Advances in Applied Mathematics 应用数学进展, 2020, 9(11), 1954-1960
Published Online November 2020 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2020.911225

Hans汉斯

基于主成分分析和长短期记忆网络的股票价格预测

刘　甲，孙德山

辽宁师范大学数学学院，辽宁　大连
Email: 1131754935@qq.com

收稿日期：2020年10月25日；录用日期：2020年11月11日；发布日期：2020年11月18日

摘　要

运用神经网络技术，建立基于主成分分析的长短期记忆神经网络(PCA-LSTM)模型并对股票开盘价格进行预测。实验采用五粮液(000858)股票，首先，利用主成分法对该股票的多个指标进行特征提取，然后利用提取的主成分建立LSTM神经网络模型，并与PCA-Elman、LSTM模型对比，结果发现PCA-LSTM模型的预测结果更好一些。

关键词

长短期记忆网络，主成分分析，Elman网络

Stock Price Prediction Based on Principal Component Analysis and Long Short-Term Memory Network

Jia Liu, Deshan Sun

School of Mathematics, Liaoning Normal University, Dalian Liaoning
Email: 1131754935@qq.com

Received: Oct. 25th, 2020; accepted: Nov. 11th, 2020; published: Nov. 18th, 2020

Abstract

Using the neural network technology, the long and short term memory neural network (PA-LSTM) model based on principal component analysis was established and the stock opening price was

文章引用: 刘甲, 孙德山. 基于主成分分析和长短期记忆网络的股票价格预测[J]. 应用数学进展, 2020, 9(11): 1954-1960. DOI: 10.12677/aam.2020.911225

predicted. In the experiment, wuliangye (000858) stock was used. First, multiple indexes of the stock were extracted by principal component method. Then, LSTM neural network model was established by using the extracted principal component and compared with THE PCA-Elman and LSTM models. The results show that the prediction results of PCA-LSTM model are better.

Keywords

Long and Short Term Memory Network, Principal Component Analysis, Elman Network

1. 引言

运用神经网络来预测金融数据成为当今热点，许多不同类型的神经网络算法也由此产生。股票市场一直受到投资者的青睐，股票价格的变化对投资者的收益有十分重大的影响，因此通过预测未来时刻股票价格进行规避风险是股票研究的一项重要内容[1]。学者们认为神经网络可以被人们广泛利用，能较好的优化网络结构。

由于股票数据的特殊性，会受到很多因素的影响，宏观上来看，社会、政治、经济和文化等方面会影响市场价格。微观上看，产业发展前景和区域经济发展状况以及市场等因素也会影响股市。这就使得我们在选取输入变量的时候，容易导致信息量过于庞大，使得神经网络结构过于复杂，加重了神经网络的训练负担，学习速度就会急剧下降；并且，主观上的选择可能影响输入变量与输出变量的相关性，从而导致神经网络运行效率和预测精度变低。

针对这一问题，引入了统计学中的主成分分析法(PCA)，对影响数据的输入变量先进行筛选，通过降维的方法挑出较少的变量，而这些变量对整体数据有 85%以上的贡献值，减轻了神经网络的训练负担，并提高了学习率。在降维后使用长短期记忆网络的输入参数减少，提高了网络的运行效率和预测精度。同样对比不降维直接将数据在LSTM中训练，我们就可以发现PCA-LSTM的优点，本文还将引入PCA-Elman网络，对比其他模型的预测精度。

2. 主成分分析和 LSTM 网络

2.1. 主成分分析

主成分分析(Principal Component Analysis, PCA)是比较常用的降维方法，它将原变量的相关性进行转换从而使得维数减少，转换以后的变量叫做主成分。

主成分分析定义如下：

设 d 维随机变量 $X=\left(X_1,X_2,\cdots,X_d\right)^{\mathrm{T}}$，对原始空间进行线性变化。变换 $d'\left(\le d\right)$ 维的随机变量 $Y=\left(Y_1,Y_2,\cdots,Y_{d'}\right)^{\mathrm{T}}$。

$$Y=W'X,$$

其中 $W\in R^{d\times d'}$ 是变换矩阵，$W=\left[W_1,W_2,\cdots,W_{d'}\right]$，$W_i$ 是 d 维列向量。

