卷积神经网络情感分析方法存在文本标记分布不平衡问题,本文提出一种基于情感分析和生成对抗网络的股票价格预测方法.首先,建立金融领域情感词典库;然后,使用基于词典的情感分析方法计算金融文本数据的情感极性和投资者每天的总体情感指数;最后,利用生成对抗网络对股市波动进行预测,其中生成器生成股票序列数据,而判别器采用卷积神经网络对生成数据和真实数据进行区分.该方法能动态地更新股票价格预测结果且误差较小.

关键词:股票价格预测;情感分析;卷积神经网络;生成对抗网络
中图分类号:TP399 **文献标志码**:A

Stock Price Prediction Method Based on Sentiment Analysis and Generative Adversarial Network

LIU Yuling[1†],ZHAO Guolong[1],ZOU Ziran[2],WU Shengting[1]
(1. College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China;
2. Business School, Hunan University, Changsha 410082, China)

Abstract: The stock price is nonstationary and volatile, the investors are easily influenced by their own sentiments, and their investment decision is irrational. Thus, the stock price is difficult to predict. Aiming at the problem of an unbalanced distribution of text labels in the sentiment analysis method based on the CNN neural network, this paper proposes a stock price prediction method based on sentiment analysis and a generative adversarial network. First, a sentiment dictionary database is established in the financial field. Then, the dictionary-based sentiment analysis method is used to calculate the sentiment polarity of financial text data and the overall sentiment trend of investors every day, that is, the sentiment index. Finally, the generative adversarial network is used to predict the stock market volatility, where the generator generates stock sequence data, and the discriminator uses a convolutional neural network to distinguish the generated data from the real data. This method can dynamically update the prediction results of stocks and obtain smaller error values.

Key words: stock price prediction; sentiment analysis; convolutional neural networks; generative adversarial network

---

* **收稿日期**:2021-08-27
基金项目:国家自然科学基金资助项目(61872134),National Natural Science Foundation of China(61872134);教育部科技发展中心项目(2019J01020),Ministry of Education Science and Technology Development Center(2019J01020);长沙市科技计划项目(Kh2005019),Science and Technology Plan of Changsha(Kh2005019)
作者简介:刘玉玲(1980—),女,湖南宁乡人,湖南大学副教授,博士
† 通信联系人,E-mail: yuling_liu@hnu.edu.cn

随着深度学习技术的不断发展,许多学者提出了新的网络模型算法来完成序列数据生成任务,其中以Goodfellow提出的生成对抗网络(GAN)最为引人注目[1],该算法在图像生成方面取得显著成果[2].金融领域相关学者也开始尝试利用GAN生成股票数据,并取得了初步成果.但是这些方法忽略了投资者情感对交易市场波动的影响.为了解决上述问题,本文采用基于词典的情感分析方法量化投资者情感倾向;然后构建用于股票价格预测的生成对抗网络框架,将量化的情感指数作为数据特征的一部分来训练网络;最后使用GAN的股票价格预测框架对上证A股指数进行实证检验,验证该方法的有效性.

1 相关工作

本节首先介绍基于词典的情感分析方法,然后简要介绍面向股票预测的GAN模型.

1.1 基于词典的情感分析方法

基于词典的情感分析方法主要利用情感词典库来统计待分析文本中正、负情感词的总权值,从而判断句子情感极性,相比于基于机器学习的情感分析方法,该方法有着很好的可解释性.目前国内外开源情感词典主要包括:知网(Hownet)[3]、大连理工大学中文情感词汇本体库(DLUTSD)[4]、台湾大学中文情感极性词典(NTUSD)[5]和普林斯顿大学的WordNet[6].通用领域的情感词汇只能表达一般心理情感,不包括"看涨/看跌"情感词,因此金融领域通常会通过加入领域词典来进行情感分析.例如:Li等人通过融合Harvard心理情感词典和Loughran - McDonald金融情感词典,探索财经新闻对股票价格上涨的影响[7].随后,UhrP等人将词汇量与情感词典资源相结合,增加了人类对常用词汇量的分析方法[8].Tan等人随后针对财经新闻文章,提出了基于规则的情感分析方法.具体为:首先用先验极性词典把新闻划分为不同类别(正面和负面),然后用句子极性比率(正面/负面)来计算整篇文章情感值,在二分类情况下,能达到75.6%的F值[9].Day和Lee根据4种数据集(NowNews、Apple-Daily、LTN和MoneyDJ)构建情感词库,并采用Suffix Array算法为情感词库扩充新词,针对不同新闻提供商提供的文本信息进行情感分析.实验结果表明,使用金融领域词典库的情感分析方法,能使投资者获得较高投资回报率[10].Zhang等人在基本情感词典库、程度副词词库、表情符号词库等6种词库的基础上构建了微博情感词库,并基于情感词库对微博文本进行情感分析[11].文献[12]和[13]建立心理情感词典库,使用基于词典的方法分析Twitter信息对股票市场的影响.在此基础上,Jiawei等人提出了一种基于词典的情感分析方法,根据情感词和句子词向量计算相似度,进行情感极性分析,并通过提取股票特征进一步预测股市走势[14].此外,傅魁等人还结合知网和金融领域的两大情感词典,建立了一套特定金融领域词典库[15].

