第50卷第8期 中国科学技术大学学报 Vol. 50, No. 8
2020年8月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Aug. 2020
文章编号:0253-2778(2020)08-1093-09

基于双编码器利用在线社交网络信息的股票价格预测

崔文泉,王青芳

(中国科学技术大学管理学院统计与金融系,安徽合肥 230026)

摘要:设计了双编码器-解码器模型,在模型的双编码器端分别对情绪变量和技术指标进行单独编码,以提高两类信息输入时编码器-解码器模型对股价的预测准确率.首先,对模型的编码和解码,基于门控循环单元(GRU)进行改进,通过去掉重置门,使用更新门代替重置门的功能,将激活函数tanh替换为ReLU激活函数,以达到提高网络训练速度和模型精度的效果.其次,将市场情绪看作离散时间的随机过程,当固定时间时,市场情绪是服从某个概率分布的变量,对其概率分布进行估计,可得市场情绪关于积极、消极和中立的概率估计.进一步的,基于构建伪标签的情感分类器,建立情绪得分公式,并基于Bagging集成的方法对市场情绪的概率分布进行估计,作为投资者情绪变量的补充.另一方面,对多个超参数调整选优,设计正交试验,大大缩短了模型选参时间.实验结果表明,两输入的双编码器-解码器,不仅提升了编码器-解码器框架的股价预测效果,还通过引入投资者情绪,提高了模型的准确率和鲁棒性.

关键词:在线社交网络;投资者情绪;双编码器-解码器;门控循环单元
中图分类号:C812 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2020.08.008

引用格式:崔文泉,王青芳.基于双编码器利用在线社交网络信息的股票价格预测[J].中国科学技术大学学报,2020,50(8):1093-1101.
CUI Wenquan, WANG Qingfang. A dual encoder-based approach to predicting stock price by leveraging online social network[J]. Journal of University of Science and Technology of China, 2020, 50(8):1093-1101.

A dual encoder-based approach to predicting stock price by leveraging online social network

CUI Wenquan, WANG Qingfang

(*Department of Statistics and Finance, School of Management, University of Science and of Technology of China, Hefei* 230026, *China*)

Abstract: We propose a dual-encoder which encodes the investor sentiment and technical indicators separately to improve the accuracy of the encoder-decoder model in predicting stock price by using two types of information. For the dual-encoder and decoder, we revise the gated recurrent unit (GRU) by removing the reset gate, using the update gate instead of the reset gate function and replacing tanh activation function with ReLU activation function to improve the speed of network training and the accuracy of the model. We regard market sentiment as a discrete-time stochastic process. When fixed time, market sentiment is a variable subject to a certain probability distribution. Sentiment score formulas are built for investor sentiment by a pseudo-label based sentiment classifier, and the market sentiment is estimated through ensemble Bagging learning. The orthogonal table experiment design is used to select parameters in our dual-encoder based model, which greatly reduces the time of parameter adjustment. Finally, experiments are conducted to show that our dual-encoder based model is more accurate than encoder-decoder model, and investor sentiment helps improve the stock forecasting in our model.

Key words: online social network; investor sentiment; dual-encoder; GRU

0 引言

股价预测一直以来都是金融时间序列分析中的热点和难点,深度学习方法在这类问题中得到广泛应用,常用的基于技术层面的策略[1-3]主要是利用股票的历史数据进行价格趋势或者收益率预测.实际上,股票价格受到众多因素的影响,除了宏观经济波动、利率变化等因素[4],市场参与者的情绪[6]也会对股价波动产生影响,因而许多研究者将市场情绪作为股票价格趋势或者收益率预测的重要影响

收稿日期:2020-06-17;**修回日期**:2020-07-02
基金项目:国家自然科学基金(71873128)资助.
作者简介:崔文泉(通讯作者),男,1964年生,博士/副教授.研究方向:数理统计.E-mail:wqcui@ustc.edu.cn

因子. Nofer 和 Hinz[7] 利用社交平台 Twitter 提取信息，建立加权社会情绪指数，从而对股票的收益率进行预测，取得了不错的效果. Peng[8] 先利用词嵌入的方法提取文本特征，再使用深度神经网络基于文本特征与历史价格来预测股票的趋势. Chen 等[9] 则基于情感词典和隐狄利克雷模型[12] (latent Dirichlet allocation，LDA)对市场情绪进行提取，之后再运用混合 RNN-boost 模型对股票趋势作预测. 从文本分析来看，上述文献中模型对市场情绪主要基于词语构建，没有关注句子的上下文语义.

传统的编码器-解码器框架模型广泛应用于机器翻译[5,11]、语音识别[10]等领域以及时间序列数据[3]中，由上述文献可知，编码器-解码器是一个框架类模型，而非具体的模型. 序列数据中，编码器-解码器的输入输出，常使用循环神经网络. Qin 等[3] 基于带注意力机制的循环神经网络，使用编码器-解码器模型，基于股票的高频数据对价格有较好的预测结果，但没有考虑市场情绪对股价预测的影响. 因此，本文尝试引入市场情绪，充分挖掘市场投资者的情绪特征，以便使股价预测能达到更优的预测效果.

本文在上述文献的基础上，考虑投资者情绪和技术指标，构建两输入的双编码器-解码器模型，模型的输入和输出建立适度优化后的循环神经网络 O-GRU(optimized-gated recurrent unit)分别进行编码和解码，在双编码器端，对投资者情绪和技术指标这两类信息生成隐藏状态，引入时间注意力机制，在时间步长上对这两类信息的隐藏状态选择有利于目标预测的状态，通过合并生成中间状态向量序列输入到解码器中，达到提升股价预测准确率的效果. 实验表明，本文的方法预测准确率比前人的策略[3]更高，同时，通过引入投资者情绪，基于帖子文本构建情感分类学习器，建立情绪得分公式，将情绪看作离散时间的随机过程，对情绪的概率进行估计，作为投资者情绪变量的补充输入到双编码器-解码器模型中以达到使模型更加稳定的效果.

1 基于双编码器的股票价格预测模型

本文的股票价格预测主要基于股票的两类信息对未来的价格进行预测. 两类信息分别是投资者情绪和技术指标.

