计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200600104

基于卷积神经网络的汽车销量预测模型

刘吉华　张梦迪　彭红霞　贾兴平
湖北大学商学院　武汉 430062
(jiujh@hubu.edu.cn)

摘　要　传统使用网络搜索数据进行销量预测时多通过人工选取关键词,难以充分考虑所有关键词的搜索量信息。通过使用卷积神经网络提取数据特征,能够解决传统预测方法存在的关键词合成问题。文章首次将深度学习理念引入汽车销量预测领域,首先通过网络爬虫方式获取汽车相关的关键词与网络搜索量,然后根据网络搜索量数据和销量数据的特点设计一种基于卷积神经网络的汽车销量预测模型,并对2019年上半年大众汽车销量做出预测。实验结果显示,与RBF模型、ARIMA模型、ARIMA+RBF混合模型对比,卷积神经网络的预测精度更高,大众品牌的预测精度达到89.51%。由于春节以及新政策出台的影响,2月份为预测误差最大的月份;随着市场的回暖,3月份为预测精度最高的月份。该预测方法为销量预测领域的研究提供了一种新思路。

关键词:卷积神经网络;大众汽车;销量预测;百度指数
中图法分类号　P315.69

Automobile Sales Forecasting Model Based on Convolutional Neural Network

LIU Ji-hua,ZHANG Meng-di,PENG Hong-xia and JIA Xing-ping
School of Business,Hubei University,Wuhan 430062,China

Abstract　Traditionally the kewords selection of web search data is a manual selection task. It is difficult to take all the keywords into consideration. Therefore,the deep learning method is introduced into the field of automobile sales prediction,and the deep learning model is used to extract features of web search data. At first,car-related keywords and online search volumes are collected through web crawlers,and then a car sales prediction model of convolutional neural network is designed,based on the characteristics of web search data and sales data. The model is adopted to predict the sales of Volkswagen in the first half of 2019. The results show that the convolutional neural network can effectively predict car sales,and the accuracy of the prediction reaches 89.51%,compared with RBF,ARIMA and ARIMA+RBF model. Due to the impact of the Spring Festival and the implementation of new policies,it has the largest forecast error in February. However,it has the highest prediction accuracy in March as the market recovers.

Keywords　Convolutional neural network,Volkswagen,Sales forecast,Baidu index

1　引言

从2019年底开始,新型冠状病毒在全国蔓延,严重影响了消费者的购买意愿。面对市场的急剧变化,国内汽车厂商在设定销售目标时非常谨慎。不合适的销售目标将导致厂家销售压力增大,渠道端库存量波动,对销售体系的整体健康发展造成不利影响。目前,随着网络的发展,大量的网络搜索数据已经积累。网络搜索数据具有数据量大、覆盖范围广、私密性强的特点,能够充分反映顾客购物前的信息搜索行为。利用网络搜索数据进行汽车销量预测,可以提前感知市场变化趋势,帮助企业制定生产计划和营销策略,提高市场绩效,减少库存压力。

在汽车销量预测的影响因素研究方面,早期研究者们主要采用历史销量数据来进行预测[1-4]。后来发现仅靠单一的历史数据无法反映市场其他因素的变动,因此将燃料价格、消费者信心指数、失业率等经济指标加入到模型中[5-7]。随着消费者行为的复杂化,覆盖信息全面的网络搜索数据成为了销量预测的有效数据[8-10],汽车销量预测方法主要分为参数化方法和非参数化方法。参数化方法有很多,包括数学模型[2-3]、时间序列模型[8-10],而非参数化方法应用较少。汽车市场的波动涉及到多方面的因素,仅使用参数化方法难以全面捕捉汽车市场的变动规律。深度学习模型具有提取数据特征的优势,相比一般模型,能够在海量数据中提取到更多、更有效的信息,在反映市场变化上具有更好的效果。

深度学习是机器学习的一个分支,它利用深层体系结构学习数据的特征[11-12],非常擅于发现高维数据中复杂的结构[13-14]。基于深度学习算法的基本模型将深度学习算法分为4类:卷积神经网络(CNN)、受限玻尔兹曼机(RBM)、自动

基金项目:国家社会科学基金(15BGL205);国家自然科学基金(71902056)
This work was supported by the National Social Science Fundation of China(15BGL205) and National Natural Science Foundation of China(71902056).
通信作者:张梦迪(13026149802@163.com)

编码器(AE)和稀疏编码(SC)[12]。其中，卷积神经网络是第一个真正成功训练多层网络结构的学习算法，并且在处理网格状的二维数据方面具有令人满意的性能[15]。

卷积神经网络问世以来，在图像识别[16-20]、语音识别方面[21-23]取得了巨大的成果。目前，学者们不仅将卷积神经网络用于传统的图像分类领域，还扩展到各个领域来进行预测方面的工作。在股票价格[24-25]、交通流量[26-27]、电价预测[28]中，卷积神经网络结构均发挥了有效的特征提取作用。卷积神经网络具有强大的特征学习能力，能够克服人工特征抽取的困难，自动挖掘数据的特征，大大提高预测的效率和准确性。如果数据量足够，卷积神经网络具有很强的非线性能力，可以逼近任何非线性函数，在预测中展示出巨大的潜力[29]。为了反映搜索数据背后的行为特征，采用深度学习方法进行预测是一个热门趋势。