1.2 面向股票预测的GAN模型

生成对抗网络(GAN)是由Goodfellow在2014年提出的机器学习架构,模型训练中生成器和判别器相互博弈.生成器生成假样本,而判别器区分真实样本和假样本.在达到一个理想点时判别器无法区分两种样本,此时生成器就能更好地捕获真实样本数据分布.目前,生成对抗网络被广泛应用于金融领域预测股票价格、优化投资组合和交易执行策略等.Zhang等人提出了一种以多层感知器(MLP)作为判别器,以长短期记忆模型(LSTM)作为生成器来预测股票收盘价的GAN体系结构,并尝试预测每日收盘价[16].Zhou等人提出了一种股票预测模型GAN-FD,采用13个技术指标作为输入数据,以LSTM作为生成器,而卷积神经网络(CNN)作为判别器,用于对抗训练以预测股票价格方向[17].王静等人结合了经验模态股票数据分解方法,解决了传统GAN预测模型不稳定的问题,并在现有股票数据集中取得较好的预测效果[18].Faraz等人在GAN模型的基础上,使用最小二乘损失函数,并采用小波变换预处理数据,减少股市噪声对模型性能的影响,该模型(LSGAN)在S&P500指数上性能优于GAN[19].

现有面向股票预测的GAN模型都借鉴了其捕获股票交易数据隐藏深层次特征的能力,但忽略了投资者情感因素,因此,本文提出一种基于情感分析和GAN的股票预测方法,并结合基于词典的情感分析方法,计算投资者情感指数,将其作为GAN预测股票收盘价的一个辅助特征.

2 方法描述

如图1所示,整个股票价格预测的框架包括两个部分:1)基于词典的金融文本情感分析,2)基于情感分析和GAN的股票价格预测.下面详细介绍这两个部分.

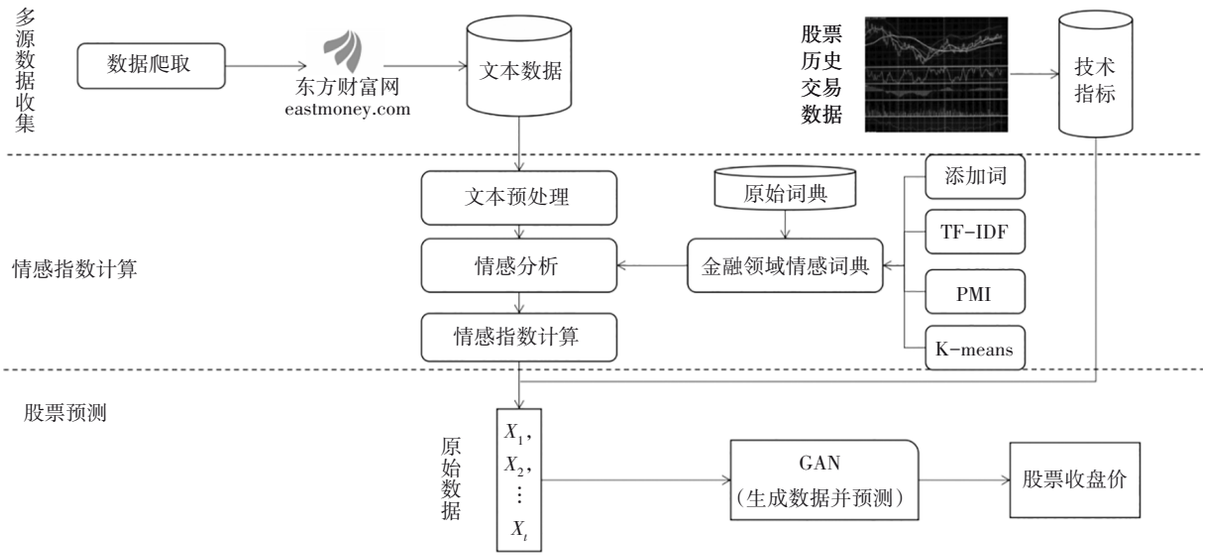


图 1　基于情感分析和GAN的股票价格预测框架

Fig.1　Stock price prediction framework based on sentiment analysis and GAN

2.1　基于词典的情感分析算法

股票投资者往往借助所获取的股市相关新闻、网络平台言论等信息来辅助其决策.股市相关新闻的情感倾向分析对后续预测方法研究有指导意义.

