# 一种基于 LSTM 模型的日销售额预测方法

吴娟娟,任 帅,张卫钢,伍 菁,李香云
(长安大学 信息工程学院,陕西 西安 710064)

**摘 要**:精准的销售额预测对于商业运营有非常大的指导意义,可以指导运营后台提前进行合理的资源配置,帮助管理者制定合理的目标。零售商店日销售额预测指从商店已有日销售额的数据资料中总结出商品销售额的变化规律,并根据该规律动态预测未来一段时间内的日销售额。预测目的是通过增加企业销量,从而完善生产模式,使企业获利。目前,现有的关于商品销售额预测方法的精度大都不高,低于85%。因此,提出了一种基于 TensorFlow 的 LSTM 模型的零售商店日销售额预测方法,能够提高预测未来一周的日销售额精度。实验结果显示,预测精度达到 90%;同时得到 LSTM 模型的 MAPE 为 0.031 932,MAE 为 168.320 7,明显高于现有模型的预测结果。

**关键词**:日销售额;预测;TensorFlow;LSTM
**中图分类号**:TP181 **文献标识码**:A **文章编号**:1673-629X(2020)02-0133-05
doi:10.3969/j.issn.1673-629X.2020.02.026

# A Daily Sales Forecasting Method Based on LSTM Model

WU Juan-juan, REN Shuai, ZHANG Wei-gang, WU Jing, LI Xiang-yun
(School of Information Engineering, Chang'an University, Xi'an 710064, China)

**Abstract**: Accurate sales forecasting has great guiding significance for business operations. It can guide the operational back-office to allocate reasonable resource in advance and help managers to set reasonable goals. The retail store daily sales forecast refers to the change rule of the sales summarized from the data of existing daily sales of stores and the dynamic prediction of daily sales in a period of time in the future. The forecasting is to improve the production model and make profits by increasing the production volume of the company. At present, the accuracy of the existing forecasting methods on the sales of goods is not high, less than 85%. Therefore, we propose a retail store daily sales forecasting method based on LSTM model of TensorFlow, which can improve the precision of daily sales forecasting for the next week. The experiment shows that the prediction accuracy reaches 90%. At the same time, the MAPE of the LSTM model is 0.031 932, and the MAE is 168.320 7, which is significantly higher than the prediction results of the existing models.

**Key words**: daily sales; forecast; TensorFlow; LSTM

## 0 引 言

零售商店日销售额预测是从已有的日销售额资料中总结出商品的日销售额规律,并且用此规律动态预测未来一段时间内的日销售额,从而指导未来的销售方法或手段,提高销量,获取更大利润。日销售额的分析是一个复杂且不规则的非线性系统。重点分析并且找到影响商品销售额的一些主要因素。但由于新产品的开发、季节性变换、促销活动、节假日、天气及政策的变化等各种原因,传统分析方法的准确性在一定程度上受到很大质疑。因此,希望建立一个非线性系统,其参数可以随预测环境的变化而变化,从而克服该缺点。

为了得到更精准的预测结果,许多学者根据市场的调研和研究,运用多种方法针对已有的数据进行分析和对比[1]。比如:Xgboost 单模型[2]、指数平滑法、ARIMA 模型和 GARCH 模型等[3-4]。但是 Xgboost 单模型采用预排序,在迭代之前,对节点的特征做预排序,遍历选择最优分割点,数据量大时,贪心法耗时,数据分割的复杂度高;指数平滑法精确率不高;ARIMA 模型对于趋势性较强的数据集预测效果比较好,但如果遇到趋势不那么强的数据集,则效果不太理想;对于非对称现象,GARCH 模型无法解释该现象,为了保证其非负性,假设模型表达式中的全部系数都大于零,这

**收稿日期**:2019-03-28 **修回日期**:2019-07-30 **网络出版时间**:2019-11-07
**基金项目**:国家自然科学基金(61702050)
**作者简介**:吴娟娟(1994-),女,硕士研究生,研究方向为机器学习、深度学习、数据挖掘;任 帅,副教授,硕导,研究方向为信息隐藏理论与模型;张卫钢,博士,教授,研究方向为计算机应用技术、智能测控技术、汽车电子技术。
**网络出版地址**:http://kns.cnki.net/kcms/detail/61.1450.TP.20191107.0908.010.html

