**Advances in Applied Mathematics** 应用数学进展, 2020, 9(11), 1893-1900
Published Online November 2020 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2020.911218

Hans汉斯

# B样条与LSTM结合的时间序列建模方法研究

## ——以股票收盘价数据为例

**李　杨，王艺舒***

青岛大学，山东　青岛
Email: 291814668@qq.com, *yishu6661@126.com

收稿日期：2020年10月16日；录用日期：2020年11月5日；发布日期：2020年11月12日

## 摘　要

**时间序列是指将某种统计指标的数值按照时间顺序排列所形成的序列，对时间序列进行分析预测有助于我们提前判断、减少潜在风险，在生产和生活中都有着重大的意义。常见的时间序列数据包括股票数据、商品日销量以及每日的气候等等。其中股票数据是一个高度复杂的非线性系统，对股票长期趋势的预测一直是一个令人感兴趣的话题。本文以股票收盘价数据为例，提出一种新的思路，对时间序列数据进行建模并预测。首先采用B样条回归对股票价格与时间变量之间的非线性关系进行建模。由于B样条曲线只能在一个提前给定的区间内进行拟合，对区间外的预测问题无能为力，因此，为解决这个问题，本文将B样条回归与长短期记忆(Long Short-Term Memory)神经网络模型相结合，构建LSTM-Bspline模型，通过训练神经网络来获取B样条曲线的参数，从而获得预测值。最后通过对某公司股票每日收盘价数据进行分析预测，并将预测结果与经典LSTM神经网络预测结果相比较，证明了LSTM-Spline模型的可行性。**

## 关键词

**人工神经网络，LSTM，B样条，股价预测**

# Study on Time Series Modeling Method of B-Spline Combined with LSTM

## —Take the Closing Share Price as an Example

**Yang Li, Yishu Wang***

Qingdao University, Qingdao Shandong
Email: 291814668@qq.com, *yishu6661@126.com

Received: Oct. 16th, 2020; accepted: Nov. 5th, 2020; published: Nov. 12th, 2020

*通讯作者。

文章引用: 李杨, 王艺舒. B 样条与 LSTM 结合的时间序列建模方法研究[J]. 应用数学进展, 2020, 9(11): 1893-1900.
DOI: 10.12677/aam.2020.911218

## Abstract

**Time series refers to the series formed by arranging the values of certain statistical indicators in chronological order. Analysis and prediction of time series can help us to judge in advance and reduce potential risks, which is of great significance in both production and life. Common time series data include stock data, daily sales of goods, daily weather, and so on. Among them, stock data is a highly complex nonlinear system, and the prediction of stock long-term trend is always an interesting topic. Taking stock closing price data as an example, this paper proposes a new way of thinking to model and predict time series data. First the B-spline regression is adopted to model the nonlinear relationship between stock price and time variable, the B-spline curve can only be a in advance within a given interval fitting, outside the range of powerless forecasting problems, therefore, to solve this problem, this paper combined the B spline regression with Long Short-Term Memory neural network model, build LSTM-Bspline model, by training neural network to obtain the parameters of the B-spline curve, the predicted value is achieved. Finally, by analyzing and forecasting the daily closing price data of a company's stock, and comparing the forecasting results with those of classical LSTM neural network, the feasibility of LSTM-Spline model is proved.**

## Keywords

**Artificial Neural Network, LSTM, B-Spline, Stock Price Forecast**



## 1. 引言

时间序列是指将某种统计指标的数值按照时间顺序排列所形成的序列，如股票收盘价、商品日销量以及每日的气候等等。而时间序列预测法就是通过处理和分析时间序列，提取出时间序列所蕴含的数据特征、发展方向和趋势等信息，进行类推或延伸，借以预测下一时间段内或以后若干时间段内数据可能达到的水平[1]。