CNN情感分析方法容易造成标签分布不均.而基于词典的情感分析方法,根据各种新闻信息、政策文件和情感词典的匹配结果来确定金融文本数据情感极性,能避免标签分布不均等问题.基于词典的情感分析方法首先通过Jieba分词算法对文本数据进行分词,并去除停用词.积极的情感词初始权值设定为1,而消极的情感词初始权值设定为-1.若情感词中存在否定词,将其权值设定为-1.若有程度副词,则其权值为情感词权值乘程度副词权值.最后,计算句子中积极词总权值和消极词总权值之和来判断句子情感极性.该算法伪代码描述如表1所示.

表 1　基于词典的情感分析方法伪代码

Tab.1　Pseudo code of sentiment analysis method based on dictionary

算法1　基于词典的情感分析方法
输入:句子,金融领域情感词典
输出:每一个句子的情感极性
步骤:
Step1:利用Jieba分词对输入的句子进行切分,同时去掉停用词,在通用停用词库中手动删除含有“不”字的词语,得到切分的词语;
Step2: 加载领域情感词典;
Step3: 根据切分之后的词语与情感词典匹配,找到句子中的情感词,并判断情感词的前面是否有否定词和程度副词,计算总的权值;
Step4: 根据总的权值判断句子的情感极性,总的权值大于0,句子的情感极性为积极的,反之,小于0则为消极的.

2.1.1　面向金融领域情感词典的构建

情感词典的构建是否恰当将直接影响情感分析结果.本文使用的词汇规模如表2所示.包括通用领域的3个情感词典,分别是HowNet(知网),DLUTSD(大连理工大学中文情感词库), NTUSD(台湾大学中文情感极性词典).此外还加入了程度副词和否定词词典库.由于这些通用领域词典都与金融领域无关,根据这些词典将文本进行分类时,效果并不理想.因此,需要将通用词典与金融领域情感词典整合.具体步骤为:合并通用词库中积极词集和消极词集,删除重复词;然后构建金融领域情感词典.详细流程如下:使用TF-IDF(词频-反向文档频率)算法选取1 000个与金融领域相关的词,并判断极性;从中手动选择50个积极情感词汇、50个消极情感词汇,并将其作为种子词,如表3所示;然后使用PMI计算出每个词与种子词的相关性,作为扩展词;手动添加金融领域的积极情感词汇和消极情感词汇;最后基于K-means特征选择方法,选择表3中积极情感词汇作为初始质心,将所有词语使用Word2vec方法映射成单词向量,映射之后的单词向量包括语法和语义信息,通过计算语料库中的词语单词向量与质心单词向量之间的距离来计算两个词语的相似性,不断迭代,直到语料库中所有词语都被归为正确类别,选择10个类别中与每一个类别相似度排在前100名的词汇加入情感词典中,消极情感词汇的操作类似.使用式(1)计算语料库中情感词与质心种子词的距离,其中m和n是语料库词向量矩阵和质心词向量矩阵的行号,j为词向量维度,然后基于式(2)计算

表2 3个通用领域词汇规模

Tab.2 Size of vocabulary in three general fields

词汇数	知网（HowNet）	大连理工大学（NLUTND）	台湾大学（NTUSD）	否定词典	程度副词
总数	6 846	22 012	11 086	36	219
积极情感词汇数	3 730	11 229	2 810	—	—
消极情感词汇数	3 116	10 783	8 276	—	—

表3 金融领域情感种子词

Tab.3 Emotional seed words in financial field

词性分类	示例
积极情感词汇	扭转 看好 反弹 安稳 爆发 暴涨 飙升 激增 ……
消极情感词汇	暗淡 负债 贬值 崩盘 暴跌 欠佳 赤字 惨淡 ……

这些距离的平均值.基于K-means的情感词汇的选择算法在表4中详细描述.面向金融领域的情感词典构建流程和算法分别如图2和表5所示.

$$d_{mn} = \sum_{i=1}^{j} |X_{mi} - X_{ni}| \tag{1}$$

$$\text{avg} = \frac{\sum_{k-1}^{m} d_{kn}}{m} \tag{2}$$

表4 基于K-means的情感词汇的选择

Tab.4 Emotional vocabulary selection based on K-means

算法2 基于K-means的情感词汇的选择
输入：金融领域语料库,初始种子词作为质心
输出:金融领域极性情感词汇
步骤：
Step1:使用Word2vec模型,金融语料库和初始种子词都从文本映射成单词向量；
Step2:对于金融语料库中的每一行,使用式(1)计算语料库中单词向量与初始质心单词向量的距离；
Step3:使用式(2),计算平均距离作为衡量标准；
Step4：选择距离每一个质心小于平均距离的100个词汇作为极性情感词汇.