1.1 问题描述

基于股票的市场情绪$\{M(t),t=1,2,\cdots,T_{tr}\}$，可看作离散时间的随机过程，其中 T_{tr} 表示 t 能取到的最大交易日. 当固定时间 t 时，将市场情绪看作服从某个概率分布的随机变量，本文利用股吧的帖子文本所反映的情绪 $\hat{M}(t)$来估计股票的市场情绪 $M(t)$. 记第 t 个交易日发表的第 i 条帖子所反映的情绪为 $\hat{M}_i(t)$，$i=1,\cdots,n(t)$，其中 $n(t)$表示第 t 个交易日帖子发表的总量. 假设市场情绪有三种取值，分别为“积极”、“消极”和“中立”，发生的概率为 $\pi(t)=(\pi_1(t),\pi_2(t),\pi_3(t))$，通过估计第 t 个交易日帖子情绪的概率分布，作为后续股价预测模型的输入.

对 $M(t)$的概率进行估计，我们可得代表投资者情绪的变量：

$$V(t)=(M(t),L(t)) \tag{1}$$

式中，$L(t)=(l_1(t),\cdots,l_K(t))$通过 LDA 主题模型[12]计算而得，$l_k(t)$表示第 t 个交易日的文本属于第 k 个主题的概率.

技术指标由股票的历史数据计算而得，记为

$$U(t)=(u_1(t),\cdots,u_m(t)) \tag{2}$$

式中，m 表示技术指标的个数.

由投资者情绪变量 $V(t)$(简记 V_t)和技术指标变量 $U(t)$(简记 U_t)，定义股价预测问题如下：

在时间窗口为 T 的一段交易日内，股票有投资者情绪的序列：

$$V_{t=1}^{T}=\{V_t,\ t=1,\cdots,T\} \tag{3}$$

技术指标变量的序列：

$$U_{t=1}^{T}=\{U_t,\ t=1,\cdots,T\} \tag{4}$$

记时刻 t 的股票价格为 Y_t，响应变量为 Y_{T+r}，

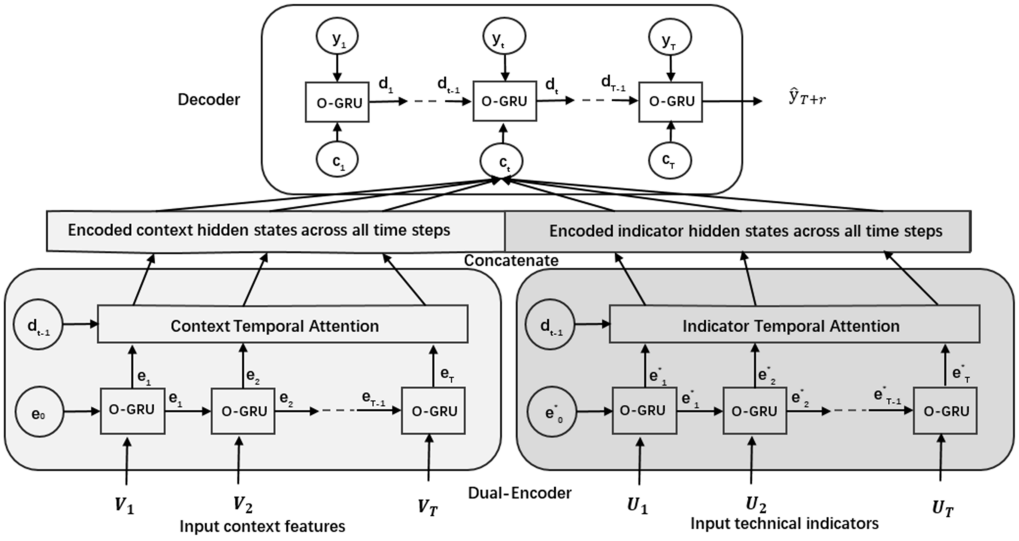


图 1 双编码器-解码器的输入层、隐藏层和输出层

Fig. 1 The input layer, hidden layer and output layer of dual-encoder-decoder

已知解释变量序列$\{V_{t=1}^T, U_{t=1}^T, Y_{t=1}^T\}$，我们有股价预测模型：

$$Y_{T+r}=f(U_1,\cdots,U_T,V_1,\cdots,V_T,Y_1,\cdots,Y_T)+\varepsilon_{T+r} \tag{5}$$

式中，ε_{T+r} 是残差序列，一般地 $r\in\{1,2,3\}$，$f(\cdot)$ 是股价预测函数，本文由双编码器-解码器学习得到.

1.2　模型与方法

基于循环神经网络的传统编码器-解码器[3]模型仅有一个编码器端，隐藏状态对应单一编码器生成，策略的实验仅使用了股票的历史高频数据来预测股价.

本文为了使股票预测适用于多类信息的输入，提出针对股票两类信息的双编码器-解码器，在双编码器端，独立地对投资者情绪和技术指标进行编码. 本文基于双编码器-解码器的股价预测模型如图 1 所示，股票的两类信息，经过各自编码器端的 O-GRU，生成隐藏状态，经过各自的时间注意力加权，合并输入到解码器端的 O-GRU，并将最后一个 O-GRU 单元的输出 $\hat{Y}_{T+r}$ 作为股价预测的结果.

1.2.1　双编码器-解码器模型中的双编码器

双编码器(dual-encoder)由文本特征编码器和技术指标编码器构成. 对文本特征编码器，在时间步 t，记 e_{t-1} 为上个时间步的隐藏状态，由本文的 O-GRU 神经网络，将投资者情绪 V_t 和 e_{t-1} 映射成当前时间步的隐藏状态 e_t，则当前时间步隐藏状态的计算公式：

$$e_t=g_1(e_{t-1},V_t) \tag{6}$$

对技术指标编码器，同文本特征编码器，当前时间步隐藏状态 e_t^* 计算公式为

$$e_t^*=g_2(e_{t-1}^*,U_t) \tag{7}$$

式中，$g_i(\cdot)$，$i=1,2$，分别表示各自编码器中 O-GRU 隐藏层的映射关系.

1.2.2　双编码器-解码器模型中的解码器

为了捕获序列的长期依赖关系，在两类输入的双编码器-解码器框架中，引入带时间注意力机制的 O-GRU，基于注意力权重，自适应地对双编码器生成的隐藏状态进行关注，计算双编码器隐藏状态在所有时间步长中的加权和，作为中间状态向量，输入给解码器端的网络中.