1927 年数学家耶尔为了预测市场变化的规律，提出的自回归(Autoregressive)模型，首次提出了系统的时间序列分析方法。紧接着，在自回归(AR)模型的基础上，1931 年 Walker，建立了滑动平均(Moving Average)模型[2]。20 世纪 70 年代 G. P. Box 和 G. M. Jenkins 发表了著作《Time Series Analysis: Forecasting and Control》[3]，在书中对于平稳的时间序列数据提出了和自回归滑动平均(Autoregressive Moving Average)模型，并且建立了一套完整的建模、估计、检验和控制的方法。随后为了应对非平稳时间序列，又出现了更为完善的差分整合滑动平均自回归(Autoregressive Integrated Moving Average)模型。

上述传统时间序列模型在大量不间断数据的基础上，拥有较高的预测精度，但是需要复杂的参数估计，而且计算出的参数不能移植[4]。

非参数回归是一种适合不确定的、非线性的动态系统的建模方法。它包括局部回归、光滑样条和正交回归等。非参数回归不需先验知识，只需足够的历史数据，寻找历史数据中与当前点相似的“近邻”，并用这些“近邻”进行预测。1991 年，Davis 和 Nihan [5]将非参数回归应用在了时间序列分析中。

20 世纪 40 年代，诞生了一种新的方法：人工神经网络。1987 年 Lapedes [6]等人最早将神经网络应用于时间序列预测，但只针对仿真数据进行了研究。1988 年 Werbos [7]，1990 年 Varfis 和 Versino [8]分别对真实时间序列数据进行了预测研究。1990 年 Weigend [9]等人针对太阳黑子的年平均活动数据，将神经网络与回归方法作了对比，得出神经网络预测优于统计预测。1991 年，Matsuba [10]等人将神经网络应用到股票预测。

循环神经网络(Recurrent Neural Network)是一类专门用于处理序列数据的神经网络，与传统神经网络相比，循环神经网络具有记忆功能，可以考虑历史序列的信息，更加注重挖掘样本之间的时序关联[10]。循环神经网络可以处理任意长度的序列数据，因此，在视频[11]、语音[12] [13]和文本[14]等序列数据中都有着广泛的应用。

但是循环神经网络也有局限性，由于其自身结构的复杂，导致其存在梯度消失的问题，当输入序列数据过长时，距离某时刻较远的序列信息会被弱化甚至忽略，因此循环神经网络无法“记忆”距当前时刻较远但重要的信息[15]。为了解决这个问题，2012 年，Graves A [16]提出了长短期记忆(Long Short-Term Memory)神经网络，通过引入线性连接和门控单元来解决梯度消失问题。

股票走势的预测问题一直是金融市场感兴趣的问题之一，一方面股票价格直接影响投资者的经济利益，另一方面股票价格也反映着国家的经济状况以及不同行业的景气情况。因此，建立适当的模型对股票价格进行预测，对于投资者把控投资方向以及国家经济政策的调整有着重大的研究意义。

本文将非参数回归中的 B 样条方法和当前流行的长短期记忆神经网络结合，构建了 LSTM-Bspline 模型以一种新的思路来对某公司股票收盘价数据进行分析并预测，并将预测结果与经典的长短期记忆神经网络相比较，证明方法的适用性。本文其余内容如下：第 2 部分，我们介绍 B 样条方法和长短期记忆神经网络的相关概念与定义。第 3 部分，将 B 样条与长短期记忆神经网络相结合，构建出 LSTM-Bspline 模型。第 4 部分以某公司股票的每日收盘价数据为例，分别使用 LSTM-Bspline 模型和 LSTM 模型进行预测，并将结果进行比较，证明本文所提出方法的有效性。第五部分对本文的研究进行总结，指出模型的不足和未来的研究方向。

## 2. 研究方法

### 2.1. 多项式样条近似

#### 2.1.1. 定义

在不失一般性的前提下，假设函数 $f(x)$ 定义域为[0, 1]，多项式样条是指在一组内部节点上平滑连接的分段多项式，本文采用多项式样条中的 B 样条，因为它具有稳定的数值性质。