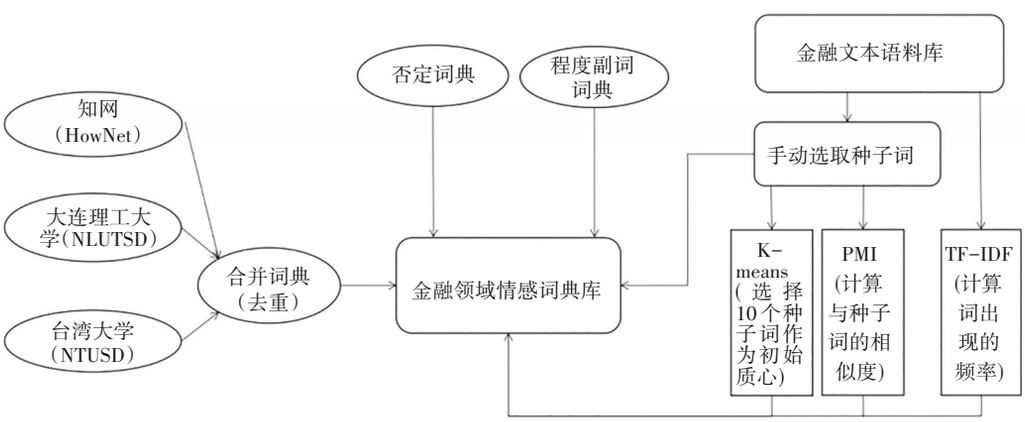


图2 面向金融领域的情感词典构建流程

Fig.2 Construction flow of emotional dictionary for financial field

表5 面向金融领域的情感词典构建算法

Tab.5 Construction algorithm of emotional dictionary for financial field

算法3 面向金融领域的情感词典构建方法
输入:通用领域情感词典,金融文本数据,程度副词词典库,否定词词典库
输出:金融领域情感词典库
步骤：
Step1:合并通用领域的情感词典,并去除重复的词；
Step2:利用Jieba分词对金融文本数据进行切分,去除停用词,在通用停用词库中手动删除含有“不”字的词语,获得金融领域语料库；
Step3：根据TF-IDF方法统计语料库中词语出现的频率,选取1 000个,手动判断情感极性,根据情感极性加入金融领域情感词典库中；
Step4：手动选择情感领域的极性种子词,包括50个积极情感词汇,50个消极情感词汇,并加入情感词典库中；
Step5：使用PMI方法计算出语料库中每一个词语与种子词的情感相关度,并排序,选取与积极种子词相关度排名前1 000的词汇,与消极种子词相关度排名前1 000的词汇,并加入金融领域情感词典中,获得最后的金融领域情感词典.
Step6：基于K-means的词典构建,从积极种子词中随机选取10个作为质心,将所有的积极情感词汇分成10类,消极情感词汇也是一样的操作,分成10类,加入情感词典中.

2.1.2　情感指数的计算

基于词典的情感分析方法只是判断每一篇新闻或者政策文件的情感极性,而股票历史数据是随时间变化的序列数据,所以也需要计算每一天情绪的波动趋势.情感指数就是计算事件和公众情绪的整体走向趋势,计算方式如式(3)所示.

$$\text{sentiment index} = \ln\frac{1+\text{num}_{\text{tpos}}}{1+\text{num}_{\text{tneg}}} \tag{3}$$

式中:num_{tpos}是第t天的积极新闻、论坛帖子和政策文件总数,而num_{tneg}是第t天的消极新闻、论坛帖子和政策文件总数.情感指数范围在-0.5~0.5之间,且情感指数低于0表示在第t天情感极性消极.

2.2　基于情感分析和GAN的股票预测

本文主要利用GAN生成股票波动曲线.不同时段的股票模式和行为或多或少会有相同之处,因此,生成数据与已有数据有相似分布.由于循环门控单元(GRU)神经网络对于时间序列数据拟合度较高,收敛速度快,本文采用GRU作为生成器,依据真实数据$X_1, X_2, \cdots, X_t$,生成与真实数据相似的数据$X_{t+1}^{\sim}$,而判别器采用CNN,对输入的序列:真实数据$X_1, X_2, \cdots, X_t, X_{t+1}$和生成数据$X_1, X_2, \cdots, X_t, X_{t+1}^{\sim}$进行分类.基于GAN的股票价格预测模型如图3所示.