在时间步 t，由上一时间步的解码器隐藏层状态 d_{t-1} 和双编码器的隐藏层状态 $[e_t;e_t^*]$，可以计算双编码器隐藏状态的注意力权重 α_t：

$$a_t^i=H_a^T\tanh(W_a[d_{t-1};e_t;e_t^*]+b_a),\ 1\leqslant i\leqslant T \tag{8}$$

$$\alpha_t^i=\frac{\exp(a_t^i)}{\sum_{j=1}^{T}\exp(a_t^j)} \tag{9}$$

式中，H_a，W_a 是参数，b_a 是偏置，α_t^i 表示双解码器的第 i 个隐藏状态的注意力权重. 由注意力机制，双编码器的每个隐藏状态都映射成中间状态向量 c_t，c_t 在每个时间步都是不同的，将 c_t 计算为双编码器所有隐藏状态的注意力加权：

$$c_t=\sum_{i=1}^{T}\alpha_t^i[e_i;e_i^*] \tag{10}$$

由中间状态向量 c_t，我们可以计算 $\tilde{y}_t$：

$$\tilde{y}_t=\tilde{W}^T[y_t;c_t]+\tilde{b}_y \tag{11}$$

式中，y_t 是解码器的输入. 用函数 g 表示 O-GRU 隐藏层的映射关系，由 $\tilde{y}_t$ 更新解码器在时间步 t 的隐藏状态 d_t：

$$d_t=g(d_{t-1},\tilde{y}_t) \tag{12}$$

最后，对股价预测模型，使用解码器最后一个时间步的输出作为股票价格的预测值：

$$\hat{y}_{T+r}=H_y^T(W_y[d_T;c_T]+b_y)+b_h \tag{13}$$

式中，H_y，W_y 是模型参数，b_y，b_h 是偏置.

1.2.3　基于门控循环单元的神经网络优化

循环神经网络广泛适用于序列数据中，但极易出现梯度消失的问题，为了改善这个问题，研究者们提出了经典的 LSTM (long-short term memory)[18]和 GRU(gated recurrent unit)[19]神经网络，这两种方法由于引入门限机制导致模型参数增多，所以它们在网络训练时都有收敛速度慢的问题，而 Józefowicz 等[20]表示 GRU 在除了语言模型的其他任务中的表现均优于 LSTM，故本文在股票预测模型中选取 GRU 作进一步研究.

本文针对标准 GRU 在一定程度上收敛速度慢、学习效率低的问题，设计基于 GRU 进行适度优化的神经网络，在原结构基础上，移除重置门、保留更新门 z_t，对候选隐藏状态 $\tilde{h}_t$ 进行更新，将 tanh 激活函数换成 ReLU 激活函数，并将原先乘以重置门的地方换成乘以 z_t. 称基于 GRU 改进后的方法为 O-GRU，则 O-GRU 神经网络的公式如下：

$$z_t=\sigma(W_{zh}h_{t-1}+W_{zx}x_t) \tag{14a}$$

$$\tilde{h}_t=\mathrm{ReLU}(W_{sh}(z_t*h_{t-1})+W_{sx}x_t) \tag{14b}$$

$$h_t=(1-z_t)*\tilde{h}_t+z_t*h_{t-1} \tag{14c}$$

式中，$*$ 表示按元素乘积，σ 即 sigmoid 函数，W_{zh}，W_{zx}，W_{sh}，W_{sx} 是相应的模型参数，为简便起见省略偏置.

本文基于设计的双编码器-解码器框架模型使用 O-GRU 对输入进行编码，同时，为了使模型适应两种输入的双编码框架，在解码器框架中使用带时间注意力机制的 O-GRU 对由双编码部分生成的针对两种输入的隐藏状态进行关注和解码，有利于提升模型的预测能力.

1.2.4　情感分类器

对投资者情绪特征的构建，当前自然语言处理中常用的是 BERT(bidirectional encoder representations from transformers)模型[13]，它采用“完形填空”和预测下一个句子的方式，有效实现了对文本的上下文语义表征，因此本文选取 BERT 预训练模型对文本进行基于上下文语义的表征. 而当标签数量所占比例较少时，半监督学习[14] (semi-supervised learning，SSL)提供了一个引入无标签数据信息的框架. Oliver 等[15]表明在理想情况下，即使标签数据非常少，SSL 也能从未标记数据中提取到有用信息.

本文将帖子的情绪划分为三类，分别为积极、消极和中立，通过构建基于帖子文本的情感分类器基学习器来预测帖子文本的情绪类别. 将 BERT 和半监督学习结合进行三分类(积极、消极和中立)的预测. 具体地，随机从无标签的帖子文本数据中抽取少量样本进行人工标注，由 BERT 构建 BERT-RCNN 模型，利用少量有标签数据对模型训练得到有监督的分类器，再对适量无标签的文本数据进行分类预测，选择预测概率最大的类作为文本数据的伪标签，并通过置信阈值过滤最大预测概率较小的伪标签数据，将有标签数据和引入无标签数据得到的伪标签数据混合起来进行半监督学习.

设帖子文本 $S=\{s_1,\cdots,s_n,s_{n+1},\cdots,s_{n+m}\}$，有标签数据为 $S_l=\{s_1,\cdots,s_n\}$，无标签的数据为 $S_u=\{s_{n+1},\cdots,s_{n+m}\}$，且 $n\ll m$. 假设 $s_i(i\in\{1,\cdots,m+n\})$是第 t 个交易日的第 i 条帖子，利用谷歌针对中文进行预训练的 BERT 模型 BERT-Base[13](它包含 12 个 Transformer 层，768 个隐含单元)，我们可以将文本映射成一个基于上下文语意的 768 维的向量，记为 $s_{i,1}$. 把 $s_{i,1}$ 作为两层双向长短期记忆网络(BiLSTM)的输入，输出维度为 $d(d<768)$的向量 $s_{i,2}$，对 $s_{i,\text{embed}}=[s_{i,1};s_{i,2}]$使用池化计算进行适度的信息过滤，减少参数和计算复杂度，防止过拟合. 由于基于金融文本的情感分类对局部的某些术语和词语会比较敏感，故使用最大池化，即在池化窗口中选取最大元素. 最后添加一层全连接层并应用 Softmax 函数，可得到帖子属于某个情感类别的概率：

$$q_i(t)=P\{M_i(t)\mid s_i(t),\Theta\} \tag{15}$$

式中，$M_i(t)$表示帖子情绪的可能结果，Θ 是模型的参数. 选取预测概率最大的作为该样本的类别，使用交叉熵作为损失函数在有标签的训练集上训练，我们得到一个 BERT-RCNN 分类器，由于是在 BERT 模型的基础上加入了循环神经网络 BiLSTM 和最大池化计算，故称其为 BERT-RCNN 分类器.