对于 B 样条，假设把定义域[0, 1]分成 $M$ 段，内节点为 $0<j_1<\cdots<j_{M-1}<1$。则对应的 B 样条基函数可以表示为 $B_1(x),\cdots,B_{M+d+1}(x)$，B 样条基的定义由 de Boor [17]导出：

$$B_{i,0}(u)=\begin{cases}1, & j_i\le u\le j_{i+1}\\ 0, & \text{other}\end{cases}$$

$$B_{i,d}(u)=\frac{u-j_i}{j_{i+d}-j_i}B_{i,d+1}(u)+\frac{j_{i+d+1}-u}{j_i-j_{i+1}}B_{i+1,d-1}(u),d>0$$

约定 $0/0=0$，其中 $d$ 是 B 样条的次数，$j_i$ 为节点。每个 B 样条基函数都具有局部支撑性，并且对于任意相邻的节点 $j_{i-1},j_i\,(1\le i\le M+1)$，除了基函数 $B_j(x),\cdots,B_{j+d+1}(x)$ 外，其余基函数在区间 $[j_{i-1},j_i]$ 都为零。

设 G 是 B 样条基函数 $\{B_j(x),j=1,\cdots,M+d+1\}$ 在[0,1]上张成的线性空间。假设 $f(x)$ 可以被 G 中的

元素近似，则：

$$f(x) \approx \sum_{j=1}^{M+d+1} \gamma_j B_{j,d}(x) \tag{2.1}$$

### 2.1.2. 基于 B 样条的最小二乘拟合

在上述(2.1)式中的参数可以通过最小二乘准则估计出，即最小化下式：

$$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{M+d+1}\gamma_j B_{j,d}(X_i)\right)^2 = \frac{1}{n}\|Y - B\gamma\|_2^2 \tag{2.2}$$

其中 $\gamma = (\gamma_1, \cdots, \gamma_{M+d+1})^{\mathrm{T}}$ 表示系数向量，B 表示矩阵 $(B(X_1), \cdots, B(X_n))^{\mathrm{T}}$，其中 $B(x) = (B_1(x), \cdots, B_{M+d+1}(x))^{\mathrm{T}}$。$Y = (Y_1, \cdots, Y_n)^{\mathrm{T}}$ 表示响应变量。如果直接通过(2.2)式对时间序列数据进行拟合，往往难以取得较好的效果，原因在于对于不同的时间点 $t$，如果使用一组统一的系数 $\gamma = (\gamma_1, \cdots, \gamma_{M+d+1})^{\mathrm{T}}$，会使得总体误差比较大。

## 2.2. LSTM 神经网络模型

LSTM (长短期记忆神经网络)是一类用于捕获时序数据中长期和短期依赖关系的 RNN (Recurrent Neural Network)模型，近年来，在语音识别，机器翻译等领域取得了巨大的成功。从数学上来讲，LSTM 是一个高度复合的非线性参数函数，它将一列向量 $(x_1, \cdots, x_n)$ 通过隐含层 $(h_1, \cdots, h_n)$ 映射到另一组向量 $(y_1, \cdots, y_n)$。LSTM 神经网络内部结构如图 1：

