Finance 金融, 2020, 10(4), 366-373
Published Online July 2020 in Hans. http://www.hanspub.org/journal/fin
https://doi.org/10.12677/fin.2020.104037

Hans汉斯

LSTM Based Quantitative Stock Forecasting

Jianqun Zhao, Qi Zhang, Yue Wang

The College of Economics and Business Administration, Guangdong Polytechnic of Science and Trade, Guangzhou Guangdong
Email: 576261250@qq.com

Received: Jun. 25th, 2020; accepted: Jul. 9th, 2020; published: Jul. 16th, 2020

Abstract

The features of stock are usually mixed with many noise data, and noisy data will affect the prediction accuracy of stock prediction model. In this paper, a quantitative coding method for stock data features is proposed, and a prediction model is constructed by using short and long term memory network to predict the quantified data. The data set uses the Shanghai and Shenzhen 300 component stocks, after the stock data quantification carries on the 3 classification rise and fall forecast. The experimental results show that the prediction effect is better than that of the original data after the stock feature is processed by quantitative coding.

Keywords

Characteristic Quantification, LSTM, Shanghai and Shenzhen 300, Forecast of Increase or Decrease

基于LSTM的量化股票预测

赵建群，张　岐，王　悦

广东科贸职业学院经济管理学院，广东　广州
Email: 576261250@qq.com

收稿日期：2020年6月25日；录用日期：2020年7月9日；发布日期：2020年7月16日

摘　要

股票特征通常夹杂较多噪声数据，而带噪数据会影响股票预测模型的预测精度。本文提出一种对股票数据特征进行量化编码的方法，并使用长短期记忆网络构建预测模型，对量化后的数据进行预测。数据集采用沪深300成分股，在对股票数据量化后进行3分类涨跌幅预测。实验结果表明，使用量化编码对股票

文章引用: 赵建群, 张岐, 王悦. 基于 LSTM 的量化股票预测[J]. 金融, 2020, 10(4): 366-373.
DOI: 10.12677/fin.2020.104037

特征处理后，预测效果优于使用原始数据预测。

关键词

特征量化，LSTM，沪深300，涨跌幅预测

1. 引言

随着国内股票市场的逐步完善，越来越多的研究机构和公司投入股票的分析和预测研究。对股票指数进行分析预测，能够快速掌握股票市场的发展动态，也能一定程度上反映国家的经济状况。由于影响股票数据的因素复杂多变，且股票序列本身具有随机游走等特性[1]，使得股票预测成为经济领域的一大难题。

股票预测已有较长的历史。统计学方法是最先用于现代股票市场预测的方法之一，其通过对股票数据的整理和分析，预测股票价格的短期走势。其中较为成功的有自回归移动平均(ARMA)模型[2]，该模型由于对股票数据优异表现沿用至今。SVM [3]、集成学习[4]等机器学习的方法也常用于股票预测，王禹等人 boosting 的方法，在添加股票的纵向变化指标后预测股票走势[5]，在股票预测的精度上取得了一定的提升。深度神经网络也是现在主流的股票预测模型之一，在股票预测上优异的表现得益于其强大的非线性拟合能力。邓凤欣利用 LSTM 网络对股票价格的走势做了预测，其结果表明使用 LSTM 的预测精度水平较高[6]。

在股票特征处理上，现有的研究大都使用特征归一化处理。同时在预测方式上，以拟合股票收盘价走势和预测股票涨跌为主。本文在已有研究的基础上，通过对股票交易的特征进行量化和编码，对股票价格的涨跌幅度进行预测。在网络结构上，选择 LSTM 作为量化股票预测的基本网络，使用量化后的数据对其进行训练，以预测股票价格的涨跌幅范围，并将结果和未将股票特征量化训练的模型做了比较分析。

2. 特征量化和模型结构

2.1. 股票数据集描述

股票价格的变动通常和多种因素有关，其中能直接影响到其变化的则是股票的基本交易指标[7]。开盘价、最高价、最低价、收盘价以及成交量等 5 维数据是最为常用的股票交易指标，使用神经网络预测股票时，常将这 5 维数据作为模型的输入，预测股票的价格走势或涨跌变化。本文也使用这 5 维特征作为股票数据的输入特征。

在股票数据的选择上，使用中证 100 指数的成分股作为基础数据来源。沪深 300 是为了反映沪深股市超大市值公司的股票价格表现而挑选出的 300 只股票[8]。由于该股票集的选取标准严格，其成份股市值高，影响大，所以研究该数据集的股票发展规律也有助于分析国内股票市场。