些约束中所隐含的任何一个滞后项的增大将会增加因而排除了的随机波动行为,导致在估计该模型时有可能会出现震荡现象;LSTM 模型[5]避免了长期依赖问题,采用特殊隐式单元,在继承了大部分 RNN 模型特性的同时解决了梯度反传过程中[6]由于逐步缩减而产生的 Vanishing Gradient 问题[7],适用于非线性回归变量,可以解决多个输入变量的问题,模型准确度高,训练速度快,并行处理能力强。LSTM 更适合用于处理与短期时间序列高度相关的问题,在 $n$ 个示例批次中不断迭代,能够快速和准确地对大量短期时间序列数据进行处理,是解决时间序列预测问题最常用的工具。

针对大数量级的序列预测,文中建立了一种 Tensorflow 框架下基于记忆机理的 LSTM 模型。以预测值和实际值的误差为优化目标,从网络结构的搭建,学习率、窗口设置上改进网络模型预测的准确性,采用 RMSProps 算法修正模型自适应率。最后,应用销售额数据进行验证,并与传统时间序列预测模型进行对比,实验结果表明建立的 LSTM 模型在销售额预测上具有良好的优越性。

## 1 LSTM 神经网络

LSTM 是一种改进的 RNN,比一般的 RNN 能够记住更长周期上的信息模式,在解决很多问题上都取得了成功,例如自然语言处理、中文文本分类研究、机器翻译等。

目前为止,实际应用中最有效的序列模型为门控 RNN(gate RNN),包括基于长短期记忆网络(long short - term memory)和基于门控循环单元(gate recurrent unit)的网络。LSTM 网络比一般的 RNN 结构更适于长期依赖,在序列处理问题上获得很好的表现[8]。LSTM 结构如图 1 所示。

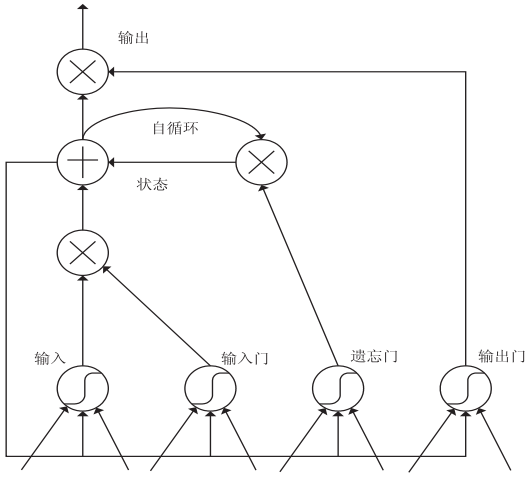


图 1 LSTM 结构

由图可知,LSTM 神经网络除了 RNN 式循环之外,还在自身“cell”(取代一般循环神经网络的隐藏单元)内部循环。每一个单元相输入输出,但也有更多的参数和控制信息流动的门控单元系统,即状态单元 $s_i^{(t)}$(时刻 $t$,细胞 $i$),外部输入门(external input gate)、遗忘门(forget gate)、输出门(output gate)。遗忘门负责从“cell”移除 LSTM 学习不重要的信息,这些信息将通过门控单元运算移除。遗忘门采取两个输入 $h_{t-1}$ 和 $x_t$。$h_{t-1}$ 是前一单元的隐藏态或输出态,$x_t$ 是特定时间步输入,为输入序列 $x$ 的第 $t$ 个元素。将给定输入向量和权重矩阵相乘,然后增加偏置项用以输入 Sigmoid 函数。函数 Sigmoid 会输出一个向量,取值的范围是 [0,1],依次对应于单元状态的每个数值。Sigmoid 函数基本可以决定保留哪些值并且忘记哪些值,如果单元状态取特定值零,遗忘门会要求单元状态将该信息完全忘记。最后输出的 Sigmoid 函数向量与单元状态相乘。相应的前向传播算法如下:

Input gates:

$$a_\ell^t = \sigma\left(\sum_{i=1}^{I}\omega_{i\ell}x_i^t + \sum_{h=1}^{H}\omega_{h\ell}b_h^{t-1} + \sum_{c=1}^{C}\omega_{c\ell}s_c^{t-1}\right) \tag{1}$$

Forget gates:

$$a_\varphi^t = \sigma\left(\sum_{i=1}^{I}\omega_{i\varphi}x_i^t + \sum_{h=1}^{H}\omega_{h\varphi}b_h^{t-1} + \sum_{c=1}^{C}\omega_{c\varphi}s_c^{t-1}\right) \tag{2}$$