目前，国内外学者们大多采用时间序列模型、回归模型、神经网络等方法来预测汽车销量，并取得了一定的成果。然而，目前很少有学者将深度学习理念引入到汽车销量预测领域。由此，本文将研究以下问题：1)如何基于卷积神经网络构建汽车销量预测模型？2)如何验证模型的有效性？

2 卷积神经网络

卷积神经网络是一种具有深度监督学习架构的多层神经网络，可以视为两部分的组合：自动特征提取器和可训练分类器[30-31]。特征提取器包含卷积滤波和下采样。特征映射上的卷积滤波内核通常为 3 * 3，并且滤波之后的下采样操作通常具有 2 的比率。分类器由全连接网络构成，通过特征提取器学习到的特征，将由全连接层进行输出，结构如图 1 所示。

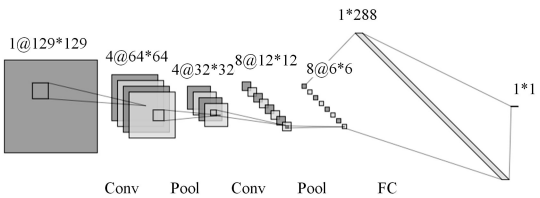


图 1 卷积神经网络结构示意图

Fig. 1 Schematic diagram of convolutional neural network structure

2.1 前向传播

2.1.1 卷积层

网络搜索数据是二维特征数据，输入数据和滤波器进行卷积操作之后，提取该数据的局部特征。本文用 X_j^l 表示卷积层 l 第 j 个通道的输出，W_{ij}^l 表示卷积层 l 第 j 个通道的权值向量，b_j^l 为该层偏置向量，M_j 为卷积核在输入数据上的感受域，“*”表示卷积运算，$f(\cdot)$ 为激活函数，则提取特征的过程如式(1)所示：

$$\begin{aligned} z_j^l &= \sum_{i\in M_j} X_i^{l-1} * W_{ij}^l + b_j^l \\ X_j^l &= f(z_j^l) \end{aligned} \tag{1}$$

在卷积层中，一个卷积核只与输入数据局部相连，与其他数据无关，减少了无关的连接，这是局部感受野；同时，这个卷积核在数据的一个通道上用同样的卷积核进行特征映射，减少了参数数量，这叫权值共享[12]。这两个特点是卷积神经网络简化操作的核心之一。

2.1.2 下采样层

对数据进行下采样，可以在保留数据有用信息的同时减少数据处理量，这是卷积神经网络减少参数计算的另一大特点。下采样一般采用 2 * 2 的窗口，如果是最大池化层，则取 2 * 2 个数中的最大值；如果是平均池化层，则取这 4 个数的平均数。X_j^l 表示 l 层第 j 个通道的特征，β_j^l 是权重，b_j^l 为偏置。$down(\cdot)$ 表示下采样，$f(\cdot)$ 为激活函数，则下采样层的原理如下：

$$\begin{aligned} z_j^l &= \beta_j^l down(X_j^{l-1}) + b_j^l \\ X_j^l &= f(z_j^l) \end{aligned} \tag{2}$$

2.1.3 全连接层

经过卷积层的数据，被拉平成一维向量，再进入全连接层进行分类或回归。全连接层的计算公式如式(3)所示。在销量预测时，为回归问题，输出为一个实数，所以损失函数采用 MSE(均方误差)。MSE 的计算公式如式(4)所示，其中 $\hat{y}_i$ 表示预测值，y_i 表示真实值，m 为样本数。

$$X^l = f(W^l X^{l-1} + b^l) \tag{3}$$

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(\hat{y}_i - y_i)^2 \tag{4}$$

2.2 反向传播

在前向传播阶段，每一轮迭代得到模型的损失，在反向传播阶段，损失通过梯度下降法反向传播，更新卷积神经网络每一层的权重与偏置。通过参数更新，模型的损失不再下降，则达到要求。卷积神经网络的反向传播主要参考文献[32]。

2.2.1 卷积层

由于卷积层的下一层通常为下采样层，因此反向传播时需要用下采样层的灵敏度来表示卷积层的灵敏度。例如，若是采用最大池化层，池化区域大小为 2 * 2，第 $l+1$ 层的每个值都对应第 l 层的 2 * 2 大小的区域，up 操作则将第 $l+1$ 层的每个值放回到这个 2 * 2 区域原始最大值的位置，其他位置补上 0。$\circ$ 代表 Hadamard 积，表示矩阵对应元素相乘，则卷积层的灵敏度为：

$$\delta_j^l = \beta_j^{l+1}(f'(z_j^l) \circ up(\delta_j^{l+1})) \tag{5}$$

在对卷积核参数求导时，需要考虑与该卷积核相乘的输入特征，$x_{u+i-1,v+j-1}^{l-1}$ 是与 W_{ij}^l 相乘的部分输入元素，η 为学习速率，则卷积层的权重更新如下：

$$W_{ij}^l = W_{ij}^l - \eta\sum_u\sum_v(\delta_{uv}^l x_{u+i-1,v+j-1}^{l-1}) \tag{6}$$