假设 $s_j(t)$是无标签数据，对无标签的数据，引入置信阈值，设帖子属于某个类别的置信阈值为 $C\in(0,1)$，若 $q_j(t)>C$ 就认为样本属于相应类别的置信度比较高，对 $q_j(t)\leqslant C$ 的无标签样本进行过滤，选取预测概率最大的作为该样本的伪标签：

$$y_j'^m(t)=\underset{M_j(t)}{\operatorname{argmax}}P\{M_j(t)\mid s_j(t),\Theta\},\ q_j(t)>C \tag{16}$$

将满足置信阈值的伪标签样本继续添加到训练集中去训练，直到总体损失不再变化为止.

上述模型，基于 BERT-RCNN 并引入了无标签数据的信息(称其为 semi-BERT-RCNN)，我们可以预测第 t 个交易日第 i 条帖子的情绪类别，预测的结果记为

$\varphi\{s_i(t)\mid\Theta\}\in\{(1,0,0),(0,1,0),(0,0,1)\}$，

分别表示积极、消极和中立情绪，其中，$\varphi(s_i(t)\mid\Theta)$表示训练好的 semi-BERT-RCNN 模型.

1.2.5 基于情绪得分的 Bagging 集成

假设第 t 个交易日的帖子有 $n(t)$ 条，利用 semi-BERT-RCNN 模型可得每条帖子的情绪类别，为简便起见，定义积极和消极情绪的得分公式分别为

$$\text{pos}(s(t))=\frac{1}{n(t)}\sum_{i=1}^{n(t)}\mathbf{1}_{\varphi\{s_i(t)\mid\Theta\}}=(1,0,0)(s_i(t)) \tag{17}$$

$$\text{neg}(s(t))=\frac{1}{n(t)}\sum_{i=1}^{n(t)}\mathbf{1}_{\varphi\{s_i(t)\mid\Theta\}}=(0,1,0)(s_i(t)) \tag{18}$$

式中，$\mathbf{1}$ 是示性函数，$s(t)$表示第 t 个交易日所有帖子.

用 D 表示第 t 个交易日所有帖子集合，由自助采样法(bootstrap sampling)，随机对数据集 D 有放回地进行一次采样操作，重复 $n(t)$次，可以得到含 $n(t)$个样本的采样集 D_{bs}. 按照上述步骤，可采样 C 个包含 $n(t)$个样本的采样集.

基于情感分类器的情绪得分公式，通过 Bagging 集成我们可以估计情绪的概率，作为后续股价预测的输入，有利于提高股价预测的准确率，具体算法步骤见算法 1.1.

算法 1.1 基于 Bagging 集成估计情绪的概率

Require：

帖子数据：D；

基学习算法：pos(s(t))，neg(s(t))；

训练轮数：C.

while $1\leqslant c\leqslant C$ **do**

$\text{pos}_c(t)=\text{pos}(D,D_{bs})$；

$\text{neg}_c(t)=\text{neg}(D,D_{bs})$；

end while

return $\pi_1(t)=\frac{1}{C}\sum_{c=1}^{L}\text{pos}_c(t)$；

$\pi_2(t)=\frac{1}{C}\sum_{l=1}^{C}\text{neg}_c(t)$.

2 相关技术

2.1 LDA 主题概率特征

文本的 LDA 特征[9]可用于股票趋势预测，本文也利用 LDA[12]方法对帖子文本进行提取，作为情绪变量的一部分. Python 的 gensim 库①采用在线变分推断计算 LDA 主题概率. 针对本文中的帖子文本，设主题有 K 个，$c_i(t)$表示第 i 条帖子的主题概率向量，$l_k(t)\in\mathbb{R}_{n(t)\times 1}$为当天帖子属于第 k 个主题的概率，设置主题的个数，计算当天的主题概率矩阵 $l(t)=[c_1(t),\cdots,c_{n(t)}(t)]^T$，其中 $l(t)\in\mathbb{R}_{n(t)\times K}$，生成矩阵 $l(t)$的一行即一条帖子的主题概率向量，向量中的每个元素即该评论属于某个特定主题的概率. 从而关于 K 个主题概率的时序特征为

$$l_k(t)=\frac{1}{n(t)}\sum_{j=1}^{n(t)}c_{i,k}(t),k=1,\cdots,K \tag{19}$$

2.2 正交试验设计

正交试验设计[21]可以在全面试验中挑选具有代表性的参数取值，由正交试验设计表进行多个对比试验，以选取最优的参数组合. 本文基于正交试

① https://radimrehurek.com/gensim.

验设计对多个调整参数选优，按验证集中预测的均方根误差最小来确定参数水平的最优组合，再在其附近进行更细致的试验，以确定出类比全局最优的超参数组合. 该试验方法有效地减少了代码运行的时间和成本.

3　上证指数股票价格预测实证分析

3.1　数据收集

本文选取 2015 年 1 月至 2019 年 1 月共 976 个交易日的数据进行实验，包括上证指数的历史 K 线数据和从东方财富股吧上证指数吧爬取的帖子文本，并选取数据集中最后 120 天的数据作为测试集. 由于上证指数吧在 2014 年发表的具有较高影响力的帖子数量仅为 2015 年的 30%，显然其在 2015 年之后的普及率和影响力更高，故实验选取 2015 年作为数据的起始时间. 而东方财富股吧是国内广受欢迎的在线社交媒体之一，与传统媒体相比较具有内容简洁、传播广泛的特点，由于是有关股票的论坛，其用户大多是对股票感兴趣的投资者. 论坛中每个用户发表的帖子都包含发表日期、内容、标题、用户影响力、帖子阅读量、帖子评论数量等特征，在爬取数据的过程中我们直接筛选出实验时间段内具有较高用户影响力、阅读量和评论量的帖子. 根据帖子内容对帖子作删除、去重等处理之后得到共计 95411 条帖子作为我们的文本数据.

3.2　技术指标

上证指数的历史数据可以通过雅虎财经获取. 如表 1 所示，上证指数的技术指标特征包括开盘价、收盘价、最高价、最低价和交易量(简称 OCLHV)，其他技术指标可以根据 OCLHV 的历史数据计算，其中 MA 表示移动平均.