在数据集时间跨度上，本文使用 2013 年~2017 年的沪深 300 指数的成分股票，并将筛选后的股票数据集划分为训练集和测试集。其数据集详细信息见表 1。

Table 1. Overview of stock data sets
表 1. 股票数据集概况

数据集	时间跨度	交易天数(天)	股票数量(只)	股票数据总量
沪深 300	2013~2017	1215	160	194,400
训练集	2013~2016	971	160	155,360
测试集	2017	244	160	39,040

2.2. 股票的特征量化

在训练时选择包含多只股票的数据集时，由于各只股票受上市时间、公司规模等因素的影响，其股票价格差别较大，这导致模型难以学习整体股票数据的变化规律；除此之外，股票市场易受各种外部因素的影响，从而造成涨跌停等特殊情况，导致该时间段内的股票特征数值过高或过低。这样的数据对模型来说等同于噪声数据，会对模型的学习造成误导。因此本文在构建整体股票数据集时，将先对每只股票数据按其价格范围进行均匀量化，然后再将其合并到整体的数据集中。

在量化之前，本文先对因部分股票停盘而造成的缺失数据进行填充。填充方式为将缺失数据上一交易日的股票数据填充为缺失数据。这样在每只股票的时间维度上可以保持统一。

具体量化时，将单只股票的 5 维特征在最大值和最小值之间划分为 4 个区间，从低到高依次编码为 1~4。然后按照每个特征值所属的区间，将其编码为对应的数值。

表 2 是原始的部分股票特征数据，表 3 则是按照上述方法进行量化编码后的特征数据。对比可以看出，在经过量化处理后，在同一时间段上，由于股票体量等因素造成的价格差异被消除了。同时，对同一只股票，其涨跌停的数据经过量化平滑后，数值变化波动较小，有利于过滤噪声数据。

除此之外，股票特征中成交量其量纲和其他 4 维数据不同，造成在数量级有较大的差异，而量化编码过程是在每一维股票特征上单独进行的，因此经过量化，原来量纲带来的影响也被消除，这样数据不用再经过归一化的处理。

Table 2. Partial stock characteristic data
表 2. 部分股票特征数据

日期	开盘价	最高价	收盘价	最低价	成交量	涨跌幅
2017/7/24	10.82	11.06	10.95	10.73	1,692,664	0.55
2017/7/25	10.98	11.27	11.0	10.95	1,954,768	0.46
2017/7/26	10.92	11.18	10.74	10.66	1,697,412	-2.36

Table 3. Some stock characteristic data after quantification
表 3. 量化后的部分股票特征数据

日期	开盘价	最高价	收盘价	最低价	成交量	涨跌幅
2017/7/24	1	1	1	1	1	0.55
2017/7/25	1	2	2	1	3	0.46
2017/7/26	1	1	1	1	2	-2.36

2.3. 神经网络结构基础

神经网络通过强大的非线性拟合能力，成为最为流行的人工智能模型之一，在分类和预测任务中都

有优异的表现。它通过模拟神经单元的构造和功能来设计相关的网络模型，再通过不同的连接方式搭建不同的神经网络[9]。

1986 年 Rumelhart 等人提出的反向传播(BP)的前馈神经网络是最早提出的神经网络，其通过引入隐藏层来实现非线性计算[10]；循环神经网络(RNN)则在序列相关的数据上做了改进，通过在神经网络层之间引入有向循环连接，使其具备时序数据的记忆功能[11]。但 RNN 在长期依赖的问题上，随着需要提取的依赖项的长度增加时，梯度将不会再更新，甚至出现梯度消失等现象，因此需要对 RNN 进行改进，以解决信息的长期依赖[12]。

长短期记忆网络(LSTM)是 RNN 的一种改进网络结构[13]，相较普通的神经网络，LSTM 加入了可控门单元。LSTM 结构如图 1 所示，其包括输入门 i_t、输出门 o_t、遗忘门 f_t 等门结构，这些特殊的设计能够实现细胞状态 C_t 的更新，式(1)、(2)、(3)是 3 个门的更新公式。其中，σ 表示激活函数，W 和 b 分别表示权重矩阵和偏置，下标 f、i、o 代表遗忘门、输入门和输出门。h_{t-1} 表示 $t-1$ 时 LSTM 细胞的输出，x_t 代表 t 时刻的输入。