由于该通道的每一个输出值都与偏置项有关，因此偏置项的梯度为该通道灵敏度元素总和，则偏置项更新如下：

$$b_j^l = b_j^l - \eta\sum_{u,v}(\delta_j^l)_{u,v} \tag{7}$$

2.2.2 下采样层

当第 l 层为下采样层时，假设第 $l+1$ 层为卷积层，则使用下一层的灵敏度来表示上一层的灵敏度，计算如下：

$$\delta_j^l = \delta^{l+1} * rot180(W_j^l) \circ f'(z_j^l) \tag{8}$$

$rot180(W_j^l)$ 表示将卷积核旋转 180°，δ^{l+1} 需要对其边界进行补 0 处理，使得大小与 z_j^l 相同。对偏置项和权重的更新如式(9)、式(10)所示：

$$b_j^l = b_j^l - \eta\sum_{u,v}(\delta_j^l)_{u,v} \tag{9}$$

$$\beta_j^l = \beta_j^l - \eta\sum_{u,v}(\delta_j^l \circ down(X_j^{l-1}))_{u,v} \tag{10}$$

2.2.3 全连接层

对于一个样本，全连接层的输出层的灵敏度为：

$$\delta^L = \frac{2}{m}(X^L - y) \circ f'(z^L) \tag{11}$$

对于中间层的灵敏度，使用下一层的灵敏度来倒推出上

一层的灵敏度，公式为：

$$\delta^l=(W^{l+1})^{\mathrm{T}}\delta^{l+1}\circ f'(z^l) \tag{12}$$

因此，第 l 层的权重和偏置的更新如下：

$$\begin{aligned}W^l&=W^l-\eta\delta^l(X^{l-1})^{\mathrm{T}}\\ b^l&=b^l-\eta\delta^l\end{aligned} \tag{13}$$

3 基于卷积神经网络的汽车销量预测模型构建

通过研究国内外相关学者的文献成果，本文构建了基于卷积神经网络的汽车销量预测模型，如图 2 所示。模型分为 4 个部分：关键词选取、数据采集与处理、CNN 模型构建、预测与结果分析。

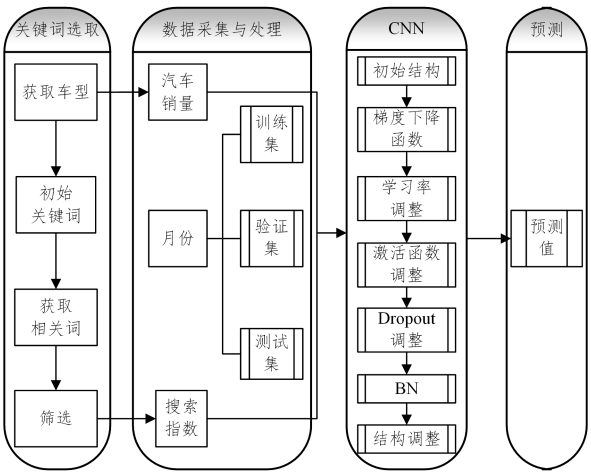


图 2 汽车销量预测模型图

Fig. 2 Sales forecast model diagram

3.1 关键词选取

在探究人们购车前的搜索行为时，关键词的选取至关重要。如果采用经验选词，关键词收集得不全面，则会对预测结果产生影响；若采用技术选词，则需要充足的时间、数据和技术设备。综合考虑，范围选词法更容易实施且准确性能够得到保障。

初始关键词来自搜狐汽车网所提供的汽车车型。采用汽车车型作为初始关键词有两个方面的优点：1）汽车车型是人们在购买汽车时都会考虑的因素，经过一系列的信息搜索之后消费者总是会检索自己想购买的车型；2）通过汽车车型检索到的相关词能够很好地覆盖汽车品牌下每一款汽车的检索，不会存在只关注热门信息而忽视与老旧车型相关信息的问题。

在乘用车市场上，大众品牌连续多年销量稳居第一，本文将大众汽车品牌作为研究对象。关键词选取过程如下：

（1）将大众汽车的车型以及“大众＋车型”的形式作为初始关键词，如大众宝来、宝来，共有 103 个初始关键词；

（2）将初始关键词在百度指数的需求图谱模块中进行检索，获得每个关键词这一年的来源相关词和去向相关词，共得到 1 238 个来源相关词和 1 409 个去向相关词，汇总去重得 1 966 个相关词；

（3）将相关词在百度搜索中检索，若第一页出现的搜索结果都是与汽车有关的信息，则保留该关键词，共保留 1 201 个关键词；

（4）将筛选后的关键词去除未被百度指数收录的关键词，剩下 409 个关键词组成关键词表。

3.2 数据采集与处理

网络搜索数据的获取采用百度指数。百度指数提供了用户在百度搜索引擎上进行信息获取的记录，具有数据量大、数据全面、信息真实的特点，适用于搜索行为的研究。通过 Python 的 Selenium 模块模拟浏览器采集网络搜索数据，获得每个关键词从 2011 年 1 月到 2019 年 6 月每日的搜索量，并将日搜索量汇总成月搜索量，部分数据如表 1 所列。

表 1 部分关键词百度指数搜索量

Table 1 Search volume of some keywords

时间	日产	福克斯	蔚来	辉昂	新迈腾
2011 年 1 月	41317	79925	0	0	5433
2011 年 2 月	43881	91953	0	0	5785
2011 年 3 月	53498	128602	0	0	6152
2011 年 4 月	49069	106069	0	0	11348
2011 年 5 月	51836	112794	61	0	37916
2011 年 6 月	49695	112007	0	0	41804
2011 年 7 月	53835	122082	0	0	88181