表 1　上证指数的技术性指标

Tab. 1　Technical indicators of SSE Composite Index

指标编号	指标	描述或公式
0	开盘价	O_t
1	收盘价	C_t
2	最低价	L_t
3	最高价	H_t
4	交易量	V_t
5	股价变化	C_t-C_{t-1}
6	股价涨跌幅	$(C_t-C_{t-1})/C_{t-1}$
7	成交量变化	V_t-V_{t-1}
8	成交量涨跌幅	$(V_t-V_{t-1})/V_{t-1}$
9	DIF	$(C_t-O_t)/C_{t-1}$
10	振幅	$(H_t-L_t)/C_{t-1}$
11	MA7	$(C_t+C_{t-1}+\cdots+C_{t-6})/7$
12	MA21	$(C_t+C_{t-1}+\cdots+C_{t-20})/21$

3.3　基于构建伪标签的半监督学习提取情感特征实验结果与分析

本文随机抽取了 2018 年之前的 2623 条帖子结合帖子的标题及内容对情绪类别进行人工标注(用于标注数据不能在上证指数预测的测试集时间范围内). 如表 2 所示为帖子标注情感特征的实例，我们可以将每条帖子标注成三种类别(积极、消极和中性)，由五位具备一定金融基础的人士通过先阅读帖子文本然后采用投票的方式标注每一条文本的情感类别.

表 2　情感类别对应帖子的实例

Tab. 2　Examples of investors' sentiment

股吧文本	情感特征
十年买股两茫茫，先亏车，再赔房. 千古跌停无处.	消极
先按反弹处理吧，但是个股最好要找离开压力位空间大一些的.	积极
盘口说：温和法破位 3300，涌现第一批抄底者！	积极
帝妖：重组是资本市场的一个永恒的主题.	中立

模型的初始化基于预训练模型①，可采取交叉验证的方式对训练集进行训练，但是由于训练 BERT 时间和资源消耗较高，本文采取随机从有标签数据中取出 20%比例的数据，保存为测试集，再将剩余有标签数据保存为训练集，让这样的方式来对模型进行微调. 训练时 Batch_size 设置为 128，初始学习率为 0.001，梯度下降的优化器采用 Adam[22]. 设置 BiLSTM 的隐藏层大小为 256，为了防止其过拟合，设置 dropout 为 0.1，表示按照 0.1 的概率，对网络中的每个节点进行随机消除，最大池化计算窗口大小和步长都设为 32. 伪标签的置信阈值不宜过高和过低，简便起见设为 0.75，对模型选取适量的无标签数据作半监督学习，观察模型在训练集上的精度曲线不再有变化时，停止添加无标签数据进行伪标签学习，如图 2 所示是模型的训练精度曲线. 模型评估采用分类正确率，即类别被正确分类的比例.

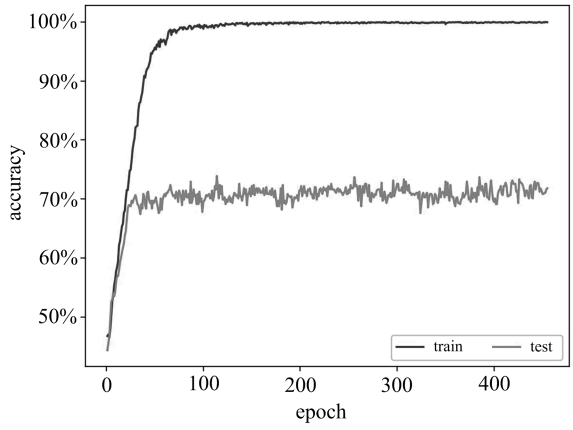


图 2　semi-BERT-RCNN 混合模型训练精度曲线

Fig. 2　The accuracy curve of semi-BERT-RCNN model training

训练 semi-BERT-RCNN 分类器，在测试集上得到的正确率为 73.9%. 训练只基于 BERT 的分类器，在测试集上的正确率为 72%. 可见 semi-BERT-RCNN 模型相比 BERT 模型增加了 1.9%的正确率. 分析模型相对于 BERT 分类器，正确率

① 谷歌提供了 6 种预训练模型，对中文只需使用 bert-base-chinese，https://github.com/google-research/bert#pre-trained-models.

没有大幅度提升,主要原因是用于训练的数据数量太小,并存在较多的噪声数据,但是总体上正确率依然有所提升,具备较好的金融文本情绪分析的能力.

3.4 双编码器-解码器的股价预测实验与分析

3.4.1 上证指数数据预处理

图 3 展示了上证指数关于时间的变动情况,可以观察到价格曲线在许多地方存在小范围内价格上的波动,Khaidem 等[23]对时间序列的历史数据做了平滑处理,过滤掉对股价趋势没有较大影响的微小波动,有利于预测股价的趋势变化,故本文对价格时序数据进行二阶指数平滑(second order exponential smoothing)[24]. 经过二阶指数平滑的价格序列如图 4 所示,通过对比平滑前后的序列变化图,可以观察到平滑后的价格序列很好地保留了原始序列的价格趋势,小范围内价格的波动情况也有所改善.

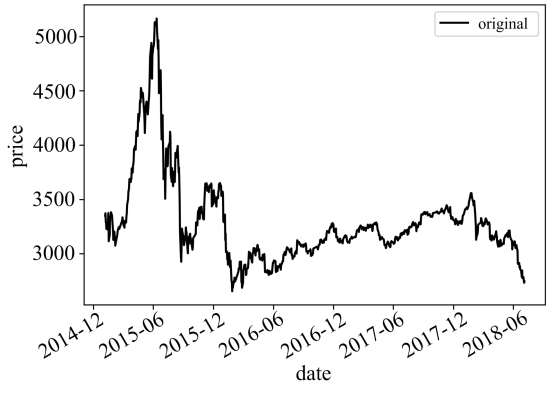


图 3 上证指数的价格序列

Fig. 3 The price curve of SSE Composite Index

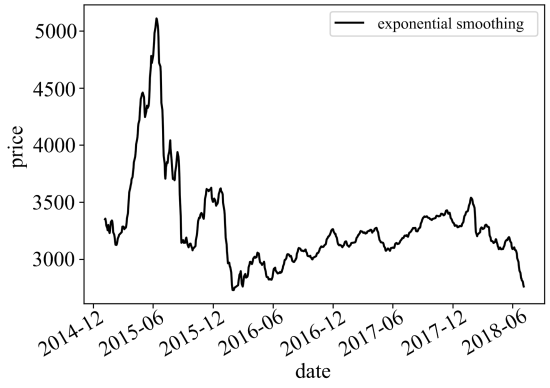


图 4 二阶指数平滑之后的上证指数价格序列

Fig. 4 The price curve of SSE Composite Index after second order exponential smoothing