设 X 的协方差矩阵为 Σ，随机变量 Y 的协方差矩阵为 $D\left(Y\right)=D\left(W^{\mathrm{T}}X\right)=W^{\mathrm{T}}D\left(X\right)W=W^{\mathrm{T}}\Sigma W$，于是优化目标为

$$\max tr\left(W^{\mathrm{T}}\Sigma W\right)$$

根据拉格朗日乘子法有

$$L(W)=tr\left(W^{\mathrm{T}}\Sigma W\right)-\lambda\left(W^{\mathrm{T}}W-I\right)$$

$$\frac{\partial L(W)}{\partial W}=2\Sigma W-2\lambda W=0$$

于是得

$$\Sigma W=\lambda W$$

这样通过求 Σ 的特征根和特征向量就可以解出 W，将求得的特征值进行排序：$\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_d$，取前 d' 个特征值对应的特征向量构成 $W=\left(w_1,w_2,\cdots,w_{d'}\right)$，这就是主成分的解。

解出 W 后，$Y_i=w_i^{\mathrm{T}}X,i=1,2,\cdots,d'$，于是有

$$D\left(Y_i\right)=D\left(w_i^{\mathrm{T}}X\right)=w_i^{\mathrm{T}}D\left(X\right)w_i=w_i^{\mathrm{T}}\Sigma w_i=w_i^{\mathrm{T}}\lambda_i w_i=\lambda_i$$

且 $\mathrm{cov}\left(Y_i,Y_j\right)=0\left(i\neq j\right)$，可知 Y_1 是方差最大的，为第一主成分，以此类推。

实际中我们用样本方差矩阵估计。

取 n 个样本，得到样本观测值为

$$X=\begin{bmatrix}x_{11}&x_{12}&\cdots&x_{1d}\\x_{21}&x_{22}&\cdots&x_{2d}\\\vdots&\vdots&\ddots&\vdots\\x_{n1}&x_{n2}&\cdots&x_{nd}\end{bmatrix}$$

每行为原始 d 维随机变量的一个观测值，这里用 X 表示数据矩阵。

确定 W 后，将原始数据矩阵 X 降维后得到：

$$Y=W^{\mathrm{T}}X^{\mathrm{T}}=\begin{bmatrix}y_{11}&y_{21}&\cdots&y_{n1}\\y_{12}&y_{22}&\cdots&y_{n2}\\\vdots&\vdots&\ddots&\vdots\\y_{1d'}&y_{2d'}&\cdots&y_{nd'}\end{bmatrix}$$

这样原始数据就由 d 维降到 d' 维

$$X=\begin{bmatrix}x_{11}&x_{12}&\cdots&x_{1d}\\x_{21}&x_{22}&\cdots&x_{2d}\\\vdots&\vdots&\ddots&\vdots\\x_{n1}&x_{n2}&\cdots&x_{nd}\end{bmatrix}\rightarrow\begin{bmatrix}y_{11}&y_{21}&\cdots&y_{n1}\\y_{12}&y_{22}&\cdots&y_{n2}\\\vdots&\vdots&\ddots&\vdots\\y_{1d'}&y_{2d'}&\cdots&y_{nd'}\end{bmatrix}$$

2.2. LSTM 网络

LSTM (Long Short-Term Memory)是长短期记忆网络，是一种特殊的递归神经网络，与常规递归网络相比能够有效地解决长时间依赖数据信息的问题，防止训练过程出现梯度消失和梯度爆炸的问题。

LSTM 神经网络含有智能网络单元，因为它可以记忆不定时间长度的数值，网络单元中的 gate 可以决定 input 能否能被记住以及 output 能否被输出，是一种通过门控状态控制传输状态的神经网络。

LSTM 在算法中加入“cell”处理器，包含三扇门：

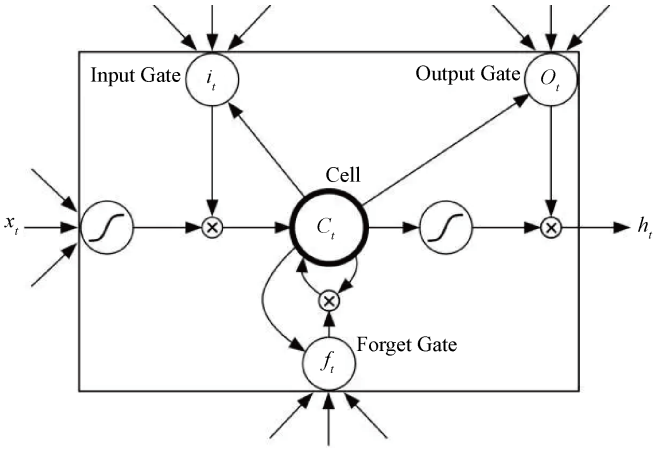
1、遗忘门(Forget Gate)，是 LSTM 的忘记阶段，对上一节点 f_t 的输入信息选择性忘记，留下有用的