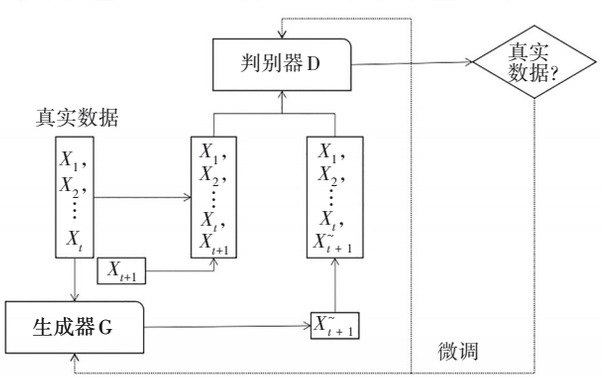


图3　基于GAN的股票价格预测模型

Fig.3　Stock price prediction model based on GAN

2.2.1　生成器

本文选择GRU神经网络作为生成器.使用9个变量来预测股票未来收盘价,这几个变量分别为开盘价,最高价,最低价,收盘价,成交量,$K\%$(随机指标),$R\%$(威廉指标),RSI(相对强弱指标),情感指数,其中$K\%$、$R\%$和RSI三个技术指标,计算方式分别如式(4)、式(5)、式(6)和式(7)所示.生成器的输入为$X=(X_1^{\mathrm{T}}, X_2^{\mathrm{T}}, \cdots, X_t^{\mathrm{T}})$,其中$X_i^{\mathrm{T}}$, $i\in(1,2,\cdots,t)$表示的是一个向量,由9个特征组成,如下所示.

$$X_i=(X_{i\mathrm{H}}, X_{i\mathrm{L}}, X_{i\mathrm{O}}, X_{i\mathrm{C}}, X_{i\mathrm{V}}, X_{iK\%}, X_{iR\%}, X_{i\mathrm{RSI}}, X_{i\mathrm{senti}})$$

$$K\% = 100\times\frac{\mathrm{C}-L_t}{H_t-L_t} \tag{4}$$

$$R\% = 100\times\frac{H_t-C}{H_t-L_t} \tag{5}$$

$$\text{RSI} = 100-\frac{100}{100+\text{RS}} \tag{6}$$

$$\text{RS} = \text{Avg}\left[\frac{t\ \text{day's up clo sin g price}}{t\ \text{day's down clo sin g price}}\right] \tag{7}$$

式中:以t天为单位,t值在本文实验中设置为5,C为收盘价,H_t和L_t分别表示最近t天的最高价和最低价.RS是t天内收盘价上涨总和的平均值与收盘价下跌平均值的比值.

生成器有两个隐藏层输出h_{1t}和h_{2t},分别表示2层的GRU隐藏层的输出,最后一个隐藏层的输出会输入到具有9个神经元的完全连接层中,生成第t+1天的数据$X_{t+1}^{\sim}$,即第t+1天的预测收盘价.生成器输出计算公式如(8)所示.∂表示全连接的激活函数"LeakyReLu",W_{f}和b_{f}分别表示全连接层的权重和偏置.

$$G_{(x)}=X_{t+1}^{\sim}=\partial(W_{\mathrm{f}}\times h_{2t})+b_{\mathrm{f}} \tag{8}$$

2.2.2　判别器

判别器的目的是对输入数据进行分类.本文选择CNN神经网络作为分类器,主要由输入层、3个卷积层、池化层、全连接层和输出层组成,卷积层激活函数都为"LeakyReLu",而输出层是一个全连接层,激活函数为"Sigmoid",所以输出为0和1,当输入为生成数据时,输出为0,而当输入为真实数据时,输出为1.同时采用交叉熵损失优化判别器的损失函数.

2.2.3　GAN模型的训练

$F(\mathrm{G},\mathrm{D})$表示GAN模型的优化过程,G表示GAN网络中的生成器,D表示判别器,损失函数如式(9)所示:

$$\begin{aligned}\min_{\mathrm{G}}\max_{\mathrm{D}}F(\mathrm{G},\mathrm{D}) = {} & E[\log(\mathrm{D}(X_{\mathrm{R}}))]+\\ & E[\log(1-\mathrm{D}(X_{\mathrm{F}}))]\end{aligned} \tag{9}$$

式中:X_{R}为真实数据,X_{F}为生成数据,生成器在使用其损失函数实现优化的同时,还结合了生成的第t+1天数据与真实的t+1天数的均方误差(MSE).

3　实验及结果分析

本文采用的生成器由2个GRU隐藏层单元组成,神经元的个数分别为512和256,隐藏层的激活函数采用"LeakyReLu",学习率为0.000 1.判别器为CNN模型,其中卷积层和全连接层的激活函数为"LeakyReLu",而输出层的激活函数为"Sigmoid",学习率为0.000 6.在GAN训练过程中,生成器采用随

机梯度优化算法,判别器采用Adam算法.本文使用上证A股指数作为实验数据集,对比不同股票价格预测方法.实验结果表明,本文提出的方法具有较好的预测效果.

3.1 数据描述

本文包含两个数据集:1)2015年3月至2021年3月期间来自东方财富网的上证指数吧的互动帖,以及从白鹿数据网站获得的政策文件.2)同时期的每日上证A股综合指数股票价格从Yahoo Finance下载获得.表6是2021年3月31日这个交易日内的股票和文本数据示例.