3.4.2 模型评估

我们使用均方根误差(RMSE)、平均绝对误差(MAE)和平均绝对百分比误差(MAPE)作为回归模型的评估指标,设 y_t 表示第 t 个交易日响应变量的真实值,$\hat{y}_t$ 表示响应变量的预测值,则计算公式如下:

$$\mathrm{RMSE}=\sqrt{\frac{1}{N}\sum_{t=1}^{N}(y_t-\hat{y}_t)^2} \tag{20}$$

$$\mathrm{MAE}=\frac{1}{N}\sum_{t=1}^{N}|y_t-\hat{y}_t| \tag{21}$$

$$\mathrm{MAPE}=\frac{1}{N}\sum_{t=1}^{N}\left|\frac{y_t-\hat{y}_t}{y_t}\right| \tag{22}$$

3.4.3 基于正交实验设计多个超参数选优

本文对多个超参数调整选优设计正交试验,通过设计以少量次数的试验,来得到最佳因素水平的组合.本文中的股票预测模型超参数有 LDA 主题个数 K,时间窗口大小 T,两个编码器的隐藏层大小分别为 L_1^E 和 L_2^E,解码器隐藏层的大小 L^D,由于是双编码器-解码器框架模型,设 $L_1^E=L_2^E=L^E$ 且 $L^D=2L^E$. 从全面试验中挑选具有代表性的参数水平 $K\in\{5,10,15,20\}$,$T\in\{5,10,15,20,25\}$和 $L^E\in\{16,32,64,128,256\}$设计正交实验 $L_{25}(5^2 4^1)$(见表 3). 通常按照经典的网格搜索,对多个参数进行选优需 100 次试验,但若按照本文的正交试验设计表,只需做 25 次实验就能得到多个超参数的最优组合,代码运行时间大大缩短.尽管牺牲了全面实验下取得全局最优超参数的最优组合的可能性,但是通过正交实验设计,我们可以确定最优参数的大致范围,在正交设计表所确定的超参数的最优组合附近,还可以设计更加全面的试验,以得到全局调优的效果,这种对多个超参数进行选优的计算量,仍旧是远低于相对经典的网格搜索.

表 3 正交设计表 $L_{25}(5^2\,4^1)$

Tab. 3 Orthogonal design table $L_{25}(5^2 4^1)$

编号	T	L^E	K	编号	T	L^E	K
0	5	16	5	1	10	32	5
2	15	64	5	3	20	128	5
4	25	256	5	5	25	128	10
6	20	64	10	7	15	32	10
8	10	16	10	9	5	256	10
10	5	128	15	11	10	64	15
12	15	256	15	13	20	16	15
14	25	32	15	15	25	16	20
16	20	32	20	17	15	128	20
18	10	256	20	19	5	64	20
20	5	32	20	21	10	128	20
22	15	16	20	23	20	256	20
24	25	64	20				

图 5 展示了 RMSE 随窗口大小 T 的变化关系,随着窗口的变大,无论编码器的隐藏层大小和 LDA 主题个数为多少,RMSE 都受其影响呈上升的趋势增加,但是当 $T=5$ 时,RMSE 值不仅是最优的而且 RMSE 值的波动范围最小最稳定.此时,编码器隐藏层的大小为 64 ,LDA 主题的个数为 15,验证集上的模型结果最优.

为了防止模型过拟合现象的发生,我们在神经网络优化的过程中,设置权值衰减为 1E−6,以调节模型复杂度对损失函数的影响,梯度下降的优化器采用 Adam[22].

3.4.4 实验结果与分析

图 6 展示了上证指数在两输入双编码器-解码

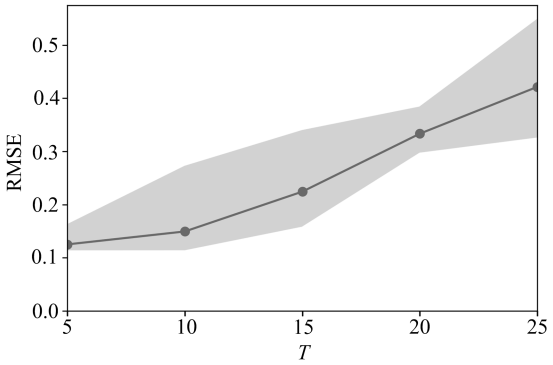


图 5　正交试验中验证集 RMSE 随窗口大小 T 的变化趋势
Fig. 5　The RMSE of validation varies over time in orthogonal experimental design

器框架模型下，在测试集上对未来第 1 天的股价预测结果，可见模型对股票价格的预测效果较好. 图 7 展示了上证指数在一般编码器-解码器模型[3]下的预测结果. 通过图 6 和图 7 预测曲线的对比可见，两输入双编码器-解码器对股价变化趋势的拟合，效果明显优于普通的编码器-解码器. 普通编码器-解码器下的预测结果，虽然也很好地反映了股价的变化趋势，但是在某些短期变化中却出现了相反预测结果，预测的结果比较不稳定，不能很好地贴合真实的股价变化趋势.

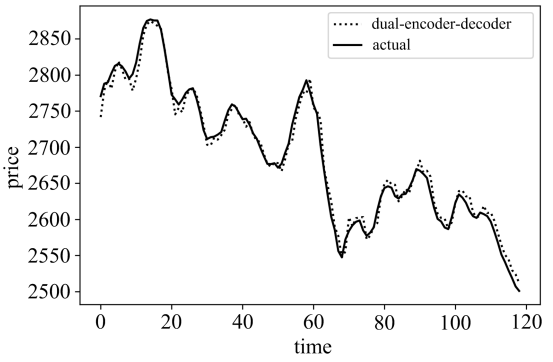


图 6　上证指数的双编码器-解码器测试集上的预测结果
Fig. 6　Predicted results of SSE Composite Index using dual-encoder-decoder on the test set