Cells:

$$a_c^t = \sum_{i=1}^{I}\omega_{ic}x_i^t + \sum_{h=1}^{H}\omega_{hc}b_h^{t-1} \tag{3}$$

$$s_c^t = a_\varphi^t s_c^{t-1} + b_\ell^t g(a_c^t) \tag{4}$$

Output gates:

$$a_\omega^t = \sigma\left(\sum_{i=1}^{I}\omega_{iw}x_i^t + \sum_{h=1}^{H}\omega_{hw}b_h^{t-1} + \sum_{c=1}^{C}\omega_{cw}s_c^t\right) \tag{5}$$

Cell outputs:

$$b_c^t = a_\omega^t h(s_c^t) \tag{6}$$

其中,$\boldsymbol{x}_i^t$ 是当前的输入向量,下标 $\ell,\varphi,c,\omega$ 表示输入门、遗忘门、细胞状态单元、输出门的相关参数。带下标 $h$ 的是泛指,由于 LSTM 的一个重要特点是具有灵活性,cell 之间能够互联,hidden units 之间也能够互联,下标 $h$ 即泛指那些连进来的东西。$\sigma$ 是 sigmoid 函数。反向传播计算如式(7)~式(13)所示:

$$\zeta_c^t \overset{\text{def}}{=} \frac{\partial\zeta}{\partial b_c^t},\ \zeta_s^t = \frac{\partial\zeta}{\partial s_c^t} \tag{7}$$

Cell outputs:

$$\zeta_c^t = \sum_{k=1}^{K}\omega_{ck}\delta_k^t + \sum_{g=1}^{G}\omega_{cg}\delta_g^{t+1} \tag{8}$$

Output gates:

$$\delta_\omega^t = f'(a_\omega^t)\sum_{c=1}^{C}h(s_c^t)\zeta_c^t \tag{9}$$

States:

$$\zeta_s^t = b_\omega^t h'(s_c^t\zeta_c^t) + b_\varphi^{t+1}\zeta_s^{t+1} + \omega_{c\ell}\delta_\ell^{t+1} +$$

$$\omega_{c\varphi}\delta_{\varphi}^{t+1} + \omega_{c\omega}\delta_{\omega}^{t} \tag{10}$$

Cells:

$$\delta_c^t = b_\ell^t g'(a_c^t)\zeta_s^t \tag{11}$$

Forget gates:

$$\delta_\varphi^t = f'(a_\varphi^t)\sum_{c=1}^{C} h(s_c^{t-1})\zeta_s^t \tag{12}$$

Input gates:

$$\delta_\ell^t = f'(a_\ell^t)\sum_{c=1}^{C} g(a_c^t)\zeta_s^t \tag{13}$$

式(8)中 $g$ 和前向传播中的 $h$ 含义相同,即泛指,由于它不一定只能输出到下一个时间的自身,或许还会输出到下一个时间其他的隐层。

## 2 LSTM 优化算法

大多数深度学习算法通常以最小化代价函数、损失函数(loss)或误差函数 $J(\theta)$ 作为优化目标。$J(\theta)$ 用来描述预测值和真实值的偏离程度,文中指的是均方根值。优化可以分成两个阶段,第一阶段是使用前向传播算法计算得到预测值,再将预测值与真实值比较得到它们之间的差距;第二阶段使用反向传播算法,计算出损失函数对所有参数的梯度,然后按照梯度和学习率(learning rate)利用梯度下降法(gradient decent)更新每一个参数。假设 $\theta$ 表示神经网络参数,优化过程可看作迭代更新并寻找一个参数 $\theta$,使得 $J(\theta)$ 最小。

梯度下降的计算过程指沿着梯度下降的方向求解极小值,迭代公式为:

$$x^* = \mathrm{argmin} f(x) \tag{14}$$

梯度下降的计算过程指沿着梯度下降的方向求解极小值,迭代公式为:

$$x_{t+1} = x_t + \eta g \tag{15}$$

其中,$g$ 表示梯度负方向,$\eta$ 表示梯度方向上的搜索步长,在 LSTM 中表示为学习率。学习率太大容易导致发散,太小收敛速度太慢,对模型精度起着至关重要的作用。