搜狐汽车网是信息最快、最全的中国汽车网站，通过 Python 编写爬虫程序，在搜狐汽车网上获得大众汽车品牌的各个车型以及各车型每月的汽车销量。我国汽车市场具有明显的季节性波动，考虑到季节性变化，在模型中加入历史销量数据。同时，汽车销量受到春节等月份的影响，再加入月份这个因子。

卷积神经网络在训练前需要对数据集进行划分，为了对汽车销量进行短期预测，将训练集的样本长度设为 6 个月，测试集的样本长度为 1 个月，验证集的样本长度为 12 个月。采用前 6 个月的网络搜索数据、去年同期的历史销量数据以及月份标识来预测该月份的汽车销量。训练集的样本长度为 6 个月是考虑了消费者最早提前 6 个月在网上搜索汽车相关的信息，而验证集的样本长度为 12 个月是将一年的平均损失作为模型选择的标准，避免有些月份预测准确度较高而春节月份预测精度较低的问题。因此，训练集的样本数为 72，验证集样本数量为 12，测试集样本数为 6。网络搜索数据和历史销量数据需提前进行空值处理以及数据归一化。

3.3 CNN 模型构建与训练

CNN 模型的输入层为网络搜索数据组成的特征矩阵，经过卷积层、池化层和全连接层不断迭代，最终降低训练损失，得到精确的预测值。损失函数采用均方误差（MSE），训练的最终目标是使验证集的均方误差最小。

卷积神经网络的训练在深度学习框架库 Pytorch 上进行。Pytorch 是 torch 的 Python 版本，由 Facebook 开源的神经网络框架，具有操作方便、代码简洁、方便调试、支持动态计算图等优点，受到科研者的欢迎。

3.3.1 *初始结构*

目前，在卷积神经网络的结构与参数设置方面，没有被广泛接受的网络参数选择策略，很大程度上取决于研究者的经验。由于数据集较少，过深的神经网络易出现过拟合，本文将采用两层卷积池化层。网络搜索数据为二维矩阵，而月份数据以及历史销量数据为单个的数值，因此，在卷积池化层提取搜索数据特征后再加入历史数据与月份数据。由于在全连接层新加入了数据，为了充分训练新数据，将全连接层设为 5 层。

输入的搜索数据为 $6*409$ 的矩阵，经过两层卷积池化层提取数据特征，在进入全连接层时将数据拉成一维向量，并加入月份数据及历史销量，最终输出一个销量预测值，数据结构

的变化如图 3 所示。

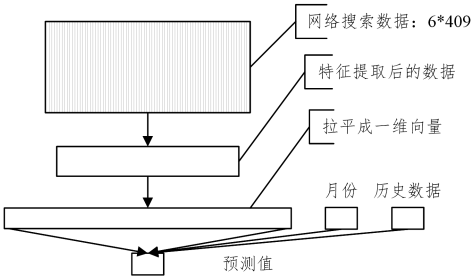


图 3 数据结构变化图

Fig. 3 Data structure change diagram

3.3.2 模型训练

初始结构不稳定且拟合效果不佳，需要对其进行参数调整。本文首先调整模型的梯度下降函数，调整的结果如表 2 所列。激活函数采用 *ReLU*，在学习率为 0.001 时随机梯度下降法（SGD）在该模型结构下取得了最好的效果，平均验证损失为 0.0730。

SGD 的初始学习率对模型有很大的影响，过大的学习率易错过全局最优解，而过小的学习率则易陷入局部最优解。将学习率设置为 0.00001～0.1，观察模型的训练效果，如表 3 所列，在激活函数为 *ReLU* 时，学习率为 0.0001 时模型的训练效果最佳，为 0.0706。

表 2 不同梯度下降函数的预测结果

Table 2 Predictive in different gradient descent functions

梯度下降函数	平均验证损失
ADAM，$lr=0.001$，*ReLU*	0.0866
Adadelta，$lr=1.0$，$rho=0.9$，$eps=1\times10^{-6}$，$weight\_decay=0$，*ReLU*	0.0940
Adagrad，$lr=0.01$，$lr\_decay=0$，weight_decay=0，*ReLU*	0.0965
Adamax，$lr=0.002$，$betas=(0.9,0.999)$，$eps=1\times10^{-8}$，$weight\_decay=0$，*ReLU*	0.0873
ASGD，$LR=0.01$，$LAMBD=0.0001$，$ALPHA=0.75$，$t_0=1\,000\,000$，*ReLU*	0.0956
RMSprop，$lr=0.01$，$alpha=0.99$，$eps=1\times10^{-8}$，$centered=$Fal，*ReLU*	0.3239
Rprop，$lr=0.01$，$etas=(0.5,1.2)$，$step\_sizes=(1\times10^{-6},50)$，*ReLU*	0.0939
SGD，$lr=0.001$，*ReLU*	**0.0730**

表 3 不同学习率与不同激活函数的预测结果

Table 3 Predictive results in different learning rates and different activation functions

学习率	激活函数	平均验证损失
$lr=0.1$	*ReLU*	346.8803
$lr=0.01$	*ReLU*	0.0900
$lr=0.001$	*ReLU*	0.0730
$Lr=0.0001$	*ReLU*	**0.0706**
	RReLU	0.0765
	ReLU6	0.0706
	RLU	0.0971
	PReLU	0.0776
	LeakyReLU	**0.0705***
$lr=0.00001$	*ReLU*	0.5320