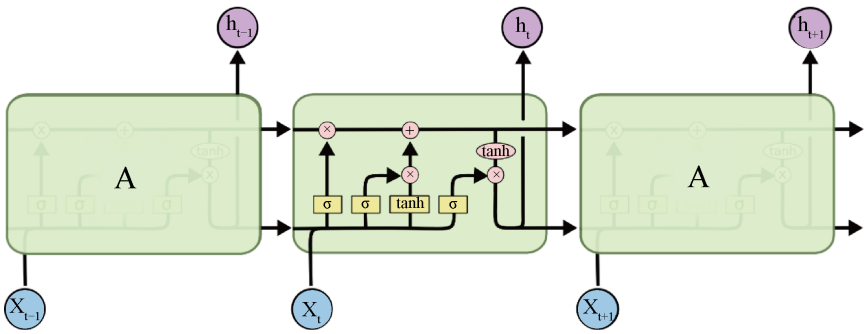


**Figure 1.** Internal structure of LSTM neural network
**图 1.** LSTM 神经网络内部结构

LSTM 神经网络由相互联系的递归子网络，即记忆模块组成，记忆模块主要包括三个门：遗忘门，输入门，输出门，和一个记忆单元。

RNN 模型是 LSTM 模型的底层结构，函数形式为：

$$h_j = \sigma_h(W_h x_j + U_h h_{j-1} + b_h)$$
$$y_j = \sigma_y(W_y h_j + b_y)$$

其中 $W_h, U_h, b_h, W_y, b_y$ 表示参数，$\sigma_h, \sigma_y$ 是非线性激活函数，上式模拟了 $x_1, \cdots, x_n$ 与 $y_1, \cdots, y_n$ 之间的非线性函数关系。可以通过将这种结构进行多次叠加，得到多层或分层 RNN 模型。

LSTM 是 RNN 结构的扩展，具体函数形式为：

$$f_j = \sigma_g\left(W_f x_j + U_f h_{j-1} + b_f\right),$$

$$i_j = \sigma_g\left(W_i x_j + U_i h_{j-1} + b_i\right),$$

$$o_j = \sigma_g\left(W_o x_j + U_o h_{j-1} + b_o\right),$$

$$g_j = \sigma_h\left(W_g x_j + U_g h_{j-1} + b_g\right),$$

$$c_j = f_j * c_{j-1} + i_j * g_j,$$

$$h_j = o_j * \sigma_h\left(c_j\right),$$

最后输出值 $y_j$ 是一个关于 $h_j$ 非线性函数，即：

$$y_j = \sigma_h\left(W_h h_j + b_y\right)$$

所有的 $W$，$U$，$b$ 代表参数，$\sigma_g, \sigma_h$ 代表非线性激活函数，本文中分别采用了 logistic 函数和 tanh 函数。Logistic 函数和 tanh 函数是神经网络中最常用的两个激活函数，这些激活函数的多个组合可以逼近输入向量和输出向量间复杂的非线性关系。

## 3. 提出模型

对于一组时间序列数据 $\{y_1, y_2, \cdots, y_n\}$，将其对应的时间变量设为 $\{1, 2, \cdots, n\}$，假设对未来 $k$ 个时间点的值 $\{y_{n+1}, y_{n+2}, \cdots, y_{n+k}\}$ 进行预测，即预测时间点 $\{n+1, n+2, \cdots, n+k\}$ 对应的值。

首先将区间 $[n+1, n+k]$ 分割成 $M$ 个子区间，其中节点为 $\{n+1 \le j_1 \le \cdots \le j_{M-1} \le n+k\}$，则对应的 B 样条基函数可以表示为 $B_1(x), \cdots, B_{M+d+1}(x)$，$d$ 为 B 样条的次数，则根据(2.1)我们可以建立如下函数关系

$$y_t = f_t = \sum_{j=1}^{M+d+1} \gamma_{t,j} B_{j,d}(t),\ t \in \{n+1, \cdots, n+k\}$$

其中 $\gamma_{t,1}, \cdots, \gamma_{t,M+d+1}$ 为时变参数，表示 $t$ 时间 B 样条基的系数。

而对于数据 $\{y_1, y_2, \cdots, y_n\}$，我们将其以 $k$ 个值为一组等分，共得到 $n/k$ 组数据，分别设为第 1，2，…，$n/k$ 组，对每组数据采用 B 样条拟合，其中每组数据的节点间隔同区间 $[n+1, n+k]$ 的节点间隔相同，并且每组数据对应一组 B 样条基函数，则对于第 $i$ 组数据，可以得到以下函数关系：

$$y_t = f_t = \sum_{j=1}^{M+d+1} \gamma_{t,j} B_{i,j,d}(t),\ t \in \{k(i-1)+1, \cdots, ik\}$$