$$f_t = \sigma\left(W_f \cdot \left[h_{t-1}, X_t\right] + b_f\right) \tag{1}$$

$$i_t = \sigma\left(W_i \cdot \left[h_{t-1}, X_t\right] + b_i\right) \tag{2}$$

$$o_t = \sigma\left(W_o \cdot \left[h_{t-1}, X_t\right] + b_o\right) \tag{3}$$

在计算时，首先从当前时刻的输入和上一时刻隐藏层状态的信息中选择出要遗弃的部分，该部分是由遗忘门来实现的。然后由输入门决定神经单元要更新的值，通过 Tanh 函数更新神经单元状态。最终神经单元输出的状态由输出门决定，其先用 Sigmoid 层决定要输出的神经单元状态，然后将这些状态用 Tanh 函数压缩在−1 到 1 之间。

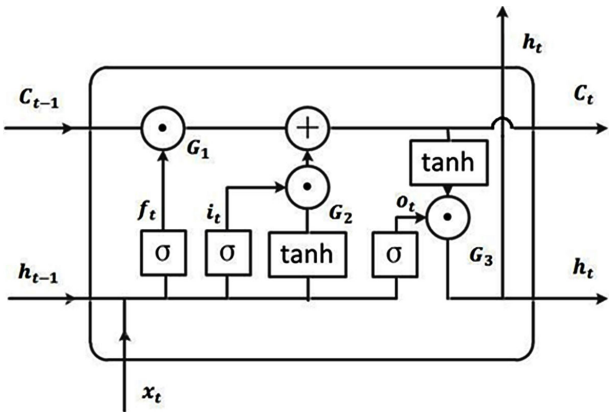

Figure 1. LSTM unit structure
图 1. LSTM 单元结构

2.4. 股票预测模型

本文使用 LSTM 网络作为基本网络结构，将量化编码后的股票特征数据作为输入变量。在模型输出上，使用涨、跌和小幅度涨跌等 3 类的分类方式做涨跌幅预测。

为了确定 3 个类别的划分界限，本文对股票数据集的涨跌幅做了统计。图 2 是股票数据集股票价格涨跌幅分布的直方图，可以看出，沪深 300 股票集的股票涨跌幅主要集中在±5%以内，也有部分样本出现了涨跌停的情况。

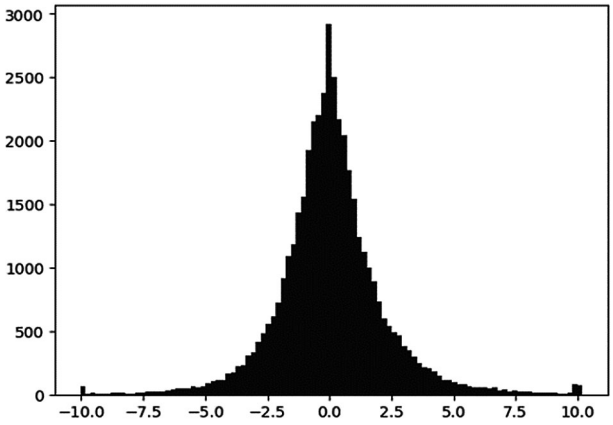

Figure 2. The rise and fall range distribution of the data set
图 2. 数据集的涨跌幅分布

图 3 是以涨跌幅 1%为界限的 3 分类直方图，从图可以看出，在涨(涨幅超过 1%)、跌(跌幅超过 1%)和小幅度涨跌(涨跌幅在±1%之间)的样本数量接近。分类预测中，多个分类的样本数均衡分布，能提高模型的精度。

因此，本文将股票的涨跌幅以上述分类方式做 3 分类，即在涨区间的即为 0；小幅度涨跌区间编码为 1；跌区间编码为 2，并将这 3 个分类用 Onehot 表示为：[0,0,1]、[0,1,0]、[1,0,0]。将编码后的涨跌分类作为标签，用以监督模型的训练。