初始函数的激活函数是 *ReLU*，但 *ReLU* 存在出现负值时梯度降为 0 的缺点，目前出现了许多改进此缺点的激活函数，如 *RreLU*，*ReLU6*，*PreLU* 和 *LeakyReLU*。从表 3 可以看出，当学习率为 0.0001 时，*LeakyRelu* 的预测效果较好，平均损失为 0.0705。

为了提高训练速度，解决过拟合问题，调整 L2 参数、Dropout 参数、添加 BN 层成为常用的做法。Dropout 是指在训练期间从神经网络中随机丢弃一些连接[33]，减弱了神经元节点间的联合适应性，增强了泛化能力。将 Dropout 进行调整，结果如图 4 所示，在 0.3 之后，随着 Dropout 值的增加，模型损失越来越大。因此，Dropout 在 0.2 与 0.3 附近取值时，模型损失最低，取 0.25 是合适的。添加 BN 层、加入 L2 正则化参数都对模型的精确度没有显著改善，但是训练速度提升较大。

接下来，探究网络结构的变化能否提升模型的预测效果。通过表 4 可以看出，训练完成的情况下，增加卷积层不能提升模型的性能，两层卷积层较为合适。同时，通过调整全连接层的层数发现，全连接层为五层时效果最佳。当网络层数达到一定程度后，要训练的参数过多，超出了机器的负荷，此时出现了训练失败的情况。

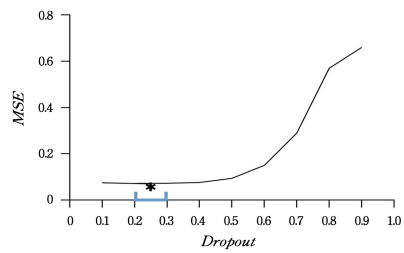


图 4 不同 Dropout 值的预测结果

Fig. 4 Predictive results in different dropout values

表 4 不同网络结构的预测结果

Table 4 Predictive results in different network structures

网络结构	平均验证损失
8conv＞16conv＞1000fc＞10fc＞1fc	0.0838
8conv＞16conv＞100fc＞10fc＞1fc	0.0857
8conv＞16conv＞500fc＞100fc＞10fc＞1fc	0.0810
8conv＞16conv＞1000fc＞500fc＞100fc＞10fc＞1fc	**0.0705**
8conv＞16conv＞32conv＞100fc＞10fc＞1fc	0.0854
8conv＞16conv＞32conv＞500fc＞10fc＞1fc	0.0778
8conv＞16conv＞32conv＞1000fc＞500fc＞10fc＞1fc	训练失败

3.4 模型确定

经过一系列的训练和调整，最终确定了一个基于卷积神经网络的大众品牌汽车销量预测模型。此模型的网络结构为两层卷积池化层、五层全连接层，最终输出汽车销量的预测值。

模型的第一个卷积池化层由 8 个 3＊3 的卷积核和 2＊2 的最大池化操作组成，第二个卷积池化层包括 16 个 3＊3 的卷积核以及 2＊2 的最大池化操作，然后将数据拉平成一维向量并加入历史销量数据和月份因子，将此数据输入到全连接层，经过长度为 1000，500，100，10，1 的全连接层，最终输出一个预测值。卷积神经网络的梯度下降函数采用随机梯度下降法（SGD），学习率为 0.0001，每一层的激活函数为 *LeakyRe-*

lu，Dropout 值为 0.25。在此参数设置下，经验证模型的预测效果较好。

4 模型的应用

本文进行短期预测，采用滚动预测法预测 2019 年 1－6 月的汽车销量。同时，考虑到不同的验证集对预测结果的影响，采用 7 折交叉验证法将每一年的数据作为验证集来选择最优模型，减少单个数据集的随机不确定性和测量误差。预测结果如表 5 所列。

表 5 大众汽车销量预测的 MAPE

Table 5 MAPE of Volkswagen sales prediction

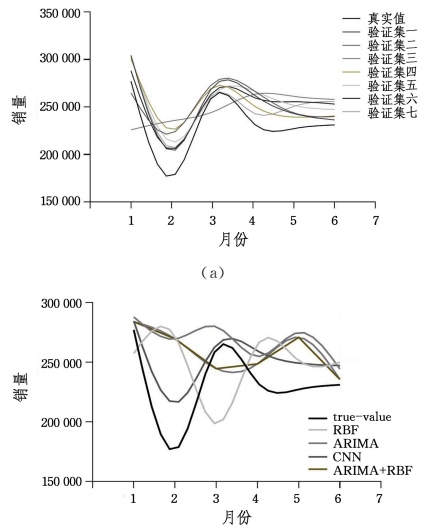
验证集	1 月	2 月	3 月	4 月	5 月	6 月	不同验证集的 *MAPE*
验证集一：2012 年	9.73	15.32	3.76	12.98	6.94	2.25	8.50
验证集二：2013 年	4.44	25.84	5.26	14.91	7.24	3.76	10.24
验证集三：2014 年	18.47	33.36	6.17	12.92	14.92	11.62	16.24
验证集四：2015 年	8.52	28.09	3.68	8.07	5.32	4.06	9.62
验证集五：2016 年	9.52	20.71	0.35	14.60	10.73	7.05	10.49
验证集六：2017 年	4.00	16.29	1.65	11.71	12.42	9.51	9.26
验证集七：2018 年	9.98	17.07	0.83	4.58	10.32	10.69	8.91
不同月份的 *MAPE*	9.24	22.38	3.10	11.40	9.70	6.99	10.47*