其中 $B_{i,j,d(t)}\ (j = 1, \cdots, M+d+1)$ 为第 $i$ 组 B 样条基函数。

通过将上式子整合，在区间 $[1, n+k]$ 上，可以用一个方程来表示 $y_t$ 与 $t$ 的关系：

设将 $y_t$ 所在的组为第 $i_t$ 组，则：

$$y_t = f_t = \sum_{j=1}^{M+d+1} \gamma_{t,j} B_{i_t,j,d}(t),\ t \in \{1, 2, \cdots, n+k\}$$

其中 $B_{i_t,j,d}(t)$ 为 $y_t$ 所在的组对应的 B 样条基函数，而 $\gamma_{t,1}, \cdots, \gamma_{t,M+d+1}$ 仍为时变参数，表示 $t$ 时间 B 样条基的系数。

下面要解决的关键问题就是时变参数 $\gamma_{t,1}, \cdots, \gamma_{t,M+d+1}$ 的获取问题，显然参数 $\gamma_{t,1}, \cdots, \gamma_{t,M+d+1}$ 依赖于过去的时间序列数据 $y_{t-1}, y_{t-2}, \cdots$，为了对这种依赖关系建模，我们从 $y_{t-1}, y_{t-2}, \cdots$ 中选取了一段固定长度的子序列来构造一个特征向量序列。假设选取的长度为 $L$，可以构造出以下长度为 $L$ 的特征向量序列：

$$x_1^t,\cdots,x_L^t=\begin{bmatrix}y_{t-L}\\(y_{t-L}-\overline{y_t})^2\\(y_{t-L}-\overline{y_t})^3\\(y_{t-L}-\overline{y_t})^4\end{bmatrix},\cdots,\begin{bmatrix}y_{t-L}\\(y_{t-L}-\overline{y_t})^2\\(y_{t-L}-\overline{y_t})^3\\(y_{t-L}-\overline{y_t})^4\end{bmatrix},\ \overline{y_t}=\frac{1}{L}\sum_{i=1}^{L}y_{t-i}$$

构造的依据是提取与过去 $L$ 个样本的第一阶，二阶，三阶，四阶中心矩有关的信息。在这种构造下，我们将 $x_1^t,\cdots,x_L^t$ 作为 LSTM 单元的输入，将 $\gamma_{t,1},\cdots,\gamma_{t,M+d+1}$ 作为 LSTM 单元的输出：

$$\left[\gamma_{t,1},\cdots,\gamma_{t,M+d+1}\right]^{\mathrm{T}}=\tanh\left(W^o h_t+b^o\right),\quad h_t=\mathrm{LSTM}_\Theta\left(x_1^t,\cdots,x_L^t\right)$$

其中 $\Theta$ 表示 LSTM 参数，$h_t$ 代表最后的隐藏状态，$W^o,b^o$ 代表输出层参数，上述一系列参数可以通过最小二乘准则获得，即：

$$\min_{\Theta,W^o,b^o}\frac{1}{n}\sum_{t=1}^{n}\left(Y_t-\sum_{j=1}^{M+d+1}\gamma_{t,j}B_{it,j,d}\left(t\right)\right)^2$$

经过对模型的训练，对于要预测的序列 $\{y_{n+1},y_{n+2},\cdots,y_{n+k}\}$ 任意 $y_t\left(t=n+1,\cdots,n+k\right)$，B 样条的时变参数 $\gamma_{t,1},\cdots,\gamma_{t,M+d+1}$ 可以通过用训练获得的模型参数 $\hat{\Theta},\hat{W}^o,\hat{b}^o$ 计算出来，从而获得预测值。

## 4. 实例分析

本文选取了某公司股票 1750 个数据进行研究，将其中五分之三的数据作为训练集，五分之一作为验证集集，五分之一作为测试集。分别用 LSTM 模型与 LSTM-Spline 模型进行预测，比较分析预测结果。

