基于 CNN-LightGBM 模型的高速公路交通量预测

张　振[1]，曾献辉[1,2]

（1. 东华大学 信息科学与技术学院，上海 201620；
2. 数字化纺织服装技术教育部工程研究中心，上海 201620）

摘　要：有效的交通流量预测对人们出行和交管部门监管都有着重要的意义。传统的交通量预测模型主要基于交通流的时间特性，未结合交通流的时间和空间特性进行深入挖掘，因此预测效果有时不佳。提出了一种基于 CNN 与 LightGBM 结合的交通流预测模型，首先利用 CNN 模型挖掘出高速公路相邻路段监测点和出入口的时间和空间关联性，实现对交通流数据的时空特征提取，然后将 CNN 提取到的特征向量输入到 LightGBM 模型中进行预测。为了验证模型的有效性，实验中使用了多种预测模型进行对比，实验结果表明，所提出的考虑到时空特性的 CNN-LightGBM 组合的模型可以明显降低预测误差，是一种有效快速的交通流预测模型。

关键词：交通流预测；CNN-LightGBM；时空关联性；高速公路

中图分类号：U491.1　　**文献标识码**：A　　DOI：10.19358/j.issn.2096-5133.2020.02.007

引用格式：张振，曾献辉. 基于 CNN-LightGBM 模型的高速公路交通量预测[J]. 信息技术与网络安全，2020，39(2)：34-39.

Prediction of highway traffic flow based on CNN-LightGBM model

Zhang Zhen[1], Zeng Xianhui[1,2]

(1. School of Information Science and Technology, Donghua University, Shanghai 201620, China;
2. Engineering Research Center of Digitalized Textile & Fashion Technology, Ministry of Education, Shanghai 201620, China)

Abstract: Effective traffic flow forecasting is of great significance to people's travel and traffic management supervision. Traditional traffic volume prediction models are mainly based on the time characteristics of traffic flow, however, these models don't combine the time and space characteristics of traffic flow for in-depth mining, so sometimes these models don't perform well. This paper proposes a traffic flow prediction model based on the combination of CNN and LightGBM. The CNN model is used to excavate the temporal and spatial correlation between the monitoring points and the entrances and exits of the adjacent sections of the highway to realize the spatiotemporal feature extraction of the traffic flow data, and then the feature vector extracted by CNN is input into the LightGBM model for prediction. In order to verify the effectiveness of the model, a variety of prediction models are used in the experiment for comparison. The experimental results show that the proposed model of CNN-LightGBM considering the spatio-temporal characteristics can significantly reduce the prediction error and is an effective and fast traffic flow forecasting model.

Key words: traffic flow prediction; CNN-LightGBM; spatiotemporal correlation; highway

0　引言

准确的交通量预测是当今智慧交通的重要基础，是交通状况判别的重要基石之一[1]。人们从上个世纪开始就在交通流预测领域做了很多交通预测研究，截止目前为止常见的交通量预测方法主要包括基于统计的预测方法、基于时间序列的交通量预测方法、基于神经网络的交通量预测方法以及基于机器学习的交通量预测方法几种。

基于统计的交通量预测方法较多，比如多元线性回归法、卡尔曼滤波器[2-3]和 K 近邻算法[4]等，这些方法主要根据历史流量数据预测未来交通流量分布，但是这些方法无法精准地预测道路短期拥堵的情况。基于时间序列的交通量预测方法如差分自回归滑动平均模型[5]，主要是将历史的流量数据按照时间排列成为时间序列，根据时间序列分析数据流的变化趋势从而预测未来的交通流量，但是这种算法的缺点是在处理数据量较大、维度较高的数据时效果一般，推广能力较差。基于神经网络交通

量预测方法如 GRU[6] 和 LSTM[7-8]，这些模型存在着计算过程中收敛速度慢、计算时间较长、容易过拟合等缺点。基于机器学习的交通量预测方法如 GBDT 模型[9]、Xgboost 模型[10] 和随机森林模型[11]，这些模型对交通流时空挖掘效果不大理想。

单一预测模型往往存在一定的缺陷，影响模型预测精度。由于高速公路交通流量变化很容易受到外界环境的影响，在空间上上下游监测点和开放路段的出入口交通流量变化对该路段交通量变化有一定的影响。深度挖掘交通流量的时间和空间特性不仅可以降低外界环境对交通流变化的影响，还可以考虑到相邻监测点和出入口之间的因果关系，提高模型的预测精度。CNN 模型主要优势是可以进行特征提取[12]，深度挖掘高速公路待预测监测点和周边检测点之间的时空上的联系。LightGBM 算法使用集成学习的方式[13]，可以快速实现梯度提升，具有较快的运行速度和较高的预测精度。本文结合这两种模型的优点提出了 CNN-LightGBM 组合模型的方式进行交通量预测。