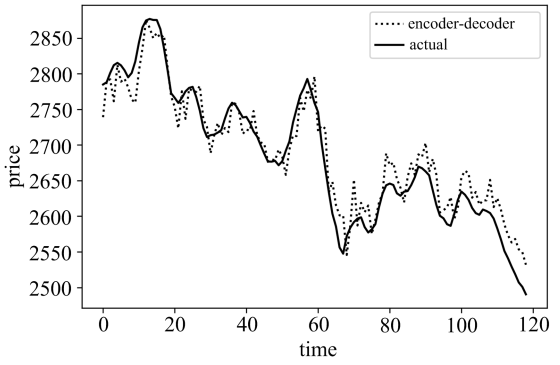


图 7　上证指数的一般编码器-解码器测试集上的预测结果
Fig. 7　Predicted results of SSE Composite Index using encoder-decoder on the test set

表 4 展示了双编码器-解码器在引入技术指标、LDA 特征和投资者情绪特征时和仅引入技术指标和 LDA 特征时对股价的预测结果，以及编码器-解码器在引入技术指标时，在引入技术指标和 LDA 特征时和在引入技术指标、LDA 特征和投资者情绪特征时对股价的预测结果. 可以观察到，无论数据输入的是什么，本文中双编码器-解码器的预测结果都优于普通编码器-解码器. 而普通编码器-解码器在加入情绪特征之后，预测效果反而下降，分析原因可能是情绪变量的信号相对技术指标较弱，并且由于是提取的变量，存在一定的噪声，使得普通编码器-解码器对信息不能很充分地进行提取. 在结构上，双编码器-解码器对两类信息的提取较充分，预测效果较好，在技术指标和 LDA 特征都输入时，引入市场情绪，模型性能在 RMSE 评估下有 12.81% 的提升.

表 4　模型在不同数据下的股价预测结果
Tab. 4　Predicted results of stock index using different models on different test sets

模型(数据)	RMSE	MAE	MAPE
Dual-Encoder-Decoder (U+L+M)	0.0993	0.076	0.2834
Dual-Encoder-Decoder (U+L)	0.1139	0.0867	0.3228
Encoder-Decoder(U)	0.1197	0.0946	0.3526
Encoder-Decoder(U+L)	0.1737	0.1441	0.5393
Encoder-Decoder(U+L+M)	0.1559	0.1271	0.4754

为了更进一步观察投资者情绪对股价预测的影响，本文在 RMSE 等评估指标的基础上，通过对上述两个模型的股价预测结果，计算其对股票趋势预测的准确率，如表 5 所示，可以看出，本文的两输入双编码器-解码器模型对股票的趋势预测结果更优，无论是 LDA 特征的引入，还是市场情绪对 LDA 特征的补充，都使模型的预测结果更准确.

表 5　模型在不同数据下对股票趋势的预测准确率对比
Tab. 5　Comparison of the accuracy of the models prediction of stock trends under different data

模型(数据)	准确率
Dual-Encoder-Decoder(U+L+M)	84.03%
Dual-Encoder-Decoder(U+L)	83.19%
Encoder-Decoder(U)	76.47%

表 6 展示了框架模型中，使用各 GRU 变体在测试集上对未来第 1 天的预测结果对比，为简便起见只对上证指数进行实验. 表中 MGU[25]、Li-GRU[26]都是基于 GRU 改进的神经网络变体. 可以看到，本文的方法 O-GRU 的预测结果较好，优于其他三种方法.

表 6　模型中各神经网络的效果对比
Tab. 6　Comparison of the predicted results using different neural networks in the model

模型中的神经网络	RMSE	MAE	MAPE
GRU	0.1112	0.085	0.3191
MGU	0.1108	0.0857	0.3191
Li-GRU	0.1301	0.1037	0.3867
O-GRU	0.0993	0.076	0.2834

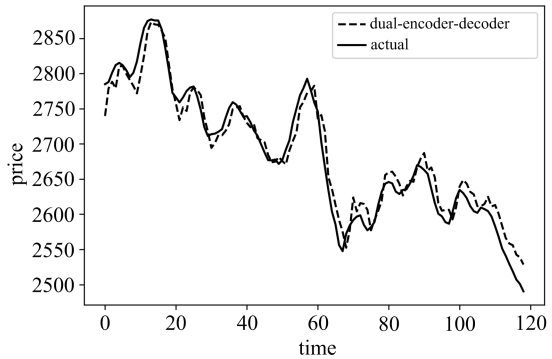
表 7 展示了两输入双编码器-解码器和普通编

码器-解码器对未来第 2 天、第 3 天股价的预测能力.可以观察到,两输入的双编码器-解码器的预测性能较优,而且单一编码器-解码器的预测差异较大.这表明两输入编码器-解码器的预测更稳定,预测结果更优.

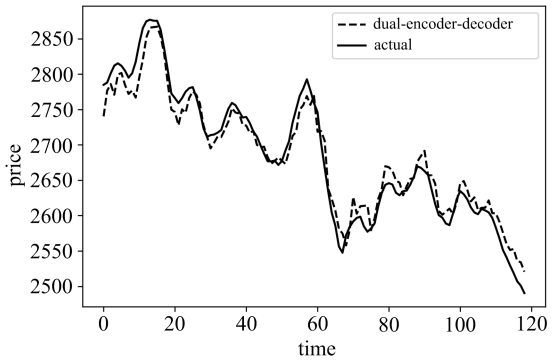
表 7 模型对未来第 2 天、第 3 天股价的预测结果

Tab. 7 Comparison of the predicted results on the 2nd and 3rd day in the future

时间	模型	RMSE	MAE	MAPE
未来第 2 天	Dual-Encoder-Decoder	0.2043	0.1658	0.6219
	Encoder-Decoder	0.2684	0.2169	0.8142
未来第 3 天	Dual-Encoder-Decoder	0.2738	0.2231	0.8338
	Encoder-Decoder	0.4170	0.3510	1.3062



(a) 未来第2天



(b) 未来第3天

图 8 两输入双编码器-解码器的预测结果

Fig. 8 Dual-encoder-decoder prediction results using two inputs compared with actual price

图 8 展示了两输入双编码器-解码器在测试集对上证指数未来第 2,3 天股价的预测结果,可以看到,由于输入的数据相对于预测的目标来说具有滞后性,导致预测结果也具有一定的滞后性,但是总体上预测结果还是较好拟合了价格的变化趋势.

为了更直观地观察,图 9 展示了两输入双编码器-解码器和单一输入编码器-解码器在测试集对上证指数未来第 3 天股价的预测结果,可以看到,双编码器-解码器框架下模型对未来第 3 天股价的变化趋势也能较好地拟合,但是相对来说普通编码器-解码器的预测结果的波动性太大而导致拟合效果不是很好.