梯度下降法要在全部训练数据上最小化损失,当样本容量非常大或是迭代次数加大时会非常消耗计算资源。随机梯度下降(stochastic gradient descent, SGD)优化损失函数,按照数据生成分布抽取 $m$ 个小批量(独立同分布的)样本,每一次迭代只计算一个样本的 loss,然后再遍历所有的小批量样本,进行一轮完整的计算。再计算其梯度的均值,最后获得梯度的无偏估计,更新如下所示:

随机梯度下降算法在第 $k$ 个训练迭代的更新方法

输入参数:学习率 $\eta$

输入参数:初始参数 $\theta$

While 满足 do,则停止

从训练数据集中采集 $m$ 个小批量样本 $\{x(1),x(2),\cdots,x(m)\}$,$x^t$ 对应输出目标 $y^t$

梯度估计计算:$\hat{g} \leftarrow + \frac{1}{m}\nabla\theta\sum_i L(f(x^i,\theta),y^i)$

应用更新:$\theta \leftarrow \theta - \eta\hat{g}$

End while

SGD 算法中关键的参数是学习率,在 LSTM 应用中会随着时间的推移逐渐改变学习率。

### 2.1 自适应学习率算法

学习率对模型的性能有显著的影响,决定了参数移动到最优值的速度。若幅度过大,会导致参数可能越过最优值;幅度过小,容易引起运算冗余,导致长时间运算无法收敛。目前常用的学习率算法有 AdaGrad[9]、RMSProp[10]、Adam[11] 等。Schaul[12] 展示了许多算法在大量学习任务上极具价值的比较。有结果表明,RMSProp 和 AdaDelta 算法表现都相当良好。文中使用如下方法改变学习率:

RMSProp 算法

输入参数:全局学习率 $\eta$,衰减速率 $\rho$

输入参数:初始化参数 $\theta$,初始化累积变量 $\gamma = 0$

While 满足 do,则停止

从训练数据集中采集 $m$ 个小批量样本 $\{x(1),x(2),\cdots,x(m)\}$,$x^t$ 对应输出目标 $y^t$

梯度计算:$g \leftarrow + \frac{1}{m}\nabla_{\tilde{\theta}}\sum_i L(f(x^i,\tilde{\theta}),y^i)$

累积梯度:$\gamma = \rho\gamma + (1-\rho)g\odot g$

计算参数更新:$\Delta\theta = -\frac{\eta}{\sqrt{\gamma}}\odot g$

应用更新:$\theta \leftarrow \theta + \Delta\theta$

End while

RMSprop 是一种自适应学习率方法,通过自动调整学习率,从而改变更新方式。

### 2.2 超参数选择

深度学习算法中使用超参数来控制计算,选择超参数有两种方法:手动选择和自动选择。手动选择超参数的主要目标是调整模型的有效容量以匹配任务的复杂性。但手动选择需要了解超参数、泛化误差、训练误差和计算环境等问题,高度依赖平台。自动选择超参数算法不需要制定学习算法的超参数,常见的有网格搜索、随机搜索。

LSTM 预测模型中包含很多参数,其中以学习率、分割窗口、状态向量大小最为关键。将这些超参数笛卡尔乘积得到一组超级参数。Bernoulli and Bengio[13] 对比了网格搜索和随机搜索,发现随机搜索能更快地减小验证误差。

## 3 实　验

### 3.1 实验流程

文中采用基于 TensorFlow[14-15] 的 LSTM 模型的

零售商店日销售额预测方法,建立的预测模型可以预测某大型连锁零售商店未来 7 天的日销售额,预测的基本步骤如图 2 所示。

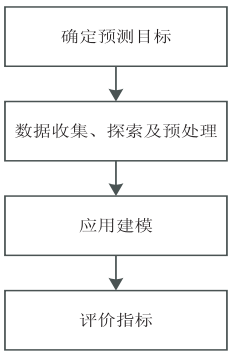


图 2 基本步骤

(1)确定预测目标。

根据相关业务背景分析原始数据源,预测未来 7 天的日销售额情况。

(2)数据收集、探索及预处理。

实验使用的数据是某大型连锁零售商店在 2018 年 1 月 1 日至 2018 年 6 月 30 日的销售数据(数据有缺失),共 267 家店,每家店 181 条数据。分析各个因素与销售额的关系,找出影响因变量的自变量,包括内部数据、外部数据以及额外数据。由于原始数据源并非完整的数据源,其中会出现缺省值或者异常值这些噪声数据,所以需要对缺省值和异常值进行处理,补全有效数据源,然后对数据进行离散化、归一化处理。以 1000011 店为例,处理得到的部分数据如表 1 所示。