1 基于 CNN-LightGBM 交通量预测模型

1.1 基于 CNN 的特征矩阵构建

卷积神经网络（Convolutional Neural Network，CNN）模型最早可追溯至 1980 年由 FUKUSHIMA K 等人提出的 Neocognitron 模型[14]，于 1998 年正式由 LE CUN Y 等人提出[15]。CNN 主要由输入层（input）、卷积层（conv）、池化层（pool）、全连接层（fc）和输出层（output）构成[16]，这种模型结构可以有效地减少权值的数量，简化网络模型，同时也可以将数据直接用作网络输入，有效地减少了特征的提取和数据重构的复杂性，该模型如图 1 所示。目前 CNN 模型大多应用于图像领域，在交通量预测方面，由于 CNN 模型具有比较差的泛化能力，如果在实际交通量发生突变的情况下，之前建立的模型可能会出现完全失效情况[17]。但是 CNN 模型具有很强的特征提取能力，大多应用于模型特征提取。

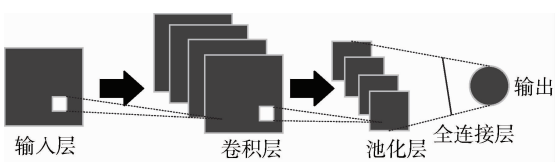

图 1 CNN 模型整体结构

高速公路的交通流量数据可以从时间和空间两个角度做出区分，在时间上，当前时刻交通流量数据是对上一时刻的数据的继承与对下一时刻数据的延伸；在空间上，对于开放的高速公路路段来说，当前监测点的交通量变化不仅仅与上下游的交通量有关，还会受到高速公路出入口交通量变化的影响。根据 CNN 捕捉的时空信息的特点，构造包含时间和空间信息的二维矩阵。

空间上监测点在 t 时刻的交通流量数据为：

$$S = \{x_{p,t}, x_{p_1,t}, \cdots, x_{p_n,t}\} \tag{1}$$

其中，$x_{p,t}$ 表示目标监测点 p 在 t 时刻的交通流量数据；$x_{p_n,t}$ 表示周边检测点 p_n 在 t 时刻交通流量数据。

构建时空特征矩阵为：

$$\boldsymbol{x} = \begin{bmatrix} x_{p,t-m} & x_{p_1,t-m} & \cdots & x_{p_n,t-m} \\ x_{p,t-(m-1)} & x_{p_1,t-(m-1)} & \cdots & x_{p_n,t-(m-1)} \\ \cdots & \cdots & \cdots & \cdots \\ x_{p,t-1} & x_{p_1,t-1} & \cdots & x_{p_n,t-1} \end{bmatrix} \tag{2}$$

其中，$x_{p,t-m}$ 表示监测点 p 在当前时间前 m 个时间统计单位的时刻交通流量数据。本实验中 m 取值为 5。

1.2 基于 CNN-LightGBM 模型的交通量预测

LightGBM 模型于 2016 年由微软亚洲研究院提出[18]，是 GBDT 模型的变体，主要用于解决 GBDT 在处理大量数据时遇到的问题。由于 LightGBM 是 GBDT 算法的提升，其基本原理与 GBDT 原理基本一致。GBDT 算法属于一种 Bossting Tree 算法，Boosting 算法可以表示为决策树的加法模型：

$$f_N(x) = \sum_{n=1}^{N} T(x;\Theta_n) \tag{3}$$

其中，$T(x,\Theta_n)$ 表示第 n 棵决策树，Θ_n 表示其参数；N 为树的棵数；x 表示输入样本。

GBDT 采用前后向分布算法的方式进行优化求解，其模型为：

$$f_n(x) = f_{n-1}(x) + T(x;\Theta_n) \tag{4}$$

其中，$f_{n-1}(x)$ 为经过 $n-1$ 步训练的提升树模型，$T(x;\Theta_n)$ 为第 n 步需要学习的决策树。前向分布算法希望将 $T(x;\Theta_n)$ 和 $f_{n-1}(x)$ 相加后能够使 $f_n(x)$ 在训练集上的经验误差最小。Bosting Tree 算法在每个步骤都会生成一个弱决策树模型，并将其累积在整个模型中，从而将模型的经验误差降至最低。GBDT 模型就是沿着当前模型误差函数的负梯度方向生成每个弱决策树模型。

LightGBM 是在 GBDT 模型的基础上结合了 Gradient-based One-Side Sampling(GOSS)算法和 Exclusive Feature Bundling(EFB)算法以增强梯度。GOSS 算法主要用于对训练样本进行采样,能够在保留大梯度样本的同时减少小梯度样本的数量,从而降低决策树生成的复杂性,同时 GOSS 算法能够提高泛化性能。EFB 算法用于减少较高特征维度下稀疏数据要素的数量,并从要素角度优化算法复杂度。