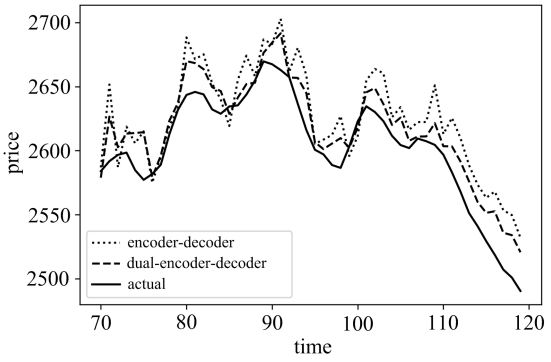


图 9 模型对未来第 3 天的预测结果对比

Fig. 9 Comparison of the model's prediction results for the third day in the future

4 结论

本文构造两输入的双编码器-解码器对股价进行预测,以单独对两类信息编码的方式,充分提取有利于股价预测的信息.本文将市场情绪看作服从某个概率分布的随机过程,基于构建伪标签的情感分类器建立积极和消极情绪的得分公式,并以得分公式提取的信息,基于 Bagging 集成的方法,对市场情绪进行估计,作为后续输入股价预测模型的特征.该方法充分利用股票多维度的信息,充分提取有利于股价预测的投资者情绪特征,将市场情绪引入到编码器-解码器模型中,为股市投资者提供了一种市场情绪估计和股价预测的可行方法.并通过在上证指数的实验结果表明,该方法不仅提高了股价预测的准确率还提高了股价预测的性能.

市场情绪是复杂多变的,具有极强的时效性,在未来的工作中,对于市场情绪的估计,还可以引入更多的数据源,例如新闻、微博和微信公众号等,尝试无监督的模型学习方法对市场情绪进行估计和提取,减少标签数据及时性和准确性对模型准确度的影响,充分发挥利用市场情绪对股价的预测作用.

参考文献(References)

[1] CHONG E, HAN C, PARK F C. Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies[J]. Expert Systems with Applications, 2017, 83: 187-205.

[2] FISCHER T, KRAUSS C. Deep learning with long short-term memory networks for financial market predictions [J]. European Journal of Operational Research, 2017, 270: 654-669.

[3] QIN Y, SONG D, CHEN H, et al. A dual-stage attention-based recurrent neural network for time series prediction[DB/OL]. [2020-03-01] https://arxiv.org/abs/1704.02971.

[4] 罗伯特·E·霍尔,马可·利伯曼.股票市场和宏观经济[M]// 经济学:原理与应用.2 版.北京:中信出版社, 2003.

[5] DE LONG J B, SHLEIFER A, SUMMER L H, et al. Positive feedback investment strategies and destabilizing rational speculation[J]. The Journal of Finance, 1990, 45: 379-395.

[6] NOFER M, HINZ O. Using Twitter to predict the stock market[J]. Business & Information Systems Engineering, 2015, 57: 229-242.
[7] PENG Y, HUI J. Leverage financial news to predict stock price movements using word embeddings and deep neural networks[DB/OL]. [2020-03-01]. https://arxiv.org/abs/1506.07220.
[8] CHEN W, YEO C K, LAU C T, et al. Leveraging social media news to predict stock index movement using RNN-boost[J]. Data & Knowledge Engineering, 2018, 118: 14-24.
[9] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3: 993-1022.
[10] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[DB/OL]. [2020-03-01]. https://arxiv.org/abs/1409.0473.
[11] CHO K, VAN MERRIENBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation [C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg, PA: Association for Computational Linguistics, 2014: 1724-1734.
[12] BAHDANAU D, CHOROWSKI J, SERDYUK D, et al. End-to-end attention-based large vocabulary speech recognition[DB/OL]. [2020-03-01]. https://arxiv.org/abs/1508.04395.
[13] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pretraining of deep bidirectional transformers for language understanding[DB/OL]. [2020-03-01]. https://arxiv.org/abs/1810.04805.
[14] LEE D H. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks[C]// ICML 2013 Workshop : Challenges in Representation Learning (WREPL), Atlanta, Georgia, USA, 2013.
[15] OLIVER A, ODENA A, RAFFEL C, et al. Realistic evaluation of deep semi-supervised learning algorithms [DB/OL]. [2020-03-01]. https://arxiv.org/abs/1804.09170,2018.
[16] LU X, NI B. BERT-CNN: A hierarchical patent classier based on a pre-trained language model[DB/OL]. [2020-03-01]. https://arxiv.org/abs/1911.06241.
[17] BENGIO Y, SIMARD P, FRASCONI P. Learning long-term dependencies with gradient descent is difficult[J]. IEEE Transactions on Neural Networks, 1994, 5(2): 157-166.
[18] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735-1780.
[19] CHUNG J, GULCLEHRE, CHO K, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[DB/OL]. [2020-03-01]. https://arxiv.org/abs/1412.3555.
[20] JOZEFOWICZ R, ZAREMBA W, SUTSKEVER I. An empirical exploration of recurrent network architectures [C]// ICML'15: Proceedings of the 32nd International Conference on International Conference on Machine Learning. JMLR.org, 2015, 37: 2342-2350.
[21] HEDAYAT A S, SLOANE N J A, STUFKEN J. Orthogonal Arrays: Theory and Applications[M]. New York: Springer, 1999.
[22] KINGMA D P, BA J. Adam: A method for stochastic optimization [DB/OL]. [2020-03-01]. https://arxiv.org/abs/1412.6980.
[23] KHAIDEM L, SAHA S, DEY S R. Predicting the direction of stock market prices using random forest[DB/OL]. [2020-03-01]. https://arxiv.org/abs/1605.00003.
[24] BROWN R G, MEYER R F. The fundamental theorem of exponential smoothing[J]. Operations Research, 1961, 9(5): 673-685.
[25] ZHOU G B, WU J, ZHANG C L, et al. Minimal gated unit for recurrent neural networks[DB/OL]. [2020-03-01]. https://arxiv.org/abs/1603.09420.
[26] RAVANELLI M, BRAKEL P, OMOLOGO M, et al. Light gated recurrent units for Speech Recognition[J]. IEEE Transactions on Emerging Topics in Computational Intelligence, 2018, 2: 92-102.