### 4.1. 数据处理

由于我们采取的 LSTM 激活函数是 tanh，输出结果为−1 到 1 之间，所以对数据进行归一化处理，同时归一化处理可以提升网络学习的速度，加快收敛。具体方法如下：

$$y=\frac{y_{\max}-y}{y_{\max}-y_{\min}}$$

### 4.2. 模型超参数选择

本文提出的模型超参数有三个：1，用于预测下个时间点的前一段时间序列的长度 $L$；2，LSTM 神经网络隐含层的维数 $H$；3，定义 B 样条基函数时的节点选取，为了计算方便，我们统一选择等距节点，节点之间的间距为 $D$。则模型可以表达为 LSTM-Spline ($H$, $L$, $D$)。

超参数的调优在以下集合中进行：$L\in\{40,60,80,100\}$，$H\in\{8,16\}$，$D\in\{k/2,k/5,k/10\}$，其中 $k$ 表示测试集的长度。

### 4.3. 结果展示

在尝试了不同的超参数组合后，选取效果最好的一组超参数，运行的结果如图 2，图中纵轴数值为归一后的值。左图为 LSTM-Spline 模型的结果，右图为 LSTM 模型的结果。其中橙色曲线代表归一化后实际值，蓝色曲线代表归一化后的预测值。

### 4.4. 结果分析

通过观察可以发现，LSTM-Spline 程序的结果可以大致预测出未来一段时间的股票的走势，虽然具

体数值有些偏差，但是趋势正确。与传统 LSTM 模型结果相比误差更小。通过计算两模型结果(归一化后)的 MSE (均方误差)，结果如表 1，证明了我们的观点。

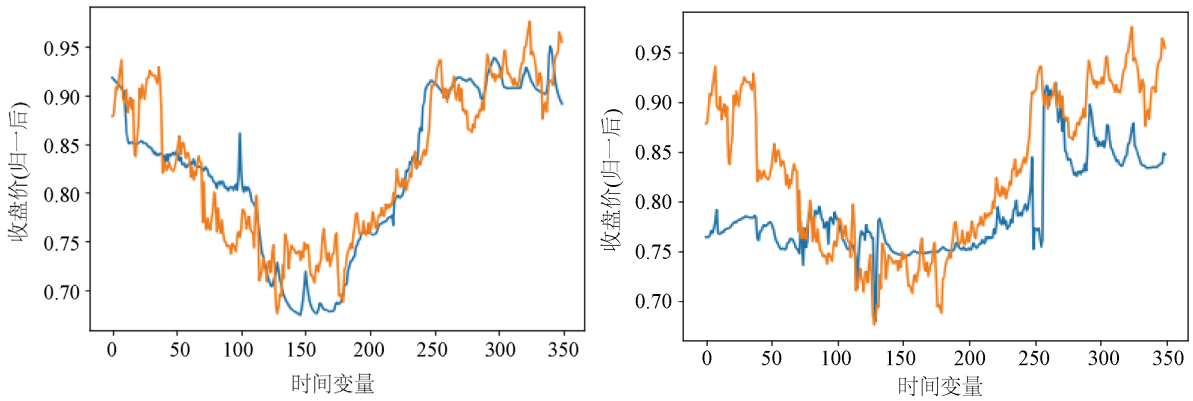


**Figure 2.** LSTM-Spline model and LSTM model running results
**图 2.** LSTM-Spline 模型与 LSTM 模型运行结果

**Table 1.** Comparison of mean square error between results of two models
**表 1.** 两模型结果均方误差对比

| 模型 | 均方误差 |
|---|---|
| LSTM-Spline | 0.0013 |
| LSTM | 0.0044 |

## 5. 总结

本文提出了一种新的思路来对股票价格进行研究，将非参数回归中常用的 B 样条与当前流行的 LSTM 神经网络结合，得到了不错的结果。但只能预测总体变化趋势，与真实值仍存在误差，股票市场中，大盘之上瞬息万变，股价的影响因子众多，因此，为了进一步提升预测效果，仍有进一步的工作要做：1. 本文只考虑了股价过去一段时间价格对自身的影响，没有加入其他变量，而股价的影响因子众多，未来应尝试加入其他变量进行预测；2. 本文中超参数的选择是在一组参数中不断尝试选取的，不一定是最优解，未来应考虑选择更为科学的办法选择超参数，提升预测效果。