表6 交易日内的股票和文本数据示例

Tab.6 Example of stock and text data during trading day

日期	开盘价/(元·手$^{-1}$)	最高价/(元·手$^{-1}$)	最低价/(元·手$^{-1}$)	前收盘价/(元·手$^{-1}$)	收盘价/(元·手$^{-1}$)	成交量/手
2021-03-31	2 570.374	2 612.653	2 560.099	2 586.224	2 612.18	252 284 835
新闻(标题)	3月31日复盘:大盘迎下跌调整,主力重点出击7股					
论坛帖子(标题)	现在看走的趋势确实不理想					
政策文件(标题)	中国人民银行、银保监会、证监会、外汇局发布《关于金融支持海南全面深化改革开放的意见》					

3.2 性能评价指标

本文预测股票的收盘价采用的评估指标有平均绝对误差(MAE),均方误差(MSE)和均方根误差(RMSE),分别由式(10)、式(11)和式(12)表示,其中f_i是第i天的预测收盘价,y_i是第i天的真实收盘价.

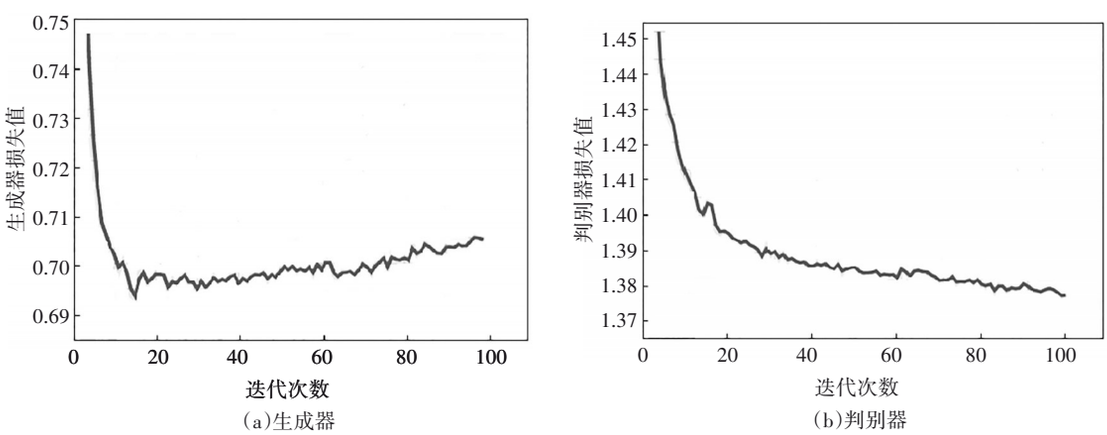
$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}|f_i-y_i| \tag{10}$$

$$\mathrm{MSE}=\frac{1}{n}\sum_{i=1}^{n}(f_i-y_i)^2 \tag{11}$$

$$\mathrm{RMS}=\sqrt{\mathrm{MSE}} \tag{12}$$

3.3 实验结果及性能比较

在训练期间,GAN网络两个模型的损失变化如图4所示.最开始生成器损失由自身损失值结合生成数据与真实数据误差值组成,而随着迭代次数不断增加,GAN网络捕获到的特征数也会增加,损失就会不断减小.生成器和判别器在对抗过程中不断实现优化直到收敛.



(a)生成器 (b)判别器

图4 生成器和判别器的损失变化

Fig.4 Loss variation of generator and discriminator

3.3.1 不同模型预测结果对比

为了比较不同模型的预测结果,本文选择在标准化的上证A股综合指数数据集上计算MAE、MSE和RMSE的平均值,并作为评估指标.同时选择常用于股票预测的经典模型ANN(人工神经网络)[20]、LSTM和WGAN-GP[21]作为基线模型,与本文提出的基于GAN的预测模型进行比较.预测结果如表7所示,字体中加粗部分表示最佳预测结果.低MAE、MSE和RMSE表示预测的收盘价接近真实数据.从表7中可以看到本文所提出的模型与其他模型相比具有竞争优势.从图5中也可以看出,基于GAN模型的预测收盘价更加接近真实收盘价的波动趋势.

表7　不同模型的预测结果对比

Tab.7　Comparison of prediction results of different models

模型	MAE	MSE	RMSE
ANN	0.183 5	0.060 0	0.227 4
LSTM	0.420 7	0.201 6	0.434 3
WGAN-GP	0.108 5	0.026 4	0.162 5
GAN	**0.064 2**	**0.007 0**	**0.081 3**

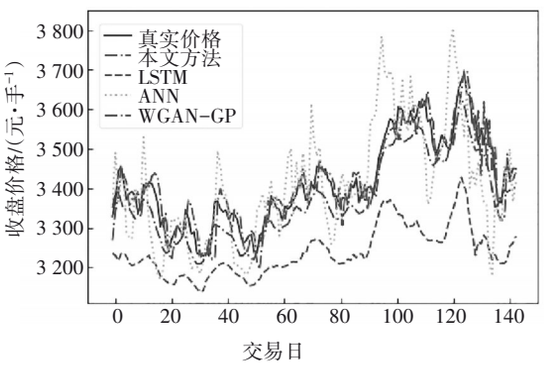


图5　多种模型预测结果对比

Fig.5　Comparison of prediction results of various models

3.3.2　不同情感词典方法的比较

针对情感词典的不同,股市收盘价预测结果如表8所示.从表中可以看出,加入金融领域自己构建的情感词典库之后,模型的预测性能得到较好的提升,实验结果证明了采用金融领域情感词典库能够提高股票价格的预测性能.