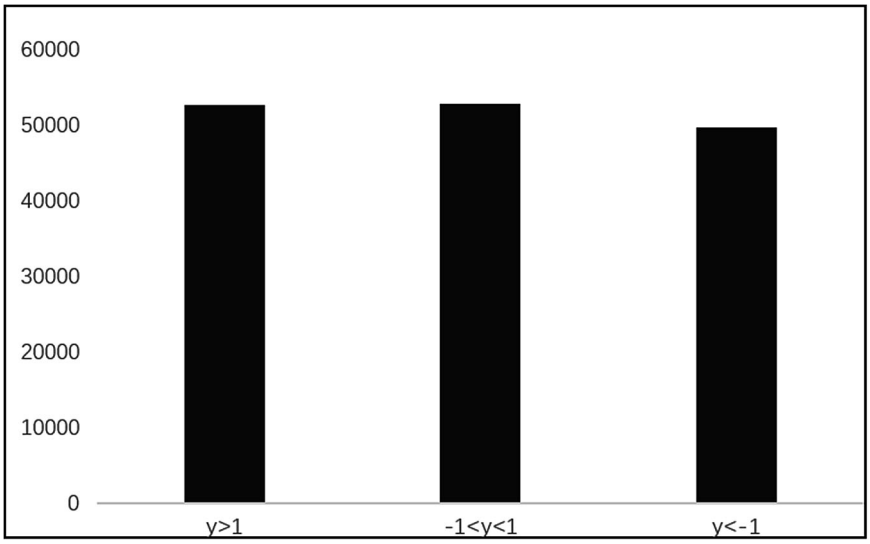

Figure 3. Sample distribution of three categories
图 3. 3 个分类的样本分布

在框架选择上，使用 Tensorflow 作为深度学习的框架，并使用 GPU 加速计算。在网络结构上，输入数据先进入一个全连接层，然后连接 1 层 LSTM 网络层。最后在输出端外接一个 Softmax 层，可将输出转换为各个类别的概率，从而输出预测的类别。神经网络内部节点数为 64，初始的学习率为 0.006。

图 4 是股票预测模型。由于使用 3 分类的预测方式，本文选择交叉熵函数作为模型的损失函数，用于评估模型预测的各类别的概率分布和实际分布之间的差异。预测值和标签值越为接近，损失函数的值就越小。在模型输出最终输出值后，需要将误差反向传播回去，不断更新模型的权重值，直至训练迭代结束。

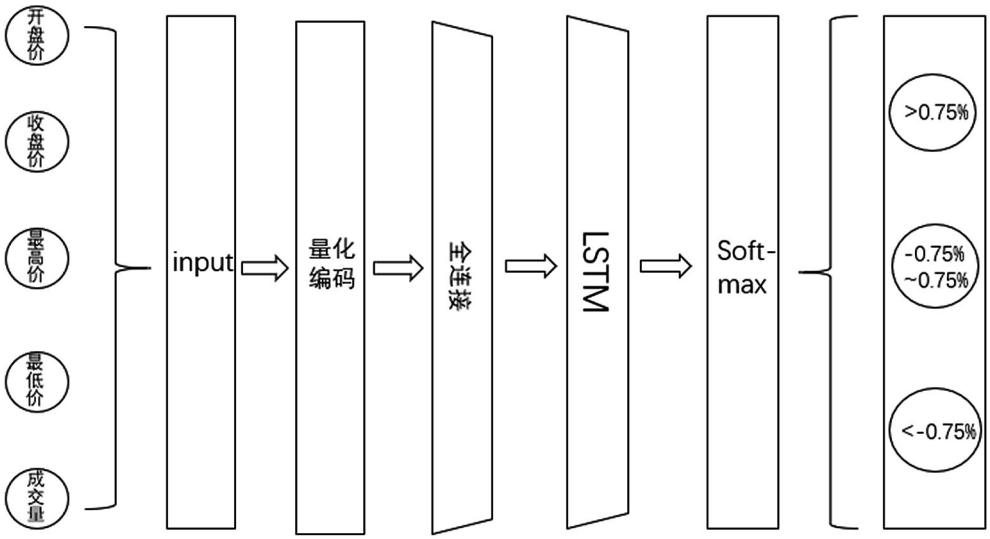

Figure 4. Stock forecasting model
图 4. 股票预测模型

3. 实验与结果

为了验证股票特征量化处理的预测效果，本文在训练数据集上分别使用经过量化编码处理后的股票特征数据和经过归一化处理的原始数据，其他的参数保持一致，训练后对应的模型分别为模型 1 和模型 2。本文将处理后的数据集划分为训练集、验证集和测试集。对每只股票，将其数据的前 80%作为训练集，剩余 20%作为测试集。训练集中，又以 8 比 2 的比例将其划分为训练集和验证集。训练集用于训练调节 LSTM 网络参数，测试集用于评测效果。验证集数据用于判断模型收敛时的迭代次数，避免过拟合。