结合 CNN 挖掘交通流空间特性与 LightGBM 快速高效预测的特性,本文构造出 CNN-LightGBM 交通量预测模型的整体结构,如图 2 所示。

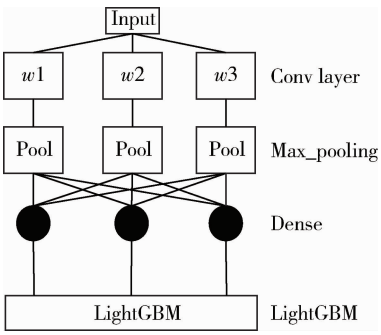

图 2 CNN-LightGB 的预测模型结构

整体思路可以分为以下 4 步:

(1)统计各个监测点的交通流量数据,并对交通流量数据进行预处理;

(2)将交通流量归一化后放入 CNN 模型进行训练,由 CNN 模型提取交通流时空特征;

(3)将 CNN 模型提取的特征向量输入到 LightGBM 模型;

(4)使用 LightGBM 模型对交通量进行预测。

2 高速公路交通量预测应用研究

2.1 数据来源

本文实验数据来源自某省交通管理局提供的某高速公路路段道路中心监测点、收费站口监测点以及服务区监测点对机动车监测的真实数据,监测点位置分布如图 3 所示。监控点记录每一辆通行车辆的车牌号、通行监测点的时间、机动车通行方向与汽车类型等。本实验采用 2018 年 11 月 1 日到 2018 年 12 月 20 日共计 300 多万条监测数据,首先根据监测点记录的机动车类型,根据国家机动车当量折算标准,将不同机动车交通量按照不同的折算系数转换成标准车型的当量交通量,当量交通量比单纯地统计汽车个数更加具有实际意义。由于短时交通流的时间跨度并没有非常标准的定义,本实验以 15 min 作为最小时间单位,统计道路每 15 min 汽车当量交通量,得到 4 800 条数据。其中,以 2018 年 11 月 1 日到 2018 年 12 月 15 日时间段的交通当量数据作为训练样本,以 2018 年 12 月 16 日到 2018 年 12 月 20 日时间段的交通当量作为测试样本,共计得到 4 320 个训练样本,480 个测试样本。

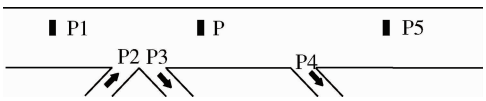

图 3 监测点位置图

2.2 数据预处理

由于高速公路监测点对数据进行实时监测传输时可能会由于监测器异常、网络异常或者存储异常等因素导致数据存在重复、缺失或错误等问题,严重影响数据的真实性,因此需要对这些数据进行合理有效的处理。

首先对监测点数据进行处理。监测点数据主要存在数据重复和数据缺失问题。对于数据重复问题,根据重复数据记录的时间排序,保留第一次出现的车辆数据,其他的重复数据进行删除处理。数据异常指的是数据库记录的数据有部分缺失,如果需要的关键数据没有缺失则不做处理,如果有缺失则做删除处理。

对处理后的监测点数据进行统计,根据我国《公路工程技术标准》[19]的规定,对各种类型的车型进行机动车当量折算,将不同的机动车车型转换为标准的机动车车型。当量能够体现出机动车在道路上的占有情况。然后根据监测点记录的机动车通行时间以 15 min 为最小时间单位,对机动车数量进行统计。

由于监测点数据缺失和数据异常问题会对统计数据造成影响,需要对统计出来的数据再次进行处理。考虑到高速公路机动车数据具有周期性、连续性和重复性的特点,采用历史值填充和平均值法进行数据填充。对于少量数据缺失问题,采用平均值填充的方式对数据进行补充。如果有大量数据缺失,则根据道路数据周期性特点,采集之前一些天数相同的时间数据进行加权平均,将缺失数据进行填充。

2.3 CNN-LightGBM 模型设计

从图 3 可以发现,待预测监测点 P 的交通流量与上游监测点 P1、出入口 P2 和 P3、下游出口 P4 和监测点 P5 都有较大的关联性,上下游的交通量变化会影响到 P 监测点的交通量变化。在模型建立前,首先对数据进行归一化处理,让数据映射到一定的范围之内,这样可以减少数据范围对预测效果的影响。CNN-LightGBM 组合模型训练过程主要分为 CNN 特征提取和 LightGBM 预测两个部分。