对不同的月份进行预测时，1 月份的预测平均损失为 9.24%，2 月份为 22.38%，3 月份为 3.10%，4 月份为 11.40%，5 月份为 9.70%，6 月份为 6.99%。可以看出，3 月份的预测精度最大，且在不同验证集上的预测效果波动不大。2 月份的损失最大，且在不同验证集上 2 月份的损失都大于其他月份。这是因为每年的 2 月份处于春节时期，且受每年新政策的影响，需求波动较大。

当采用不同的验证集时，2012 年的数据作为验证集一预测得到的损失为 8.50%，2013 年作为验证集二的预测损失为 10.24%，2014 年作为验证集三的预测损失为16.24%，2015 年的数据作为验证集四的预测损失为 9.62%，2016 年作为验证集五的预测损失为 10.49%，2017 年作为验证集六的预测损失为 9.26%，2018 年作为验证集七的预测损失为 8.91%。当采用 2014 年作为验证集时，预测的平均损失最大，为 16.24%，远高于其他验证集的损失，且 1 月、2 月、5 月、6 月的预测损失均大于采用其他验证集进行预测的损失。出现这种情况，可能是因为 2014 年的数据相较于其他年份的数据存在异常的波动，而 2014 年的数据未参与到模型训练，仅仅作为验证集确定模型参数，此时模型未学习所有的波动情况，预测能力较弱。

最终，对所有验证集的预测效果进行平均，得到预测损失为 10.47%，即 2019 年上半年的预测准确性达到 89.51%。通过各月份的预测效果看出，每年的 3 月份市场波动最平稳，此时中国的春节假期已过，市场逐渐回暖，消费者的需求上升。每年的 2 月份市场波动最剧烈，若是春节处于 2 月中下旬，则 2 月上旬消费者的购车需求仍较大，若是春节处于 2 月上旬，则此时消费者的购车需求已经较低。此外，每年的 1 月份为新政策出台时期，政策一般具有滞后影响，则 2 月份的汽车市场将受到政策影响。

如图 5(a)所示，除了验证集三以外，各验证集对 1－6 月的预测值均高于实际值。这是因为一些城市将在 2019 年 7 月份实行国六排放政策，在近几个月内，消费者购买汽车会持观望态度，导致市场低于预期。采用 2014 年的数据作为验证集三时，预测效果与其他验证集不同。2014 年国家出台了各种有关汽车的新政策，如免征新能源车购置税、买优惠车可享“实时核定”购置税、汽车“双反”政策到期，这些政策对大众品牌的汽车市场造成冲击，致使 2014 年大众汽车销量波动较大。除了验证集三以外，其他验证集的预测值的波动趋势基本与真实值相吻合。



(a)

(b)

图 5 不同验证集和不同模型的销量预测结果

Fig. 5 Sales prediction different validation sets and different model

为了比较不同模型的预测结果，采用文献[2]的汽车销量预测方法对大众汽车销量进行预测。使用 ARIMA 和 RBF 模型进行滚动预测，预测结果如表 6 所列。RBF 模型、ARIMA 模型以及 ARIMA＋RBF 混合模型的 *MAPE* 分别为 19.61%，16.69%，14.90%，预测误差均大于 CNN（为 10.47%）。对 2019 年 1 月到 6 月的汽车销量进行预测时，CNN 模型各月的预测误差均小于 RBF 模型。ARIMA 模型在 2 月份的预测误差较大，为 52.75%，这是因为线性模型难以捕捉现实世界中极端的影响因素。ARIMA＋RBF 混合模型的预测效果好于 RBF 模型与 ARIMA 模型，预测了销量的线性变动趋势，同时对随机波动进行拟合，但仍然难以应对 2 月份的不规则变动。

表 6 不同模型的预测误差

Table 6 MAPE of different model

	1 月	2 月	3 月	4 月	5 月	6 月	*MAPE*
CNN	9.24	22.38	3.10	11.40	9.70	6.99	10.47
RBF	7.12	54.12	23.94	13.71	10.63	8.16	19.61
ARIMA	3.91	52.75	7.31	9.36	21.01	5.77	16.69
ARIMA＋RBF	2.48	52.84	6.16	6.71	19.23	1.99	14.90

各模型的预测效果如图 5(b)所示。相较于 ARIMA 模型、RBF 模型以及 ARIMA＋RBF 的混合模型，CNN 模型能够拟合真实值的波动趋势。

结束语 本文在充分挖掘网络搜索数据的基础上，提出了一个基于卷积神经网络的汽车销量预测模型，并对大众汽车品牌进行预测，其预测准确度达到 89.51%。此方法具有一定的理论和实践意义。在理论层面，构建了网络搜索数据与卷积神经网络结合的汽车销量预测模型，将深度学习理论引入到传统的汽车销量预测研究中，为汽车销量预测以及搜索数据的利用提供了新的思路。同时，使用卷积神经网络代

替人工合成关键词，克服了人工抽取特征时权重选择困难的问题。在实践层面，有利于汽车公司充分利用用户的搜索行为，发现用户的关注重点，从而在营销策略以及产品设计上做出符合顾客需求的改动。