表8　采用不同情感词典预测股票价格的结果对比

Tab.8　Comparison of stock price prediction results using different emotional dictionaries

方法	MAE	MSE	RMSE
通用情感词典	0.074 9	0.008 8	0.091 9
通用情感词典+领域情感词典	**0.064 2**	**0.007 0**	**0.081 3**

3.3.3　加入多源情感特征对股票价格的影响

实验也验证了投资者情感因素对于股票价格有影响,选取的金融文本数据包括公司财经新闻、论坛帖子和国家发布的政策文件,探究这些金融文本对投资者情感的影响,作为多源情感特征.我们将股票历史数据作为基线进行结果对比.当加入了3个技术指标之后,评价指标都有一定程度的降低.这说明3个技术指标反映了平均波动方向,可以作为股票预测的一个辅助特征.当在分析股票历史数据基础上加入多源情感特征之后,使用GAN模型来预测股票收盘价,预测误差结果有较大的减少,投资者情感会影响股票波动,多源情感特征可以作为股票预测的一个辅助特征,进一步提升预测效果.而当结合技术特征和多源情感特征时,预测结果又稍微比只加入一个特征的效果有所提升,这也证实了股票波动受到多因素影响,需要从多个方面考虑,从而获得更好的效果.表9为选取不同特征预测股票价格的结果对比.

表9　选取不同特征预测股票价格的结果对比

Tab.9　Comparison of results of predicting stock price by selecting different feature

方法	MAE	MSE	RMSE
只有股票历史数据	0.081 1	0.010 7	0.098 4
股票历史数据+技术指标	0.071 5	0.007 6	0.086 1
股票历史数据+多源情感特征	0.067 6	0.007 9	0.086 0
股票历史数据+技术指标+多源情感特征	**0.064 2**	**0.007 0**	**0.081 3**

3.3.4　所提方法的稳定性

本文绘制了股票收盘价对比图,如图6所示.最开始的预测中,预测的价格与真实价格很接近,说明在模型的训练期间,预测模型在生成股票数据时,充分提取了金融文本数据的股票交易数据的特征,并且能够很好地学习参数.部分预测的收盘价有较大的出入,整体上,股票大致上涨和下跌趋势还是比较明显,充分证明了GAN用于预测股票收盘价的稳定性.

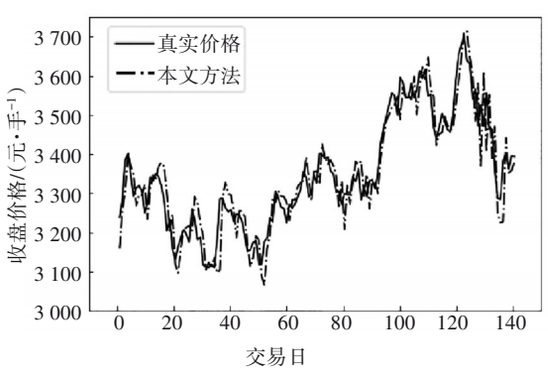


图6　上证A股综合指数收盘价预测结果

Fig.6　Prediction results of closing price of SSE A-share composite index

4　结　论

本文提出了一种基于情感分析和GAN的股票价格预测方法.通过构建金融领域的情感词典库,采用基于词典的方法对金融文本数据进行情感分析.

同时,本文将GAN应用到股票数据生成中,在上证A股指数的数据集上进行实验验证.本文结合投资者情感指数的GAN预测模型具有更好的稳定性和较小的误差值.未来的研究工作可在如下2个方面展开:

1)在多源数据中添加金融报表数据和股票波动折线图,通过处理跨模态数据,更好地捕获特征,提高最终预测效果;

2)本文提出的方法是根据前几天数据来预测后一天数据,整体时间周期较长,粒度也较大,可以采取某天中开市期间数据,研究当天收盘价,进一步比较多源数据对于股票市场波动的持久度影响.

参考文献

[1] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, *et al*. Generative adversarial nets[C]//Advances in Neural Information Processing Systems. New York, NY, USA: ACM, 2014: 2672-2680.

[2] 王军,夏利民. 基于深度学习特征的异常行为检测[J]. 湖南大学学报(自然科学版),2017,44(10):130-138.
WANG J, XIA L M. Abnormal behavior detection based on deep-learned features[J]. Journal of Hunan University (Natural Sciences), 2017, 44(10): 130-138. (In Chinese)

[3] DONG Z D, DONG Q. How Net-a hybrid language and knowledge resource[C]//International Conference on Natural Language Processing and Knowledge Engineering. Beijing, China: IEEE, 2003: 820-824.