在模型评估上，本文除使用分类准确率作为判断模型优劣的标准外，还使用了精确率、召回率和 F1 值等指标，以较为全面的比对两个模型在预测性能上的差异。

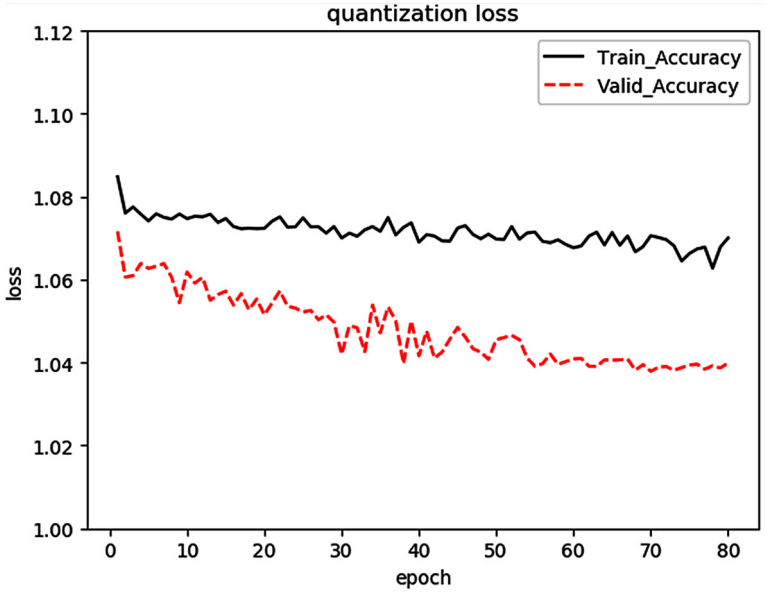

Figure 5. Model 1 loss function variation
图 5. 模型 1 损失函数变化

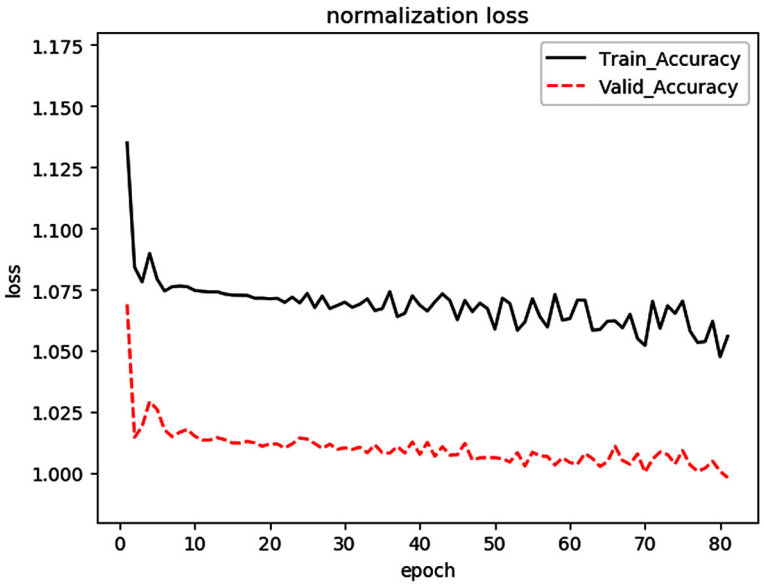

Figure 6. Model 2 loss function change
图 6. 模型 2 损失函数变化

图 5 和图 6 分别是模型 1 和模型 2 训练过程中损失函数的变化图象。由图可以看出，使用量化的股票交易数据训练模型时，在迭代了 60 次左右收敛。相比使用非量化数据，模型收敛速度较快，在迭代 20 次时就已经收敛。在模型保存上，保存验证集损失函数最小时的模型，作为测试集的预测模型。

表 4 是模型 1 和模型 2 在测试集上的表现，由表可以看出，模型 1 在准确率和 F1 上的表现都优于模型 2，这说明对股票交易数据进行量化处理可以有效的过滤噪声数据，使特征数据更加平滑，利于模型对股票价格变动规律的学习。

Table 4. Performance of different models in test set
表 4. 不同模型在测试集的表现

模型名称	准确率	精度	召回率	F1
模型 1	0.56	0.58	0.49	0.55
模型 2	0.53	0.52	0.41	0.46

为了测试模型单只股票预测效果，本文从数据集中随机抽选了 3 只个股，分别在两个模型上测试预测精度。

Table 5. Performance of single stock on Model 1
表 5. 单只股票在模型 1 上的表现

股票名称	准确率	精度	召回率	F1
浦发银行	0.60	0.72	0.49	0.56
中信证券	0.57	0.39	0.35	0.36
中兴通讯	0.56	0.41	0.33	0.34