通过对预测模型的结构进行调整，发现模型的第一个卷积池化层包含 8 个 3 * 3 的卷积核和 2 * 2 的最大池化操作，第二个卷积池化层包含 16 个 3 * 3 的卷积核和 2 * 2 的最大池化操作，后接长度为 1 000，500，100，10，1 的全连接层所得到的平均损失最小。对模型的激活函数、梯度下降函数、学习率、Dropout 参数等进行调整，发现 *LeakyReLU* 比 *ReLU* 更适合本文模型，且随机梯度下降法在学习率为 0.0001 时能取得较好的效果，Dropout 值在 0.2 到 0.3 之间时模型的损失较低，添加 BN 层以及加入 L2 正则化参数对结果没有太大影响。

本文在预测汽车销量时仍然存在不足：1）汽车销量数据为月度数据，粒度较大，而上一周的搜索量可能已经影响到这周的汽车购买意愿，因此无法研究短时间内网络搜索数据对汽车销量的影响；2）由于实验条件有限，未能对卷积神经网络结构更复杂的模型进行研究。

参 考 文 献

[1] LI X，ZONG Q，TONG L. Hybrid Forecasting Method for Automobile Sale[J]. Journal of Tianjin University (Social Sciences)，2006(3)：175-178.

[2] ZHAO H，XV H，LIANG P，et al. Application of an Optimal Combined Forecasting Method to Prediction of Demand for Private Cars[J]. Industrial Engineering Journal，2008，11(1)：126-128.

[3] LIU Y Q，WANG M，WANG J Y. The Predictive Research on China's New Energy Vehicles Market[J]. Research on Economics and Management，2016，37(4)：86-91.

[4] CHEN D. Chinese automobile demand prediction based on ARIMA model[C]// 2011 4th International Conference on Biomedical Engineering and Informatics (BMEI). Shanghai：IEEE，2011：2197-2201.

[5] SA-NGASOONGSONG A，BUKKAPATNAM S T S，KIM J，et al. Multi-step sales forecasting in automotive industry based on structural relationship identification[J]. International Journal of Production Economics，2012，140(2)：875-887.

[6] GAO J J，XIE Y A. Chinese automobile sales forecasting using economic indicators and typical domestic brand automobile sales data：A method based on econometric model[J]. Advances in Mechanical Engineering，2018，10(2)：168781401774932.

[7] ZIROGIANNIS N，DUNCAN D. The effect of CAFE standards on vehicle sales projections：A Total Cost of Ownership approach[J]. Transport Policy，2019，75(5)：70-87.

[8] YAN C S，FELIPE L. Nowcasting with Google Trends in an Emerging Market[J]. Working Papers Central Bank of Chile，2010，32(4)：289-298.

[9] FANTAZZINI D，TOKTAMYSOVA Z. Forecasting german car sales using google data and multivariate models[J]. International Journal of Production Economics，2015，170：97-135.

[10] YONG Z，MINER Z，NANA G，et al. Forecasting electric vehicles sales with univariate and multivariate time series models：The case of China[J]. PLOS ONE，2017，12(5)：e0176729.

[11] CHEN X W，LIN X. Big Data Deep Learning：Challenges and Perspectives[J]. Quality Control Transactions，2014，2(2)：514-525.

[12] GUO Y，LIU Y，OERLEMANS A，et al. Deep learning for visual understanding：A review[J]. Neurocomputing，2016，187(26)：27-48.

[13] LECUN Y，BENGIO Y，HINTON G. Deep learning[J]. Nature，2015，521(7553)：436-444.

[14] LITJENS G，KOOI T，BEJNORDI B E，et al. A survey on deep learning in medical image analysis[J]. Medical Image Analysis，2017，42(9)：60-88.

[15] LIU W，WANG Z，LIU X，et al. A survey of deep neural network architectures and their applications[J]. Neurocomputing，2017，234(19)：11-26.

[16] KRIZHEVSKY A，SUTSKEVER I，HINTON G. ImageNet Classification with Deep Convolutional Neural Networks[J]. Advances in Neural Information Processing Systems，2012，25(2)：1097-1105.

[17] ZHANG J，SHAO K，LUO X. Small sample image recognition using improved Convolutional Neural Network[J]. Journal of Visual Communication & Image Representation，2018，55(8)：640-647.

[18] TATSUMA A，AONO M. Food Image Recognition Using Covariance of Convolutional Layer Feature Maps[J]. Ieice Transaction on Informathion and Systems，2016，E99D(6)：1711-1715

[19] SHIN H C，ROTH H R，GAO M，et al. Deep Convolutional Neural Networks for Computer-Aided Detection：CNN，Architectures，Dataset Characteristics and Transfer Learning[J]. IEEE Transactions on Medical Imaging，2016，35(5)：1285-1298.

[20] ABD MUBIN N，NADARAJOO E，et al. Young and mature oil palm tree detection and counting using convolutional neural network deep learning method[J]. International Journal of Remote Sensing，2019，40：7500-7515.

[21] SAINATH T N，KINGSBURY B，et al. Deep Convolutional Neural Networks for Large-scale Speech Tasks[J]. Neural Networks，2015，64：39-48.

[22] ABDEL-HAMID O，MOHAMED A R，JIANG H，et al. Convolutional Neural Networks for Speech Recognition[J]. IEEE/ACM Transactions on Audio，Speech，and Language Processing，2014，22(10)：1533-1545.