## 参考文献

[1] 张美英, 何杰. 时间序列预测模型研究综述[J]. 数学的实践与认识, 2011, 41(18): 189-195.

[2] 王洋. 基于时间序列分析的 IP 语音收入预测[D]: [硕士学位论文]. 长春: 吉林大学, 2004.

[3] Box, G.E.P., Jenkins, G.M., Reinsel, G.C., *et al.* (2015) Time Series Analysis: Forecasting and Control. John Wiley & Sons, Hoboken.

[4] 刘静, 关伟. 交通流预测方法综述[J]. 公路交通科技, 2004, 21(3): 82-85.

[5] Davis, G.A. and Nihan, N.L. (1991) Nonparametric Regression and Short-Term Freeway Traffic Forecasting. *Journal of Transportation Engineering*, **117**, 178-188. https://doi.org/10.1061/(ASCE)0733-947X(1991)117:2(178)

[6] Lapedes, A. and Farber, R. (1987) Nonlinear Signal Processing Using Neural Networks: Prediction and System Modelling. *IEEE International Conference on Neural Networks*, San Diego, CA, 21 June 1987.

[7] Werbos, P.J. (1988) Generalization of Backpropagation with Application to a Recurrent Gas Market Model. *Neural networks*, **1**, 339-356. https://doi.org/10.1016/0893-6080(88)90007-X

[8] Varfis, A. and Versino, C. (1990) Univariate Economic Time Series Forecasting by Connectionist Methods. *Proceed-*

*ings of INNC*, **90**, 342-345.

[9] Weigend, A.S., Huberman, B.A. and Rumelhart, D.E. (1990) Predicting the Future: A Connectionist Approach. *International Journal of Neural Systems*, **1**, 193-209. https://doi.org/10.1142/S0129065790000102

[10] Matsuba, I. (1991) Application of Neural Sequential Associator to Long-Term Stock Price Prediction. *Proceedings of the* 1991 *IEEE International Joint Conference on Neural Networks*, Singapore, 18-21 November 1991, 1196-1201. https://doi.org/10.1109/IJCNN.1991.170559

[11] 张斌, 魏维, 高联欣, 等. 基于时空域深度神经网络的野火视频烟雾检测[J]. 计算机应用与软件, 2019, 36(9): 236-242.

[12] 冯天艺, 杨震. 采用多任务学习和循环神经网络的语音情感识别算法[J]. 信号处理, 2019, 35(7): 1133-1140.

[13] 卢艳. 基于神经网络与注意力机制结合的语音情感识别研究[D]: [硕士学位论文]. 北京: 北京邮电大学, 2019.

[14] 唐贤伦, 林文星, 杜一铭, 等. 基于串并行卷积门阀循环神经网络的短文本特征提取与分类[J]. 工程科学与技术, 2019, 51(4): 125-132.

[15] 鞠春雷, 聂方超, 刘文岗, 郭金山, 张江石. 基于长短期记忆网络的矿工不安全行为研究[J]. 煤矿安全, 2020, 51(9): 260-264.

[16] Graves, A. (2012) Supervised Sequence Labelling. In: *Supervised Sequence Labelling with Recurrent Neural Networks*, Springer, Berlin, Heidelberg, 5-13. https://doi.org/10.1007/978-3-642-24797-2_2

[17] De Boor, C. (1978) A Practical Guide to Splines. In: *Applied Mathematical Sciences*, Vol. 27, Springer-Verlag, New York, 157.