Table 6. Performance of single stock on model 2
表 6. 单只股票在模型 2 上的表现

股票名称	准确率	精度	召回率	F1
浦发银行	0.59	0.58	0.49	0.55
中信证券	0.55	0.42	0.19	0.24
中兴通讯	0.53	0.25	0.21	0.29

从表 5 和表 6 可以看出，在单只股票预测上，由于个股的发展规律各不相同，预测精度也各有差别。在准确率上，单只股票在两个模型的表现较为接近，但在 F1 上，模型 1 要优于模型 2。因此，在实际对单只股票进行预测时，可以使用对股票特征进行量化的方式训练模型。

4. 总结

在股票涨跌预测上，本文使用对股票特征量化的方式，对股票特征数据进行过滤和平滑，使用量化编码后的数据和原始数据分别训练股票预测模型。结果表明，在沪深 300 数据集上，经过量化处理的股票特征更容易让模型学习到股票价格发展的规律，其测试集的各项指标都优于使用原始数据训练的模型。在单只股票预测上，使用量化数据训练的模型表现也优于使用原始数据训练的模型。

但是，金融时间序列预测的数据来源于市场的真实数据，而市场运行于博弈过程中。如果某种预测方法明显优于其他方法，且在相应的市场操作过程中该预测方法起到一定作用时，则会随着时间的推移，其性能会下降，因此，也就不会有预测准确性能特别优异的算法存在。而本文提出的基于股票数据特征量化编码的数据处理方法，并结合 LSTM 网络构建的预测模型，其主要目的是在没有特别突发事件的影响下，在仅仅基于主要的交易数据本身(不考虑其他事件)，尽可能获得一种适应能力更强的、预测准确度较高、性能稳定的预测方法。

参考文献

[1] 刘长虎, 陶建格, 崔衍秋. 股票价格指数的投资功能[J]. 市场论坛, 2004(3): 71-72.

[2] 丁玮珂. 基于 ARMA 模型预测股票价格的实证分析[J]. 广西质量监督导报, 2019(5): 151-153.

[3] 高雯. 基于支持向量机参数优化算法的股票智能投顾策略研究[D]: [硕士学位论文]. 上海: 上海师范大学, 2018.

[4] Lai, R.K., Fan, C.Y., Huang, W.H., *et al*. (2009) Evolving and Clustering Fuzzy Decision Tree for Financial Time Series Data Forecasting. *Expert Systems with Applications*, **36**, 3761-3773. https://doi.org/10.1016/j.eswa.2008.02.025

[5] 王禹. 基于 Cart 树和 Boosting 算法的股票预测模型[D]: [硕士学位论文]. 哈尔滨: 哈尔滨理工大学, 2018.

[6] 邓凤欣. LSTM 神经网络在股票价格趋势预测上的应用[D]: [硕士学位论文]. 广州: 广东外语外贸大学, 2018.

[7] 吴贻鼎, 朱翔, 黄继瑜, 明海山. 基于神经网络的证券市场预测[J]. 计算机应用, 2002(5): 31-33.

[8] 唐勇, 洪晓梅, 朱鹏飞. 投资者情绪与股票价格之间的信息溢出效应研究——基于行业差异视角[J]. 武汉金融, 2019(9): 49-57.

[9] 阎平凡, 张长水. 人工神经网络与模拟进化计算[M]. 北京: 清华大学出版社, 1900.

[10] Rumelhart, D.E., Hinton, G.E. and Williams, R.J. (1986) Learning Representations by Back-Propagating Errors. *Nature*, **323**, 533-536. https://doi.org/10.1038/323533a0

[11] Chung, J., Gulcehre, C., Cho, K., *et al*. (2015) Gated Feedback Recurrent Neural Networks. 32*nd International Conference on Machine Learning*, 2067-2075.

[12] Bengio, Y. (1994) Learning Long-Term Dependencies with Gradient Descent Is Difficult. *IEEE Transactions on Neural Networks*, **5**, 157-166. https://doi.org/10.1109/72.279181

[13] Hochreiter, S. and Schmidhuber, J. (1997) Long Short-Term Memory. *Neural Computation*, **9**, 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735