[23] SWIETOJANSKI P，GHOSHAL A. Convolutional Neural Networks for Distant Speech Recognition[J]. IEEE Signal Processing Letters，2014，21(9)：1120-1124.

[24] JIASHENG C，JINGHAN W. Stock price forecasting model based on modified convolution neural network and financial time series analysis[J]. International Journal of Communication Systems，2019，32(12)：e3987. 1-e3987. 13.

[25] GUNDUZ H，YASLAN Y. Intraday prediction of Borsa Istanbul using convolutional neural networks and feature correlations[J]. Knowledge-based Systems，2017，137(dec. 1)：138-148.

[26] NIU K，CHENG C，CHANG J，et al. Real-Time Taxi-Passenger Prediction with L-CNN[J]. IEEE Transactions on Vehicular Technology，2019，68(5)：4122-4129.

（下转第 189 页）

[5] SCHERE K L. U. S. EPA MODELS-3/CMAQ - STATUS AND APPLICATIONS[R]. Presented at US/German Ozone/Fine Particle Science and Environmental Chamber Workshop, Riverside, CA, 1999.

[6] BAI S N,SHEN X L. PM2.5 PREDICTION BASED ON LSTM RECURRENT NEURAL NETWORK [J]. Computer Applications and Software ,2019,36(1):73-76,110.

[7] HAN W, WU Y L, REN F. The Prediction of Air Pollutants Based on Full Connection and LSTM Neural Network [J]. Geomatics World ,2018,25(3):34-40.

[8] MENG Q. Air Quality Classification Prediction Based on Random Forest Model[J]. J Chongqing Technol Business Univ. (Nat Sci Ed) 2018,179(3):33-37.

[9] DU X, F J Y, L S Q, S W. PM2.5 concentration prediction model based on random forest regression analysis[J]. Telecommunication Science 2017,33(7):66-75.

[10] KULKARNI G E, MULEY A A, DESHMUKH N K, et al. Autoregressive integrated moving average time series model for forecasting air pollution in Nanded city, Maharashtra, India [J]. Modeling Earth Systems & Environment, 2018, 4(4): 1435-1444.

[11] PENG S J, SHEN J C, ZHU X. Forecast of PM2.5 Based on the ARIMA Model[J]. Safety and Environmental Engineering , 2014,21(6):125-128.

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

[13] SRIVASTAVA N,HINTON G,KRIZHEVSKY A,et al. Dropout: A Simple Way to Prevent Neural Networks from Overfitting[J]. Journal of Machine Learning Research, 2014, 15(1): 1929-1958.

[14] KINGMA D,BA J. Adam: A Method for Stochastic Optimization[J]. Computer Science, 2014.

[15] ZHAI B, CHEN J. Development of a stacked ensemble model for forecasting and analyzing daily average PM2.5 concentrations in Beijing, China[J]. Science of the Total Environment, 2018,635:644-658.

[16] GAN K,SUN S,WANG S,et al. A secondary-decomposition-ensemble learning paradigm for forecasting PM2.5,concentration [J]. Atmospheric Pollution Research, 2018: S1309104217306190.

[17] WANG P,ZHANG H,QIN Z,et al. A novel hybrid-Garch model based on ARIMA and SVM for PM2.5 concentrations forecasting[J]. Atmospheric Pollution Research, 2017: S1309104216302616.

[18] NG K Y,AWANG N. Multiple linear regression and regression with time series error models in forecasting PM10 concentrations in Peninsular Malaysia[J]. Environmental Monitoring and Assessment,2018,190(2):63.

LIU Meng-yang, born in 1998, undergraduate. His main rersearch interests include machine learning, deep learning and high performance computing.

WU Li-juan, born in 1987, postgraduate, engineer. Her main research interestsinclude intelligent information processing, prediction of air pollutants and big data analysis.

(上接第 183 页)

[27] ZHANG W B,YU Y H. Short-term traffic flow prediction based on spatio-temporal analysis and CNN deep learning[J]. Transportmetrica A-transport Science,2019,15:1688-1711.

[28] KUO P H,HUANG C J. An Electricity Price Forecasting Model by Hybrid Structured Deep Neural Networks[J]. Sustainability, 2018,10(4):1280-1296.

[29] HUANG Z,HUANG G,CHEN Z,et al. Multi-Regional Online Car-Hailing Order Quantity Forecasting Based on the Convolutional Neural Network[J]. Information,2019,10(6):193-206.

[30] NIU X X,SUEN C Y. A novel hybrid CNN-SVM classifier for recognizing handwritten digits[J]. Pattern Recognition, 2012, 45(4):1318-1325.

[31] COLLOBERT R,WESTON J,BOTTOU L,et al. Natural Language Processing (Almost) from Scratch[J]. Journal of Machine Learning Research,2011(12).

[32] BOUVRIE J. Notes on Convolutional Neural Networks [R]. MIT CBCL Tech Report,Cambridge,MA,2006.

[33] SRIVASTAVA N,HINTON G,KRIZHEVSKY A,et al. Dropout: A Simple Way to Prevent Neural Networks from Overfitting[J]. Journal of Machine Learning Research,2014,15(1): 1929-1958.

LIU Ji-hua, born in 1971, Ph. D. His main research interests include data mining and data analysis.

ZHANG Meng-di, born in 1995, master. Her main research interests include data mining and data analysis.