信息，优化了数据信息。

2、输入门(Input Gate)，是 LSTM 的记忆阶段，随着时间对 i_t 来不断更新状态值 Ct，选择留下有用的状态值，抛弃无用的数值。

3、输出门(Output Gate)，是 LSTM 的输出阶段，经遗忘门和输入门对信息 o_t 的不断筛选过滤，最终决定当前状态的输出 h_t。

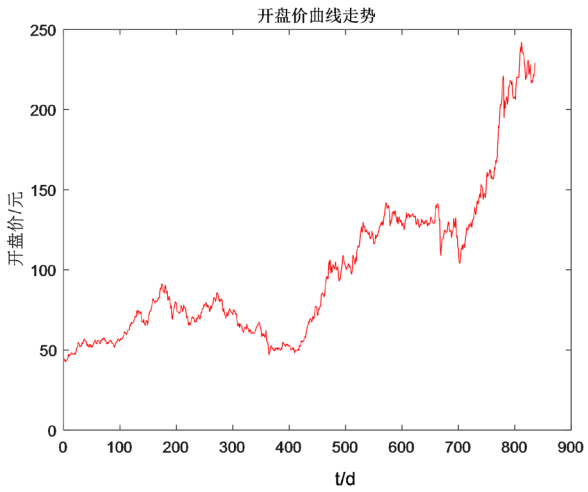
最重要的是遗忘门(Forget gate)，遗忘门不受人为因素的干扰，自行决定保留多久的记忆[2]。



3. 实证分析

3.1. 数据选取

本文选取五粮液(000858) 2017 年 5 月 5 日到 2020 年 10 月 12 日共 836 个交易日的数据。分别用开盘价、收盘价、最高价、最低价、成交量、MA.MA1、OBV.OBV、MACD.DIFF、KDJ.K、等 9 个指标预测未来 5 日股票的开盘价。我们选取开盘价为输出变量，其他指标作为输入变量。图中可以看出五粮液股票的开盘价格浮动没有明显的规律。

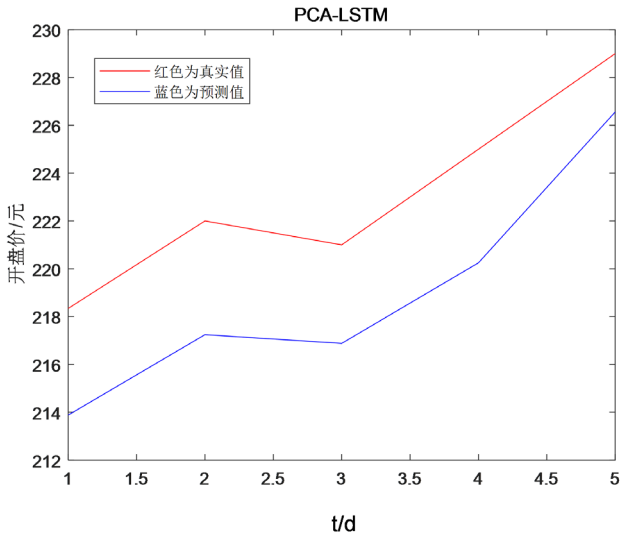


3.2. PCA-LSTM 模型

首先把数据作标准化处理，采用“max-min 标准化”方法，这样处理数据消除了量纲的影响，但不

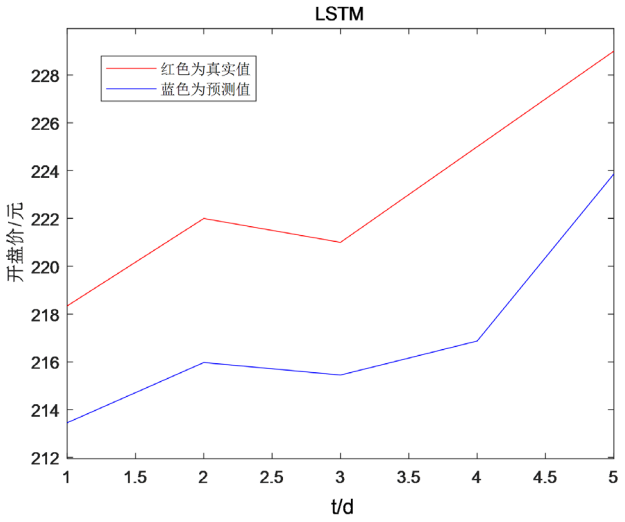
改变数据的原始意义。

标准化后的数据进行主成分分析将原来 9 维的数据降到了 3 维，贡献率已达到 95%。降维后将数据输入到 LSTM 神经网络中进行训练，选取前 831 个数据作为训练样本，后 5 个数据作为测试样本，采用前一日的特征指标来预测下一日的股票开盘价格。PCA-LSTM 选取 Adam 作为优化器，经过多次尝试训练，发现当初始学习率为 0.008，迭代次数为 200 次的时候效果最佳。结果如图：