[4] 陈建美,林鸿飞,杨志豪. 基于语法的情感词汇自动获取[J]. 智能系统学报,2009,4(2):100-106.
CHEN J M, LIN H F, YANG Z H. Automatic acquisition of emotional vocabulary based on syntax[J]. Caai Transactions on Intelligent Systems, 2009, 4(2): 100-106. (In Chinese)

[5] HUANG W C, LIN M C, WU S H. Opinion sentences extraction and polarity classification using automatically generated templates [C]//Proceedings of NTCIR-8 Workshop Meetings. Tokyo, Japan, 2010: 255-257.

[6] DU W F, TAN S B, CHENG X Q, *et al*. Adapting information bottleneck method for automatic construction of domain-oriented sentiment lexicon[C]//WSDM'10: Proceedings of the Third ACM International Conference on Web Search and Data Mining. New York, NY, USA: ACM, 2010: 111-120.

[7] LI X, XIE H, CHEN L, *et al*. News impact on stock price return via sentiment analysis[J]. Knowledge-Based Systems, 2014, 69: 14-23.

[8] UHR P, ZENKERT J, FATHI M. Sentiment analysis in financial markets A framework to utilize the human ability of word association for analyzing stock market news reports[C]//2014 IEEE International Conference on Systems, Man, and Cybernetics. San Diego, CA, USA: IEEE, 2014: 912-917.

[9] TAN L I, PHANG W S, CHIN K O, *et al*. Rule-based sentiment analysis for financial news[C]//2015 IEEE International Conference on Systems, Man, and Cybernetics. Hong Kong, China: IEEE, 2015: 1601-1606.

[10] DAY M Y, LEE C C. Deep learning for financial sentiment analysis on finance news providers[C]//2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). San Francisco, CA, USA: IEEE, 2016: 1127-1134.

[11] ZHANG S X, WEI Z L, WANG Y, *et al*. Sentiment analysis of Chinese micro-blog text based on extended sentiment dictionary [J]. Future Generation Computer Systems, 2018, 81: 395-403.

[12] ZHANG C Y, JI Z, ZHANG J X, *et al*. Predicting Chinese stock market price trend using machine learning approach[C]//CSAE, 18: Proceedings of the 2nd International Conference on Computer Science and Application Engineering. New York, NY, USA: ACM, 2018: 1-5.

[13] LI M, YANG C, ZHANG J, *et al*. Stock market analysis using social networks[C]//Proceedings of the Australasian Computer Science Week Multiconference. New York, NY, USA: ACM, 2018: 1-10.

[14] JIAWEI X, MURATA T. Stock market trend prediction with sentiment analysis based on LSTM neural network [C]//International Multiconference of Engineers and Computer Scientists. Hong Kong, 2019: 475-479.

[15] 傅魁,刘玉洁,陈美丽. 基于财经新闻情感倾向值的股票价格预测[J]. 北京邮电大学学报(社会科学版),2019,21(1):87-100.
FU K, LIU Y J, CHEN M L. Stock price forecasting based on emotion tendency of financial news[J]. Journal of Beijing University of Posts and Telecommunications (Social Sciences Edition), 2019, 21(1): 87-100. (In Chinese)

[16] ZHANG K, ZHONG G Q, DONG J Y, *et al*. Stock market prediction based on generative adversarial network[J]. Procedia Computer Science, 2019, 147: 400-406.

[17] ZHOU X Y, PAN Z S, HU G Y, *et al*. Stock market prediction on high-frequency data using generative adversarial nets[J]. Mathematical Problems in Engineering, 2018, 2018: 4907423.

[18] 王静,邹慧敏,曲东东,等. 基于经验模态分解生成对抗网络的金融时间序列预测[J]. 计算机应用与软件,2020,37(5):293-297.
WANG J, ZOU H M, QU D D, *et al*. Financial time series prediction based on empirical mode decomposition to generate adversarial networks[J]. Computer Applications and Software, 2020, 37(5): 293-297. (In Chinese)

[19] FARAZ M, KHALOOZADEH H. Multi-step-ahead stock market prediction based on least squares generative adversarial network [C]//2020 28th Iranian Conference on Electrical Engineering (ICEE). Tabriz, Iran: IEEE, 2020: 1-6.

[20] VIJH M, CHANDOLA D, TIKKIWAL V A, *et al*. Stock closing price prediction using machine learning techniques[J]. Procedia Computer Science, 2020, 167: 599-606.

[21] LIN H, CHEN C, HUANG G F, *et al*. Stock price prediction using Generative Adversarial Networks[J]. Journal of Computer Science, 2021, 17(3): 188-196.