表 1 销售额数据

| gid | fildate | $w_1$ | $w_2$ | $w_3$ | $d_1$ | $d_2$ | $d_3$ | $d_4$ | $d_5$ | $d_6$ | $d_7$ | voc | realamt |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1000011 | 2018/1/1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 7 734.42 |
| 1000011 | 2018/1/2 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 4 927 |
| 1000011 | 2018/1/3 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 4 799.8 |
| 1000011 | 2018/1/4 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 5 445.2 |
| 1000011 | 2018/1/5 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 5 664.5 |
| 1000011 | 2018/1/6 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 5 055.8 |
| 1000011 | 2018/1/7 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 4 294.1 |
| 1000011 | 2018/1/8 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 587.2 |
| 1000011 | 2018/1/9 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 5 393.1 |

其中,gid 表示商店 ID,共 267 家(181 条数据);fildate 表示销售时间;$d$(dayofweek)表示当天处于周几;voc 表示国家法定节假日;$w$(weather)表示天气(100-晴,010-雨,001-雪)。

文中利用天气、处在周几以及是否为节假日等特征因素来分析对销售额的影响,根据特征建立模型实现对销售额的预测。

(3)应用建模。

首先,将数据集划分为训练集和测试集;其次,对于训练集做特征筛选,提取有信息量的特征变量,并且去掉无信息量等的干扰特征变量;最后,应用算法建立 LSTM 模型后,结合测试集对算法模型的输出参数进行优化,提高泛化能力,从而提高预测精度,得到最终的训练模型。

(4)评价指标。

预测回归类模型精度常用的评价方法[16]有 RMSE(root mean squared error,均方根误差)、MAPE(mean absolute percentage error,平均绝对百分比误差)、MAE(mean absolute error,MAE)和 MPE(mean percentage error,平均百分比误差)。文中将选取 MAPE 和 MAE 作为衡量标准,MAPE 的大小用来衡量一个模型预测结果的好坏,MAPE 的值越小,则模型的预测结果越好,MAE 的值更好地反映了预测值误差的实际情况,MAE 的值越小,模型预测的误差越小。

### 3.2 实验结果分析

分别采用 LSTM 和 Xgboost 对 1001281 店 2018 年 6 月 24 日—2018 年 6 月 30 日的每日销售额进行预测,LSTM 和 Xgboost 的预测结果如图 3 所示。文中采用绝对百分误差[17]和平均绝对误差[18]作为最终算法质量的衡量标准,MAPE 和 MAE 越低则表明算法误差越小,公式如下:

$$\mathrm{MAPE}=\frac{\sum_{i}^{n}\frac{|y_i-\hat{y}_i|}{y_i}}{n} \tag{16}$$

$$\mathrm{MAE}=\frac{\sum_{i}^{n}|y_i-\hat{y}_i|}{n} \tag{17}$$

其中,$n$ 为预测结果总个数,$y_i$ 为真实值,$\hat{y}_i$ 为预

测值。

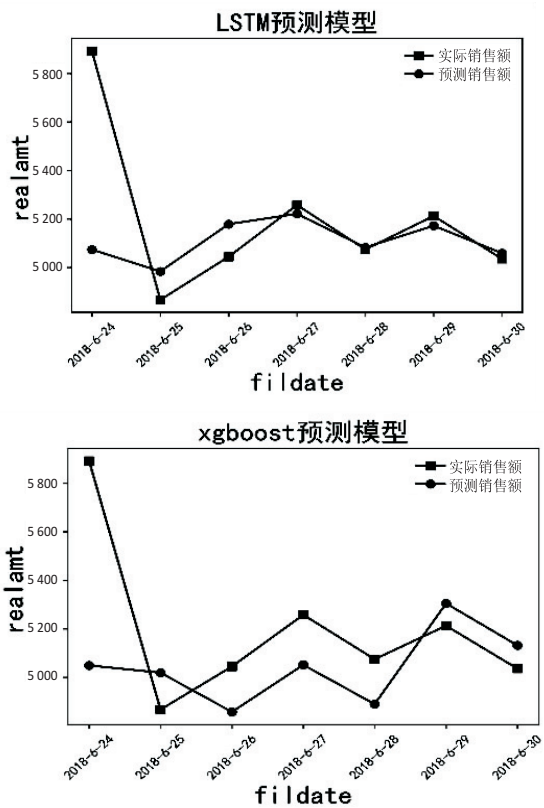


图 3　LSTM、Xgboost 一周日销售额预测结果

LSTM 和 Xgboost 零售商店日销售额预测模型的预测性能指标如表 2 所示。

表 2　两种方法的预测性能指标

| 模型 | MAPE | MAE |
|---|---|---|
| LSTM | 0.031 932 | 168.320 7 |
| Xgboost | 0.056 097 | 251.710 8 |

可见,无论是 MAPE 还是 MAE,LSTM 模型的预测效果都稍优于 Xgboost。

## 4　结束语

文中基于 TensorFlow 框架建立了 LSTM 模型并预测了商品销售额,然后与 Xgboost 模型的预测结果进行了比较。在保证参数调优的情况下,根据 MAPE 和 MAE 评价标准,对比二者的预测结果,发现 LSTM 模型的 MAPE 为 0.031 932,MAE 为 168.320 7,而 Xgboost 模型的 MAPE 为 0.056 097,MAE 为 251.710 8。结果表明,LSTM 模型的准确性更好,用其预测商品的销售额是可行的。该方法具有较高的实用价值。但实际商品销售往往受到政治、经济、文化等多种因素的影响,用该方法预测仍存在不足,如何对各种因素进行取舍,将是今后努力的方向。

**参考文献:**

[1] 叶志祥. 基于 BP 神经网络和 KNN 的店铺销售额预测研究[D]. 淮南:安徽理工大学,2018.

[2] 叶倩怡. 基于 Xgboost 方法的实体零售业销售额预测研究[D]. 南昌:南昌大学,2016.

[3] 孔琳琳,刘　澜,许文秀,等. 基于时间序列分析的港口集装箱吞吐量预测分析[J]. 森林工程,2016,32(5):106-110.

[4] 刘　震,党耀国,钱吴永,等. 基于面板数据的灰色网格关联度模型[J]. 系统工程理论与实践,2014,34(4):991-996.

[5] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation,1997,9(8):1735-1780.

[6] BENGIO Y,SIMARD P,FRASCONI P. Learning long-term dependencies with gradient descent is difficult[J]. IEEE Transactions on Neural Netwoks,1994,5(2):157-166.

[7] KOLEN J F,KREMER S C. Gradient flow in recurrent nets the difficulty of learning long term dependencies[M]//A field guide to dynamical recurrent networks. [s. l.]:Wiley-IEEE Press,2001:237-243.

[8] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks[J]. Neural Information Processing Systems Foundation,2014,4:3104-3112.

[9] DUCHI J, HAZAN E, SINGER Y. Adaptive subgradient methods for online learning and stochastic optimization[J]. Journal of Machine Learning Research,2011,12(7):257-269.

[10] HINTON G E,SRIVASTAVA N,KRIZHEVSKY A,et al. Improving neural networks by preventing co-adaptation of feature detectors[J]. CoRR, 2012, abs/1207. 0580:212-223.

[11] KINGMA D P,BA J. Adam:a method for stochastic optimization[C]//International conference on learning representations(ICLR 2015). [s. l.]:[s. n.],2015:1-13.

[12] HOCKE J,MARTINETZ T. Global metric learning by gradient descent[M]. Germany: Springer International Publishing,2014:129-135.

[13] BERGSTRA J,BENGIO Y. Random search for hyper-parameter optimization[J]. Journal of Machine Learning Research,2012,13(1):281-305.

[14] 李剑风. 融合外部知识的中文命名实体识别研究及其医疗领域应用[D]. 哈尔滨:哈尔滨工业大学,2016.

[15] 杨晓峰. 多层随机森林算法在电信离网预测中的应用[D]. 苏州:苏州大学,2016.

[16] 张　婧. 基于数据挖掘的零售业商品销售预测研究[D]. 成都:四川师范大学,2008.

[17] 谢睿航,申慧涛,段绪晨. 基于 LSTM 网络的地铁线路客流量预测[J]. 中国战略新兴产业,2018(22):106.

[18] 马超群,王晓峰. 基于 LSTM 网络模型的菜品销量预测[J]. 现代计算机,2018(23):26-30.