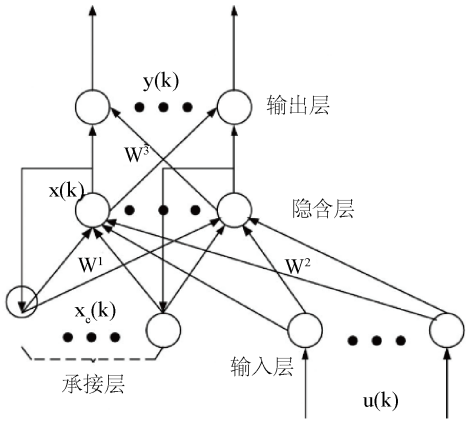


3.3. 模型对比

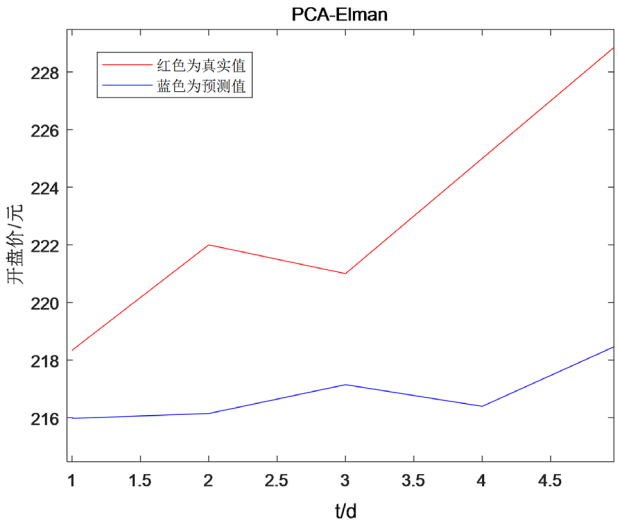
同样，若不对数据进行主成分分析，在数据标准化之后直接在 LSTM 神经网络里进行训练，选取参数不变的情况下，得到如图所示的预测效果：



Elman 网络包含输入层、中间层(隐藏层)、承接层和输出层。Elman 对比静态的 BP 网络，由于它增加了一个承接层当作一种延时算子，具有记忆的功能[3]，记忆隐含层前一时刻的输出，对历史的状态比较敏感，由于不断更新网络状态，从而增加了网络处理动态信息的能力。Elman 网络结构如图：



将数据用 PCA-Elman 神经网络训练，结果如下：



我们预测了五粮液 2020 年 9 月 28 日到 10 月 12 日共 5 日的开盘价格，将三种模型的预测值放到如下表格：

	日期	真实值	PCA-LSTM	LSTM	PCA-Elman
	2020.09.28	218.3400	213.8827	213.4512	215.9741
开	2020.09.29	217.0100	217.2447	215.9735	216.1446
盘	2020.09.30	221.0000	216.8814	215.4484	217.1460
价	2020.10.09	225.0000	220.2482	216.8713	216.3934
	2020.10.12	229.0000	226.5597	223.8607	218.5380
RMSE			0.0875	0.1290	0.1470

从表格中我们也可以知道 PCA-LSTM 模型的拟合效果最好，较比于 LSTM 和 PCA-Elman 模型误差也是最小的，对五粮液股票数据的开盘价格预测效果最佳。

4. 总结

本文建立了 PCA-LSTM、PCA-Elman 和 LSTM 神经网络模型，LSTM 是解决长序依赖问题的有效技术。对比试验发现 PCA-LSTM 对五粮液股票价格的预测精度更高，对于数据波动较大的开盘价格有较好的预测。由于长短记忆网络的适用性比较广，导致模型的变化程度也比较高。后期考虑加入对股票影响大的新闻、市场指数等特征来训练模型，希望能提高模型对股价预测的精准度，给股民的选择带来更有价值的参考[4]。

基金项目

辽宁省自然科学基金指导计划项目(编号：2019-ZD-0471)。

参考文献

[1] 文宝石, 颜七笙. 数据多维处理 LSTM 股票价格预测模型[J]. 江西科学, 2020, 38(4): 443-449+472.

[2] 王悦霖, 徐野. 基于 LSTM 的多元股票信息特征提取与预测研究[J]. 现代经济信息, 2019(3): 325.

[3] 宋明达, 赵宇红. 基于 Elman 神经网络在电力负荷预测中的研究[J]. 科技风, 2020(11): 200-201.

[4] 彭燕, 刘宇红, 张荣芬. 基于 LSTM 的股票价格预测建模与分析[J]. 计算机工程与应用, 2019, 55(11): 209-212.