(1)使用 CNN 提取特征。CNN 特征提取使用两次卷积和两次池化。第一次 CNN 模型利用 3×3 的卷积方式将数据转换成[2,6,6]的特征图,池化之后得到[2,3,3]的特征图。第二次 CNN 使用 3×3 的卷积将数据转换成[4,3,3]的特征图,池化之后得到[4,2,2]的特征图。之后使用 Reshape 对数据重组,将二维矩阵转换成一维向量。使用全连接方式将数据变成 128 维度数据。实验中 CNN 使用 MSE 作为损失函数,卷积核大小为 3×3,步长为 1,padding 为 1,第一次卷积核的个数由 1 变成 2,第二次卷积核由 2 变成 4。池化的窗口大小为 3×3,步长为 2,padding 为 1。

(2)将高维特征向量输入到 LightGBM 模型进行预测。将 CNN 模型得到的特征向量作为输入传送给 LightGBM 模型,由 LightGBM 进行交通量预测。LightGBM 模型学习率为 0.1,最大深度为 8,num_leaves 为 50,其他均使用默认值。

CNN-LightGBM 模型设计流程如图 4 所示。

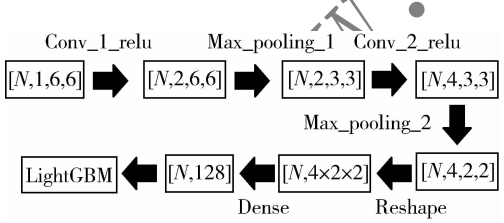

图 4 CNN-LightGBM 模型设计

2.4 模型评价标准

为了验证模型的有效性,实验采用了均方根误差(Root Mean Squart Error,RMSE)、平均绝对百分比误差(Mean Absolute Percentage Error,MAPE)和平均绝对差值(Mean Absolute Error,MAE)作为评价标准[20],用于判断预测交通当量数据的准确性。

$$RMSE = \left[\frac{1}{n}\sum_{i=1}^{n}(|y_i - \hat{y}_i|^2)\right]^{\frac{1}{2}} \quad (5)$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i} \quad (6)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (7)$$

其中,$\hat{y}_i$ 表示经过预测模型预测出来的交通当量数据,y_i 表示经过监测点采集并统计处理后的真实交通当量数据,n 为交通当量数据的个数。

2.5 研究结果

为了验证高速公路空间特性对交通量的影响,本实验将考虑空间特性的 LightGBM-1 模型与未考虑空间特性的 LightGBM-2 模型作对比,如图 5 所示。通过表 1 对比发现考虑到空间特性的 LightGBM-1 模型比未考虑到空间特性的 LightGBM-2 模型的 MAPE 降低了 12.75%,RMSE 降低了 16.20%,MAE 降低了 16.91%。说明考虑空间特性的 LightGBM-1 模型预测效果更好,这表明,挖掘交通流的空间特性可以降低预测误差。

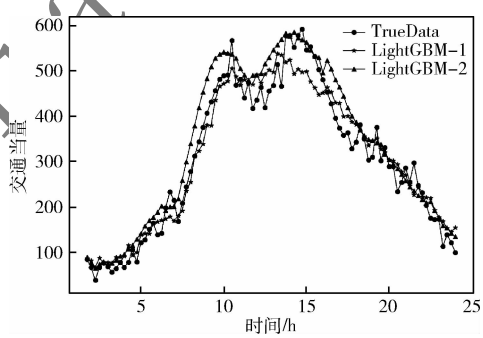

图 5 LightGBM-1 与 LightGBM-2 模型预测结果

将 CNN 提取的特征向量全连接后得到 CNN 预测模型,将 CNN 的预测结果与真实值进行对比,如图 6 所示。结合 CNN 模型与 LightGBM 模型优势的 CNN-LightGBM 模型的预测结果如图 7 所示。由表 1 可知,考虑到时空特性的 CNN-LightGBM 与未考虑交通流空间特性的 LightGBM-2 相比,MAPE 降低了 28.28%,RMSE 降低了 30.87%,MAE 降低了 33.56%。CNN-LightGBM 模型与考虑交通流空间特性的 LightGBM-1 模型相比,MAPE 降低了 17.80%,RMSE 降低了 17.51%,MAE 降低了 20.04%。CNN-LightGBM 模型与单独 CNN 模型相比,MAPE 降低了 23.69%,RMSE 降低了 15.69%,MAE 降低了 21.53%。这表明基于深度学习的 CNN-LightGBM 模型可以更加深入地挖掘交通流的时空特性,相比于组合中的单一模型,可以明显降低预测误差。

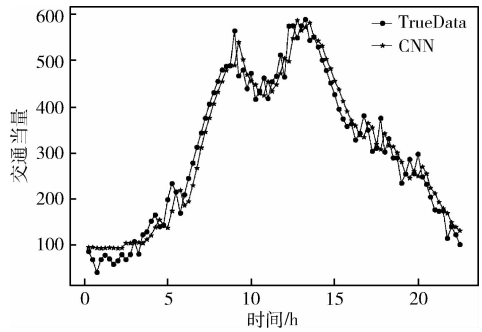

图 6　CNN 模型预测结果

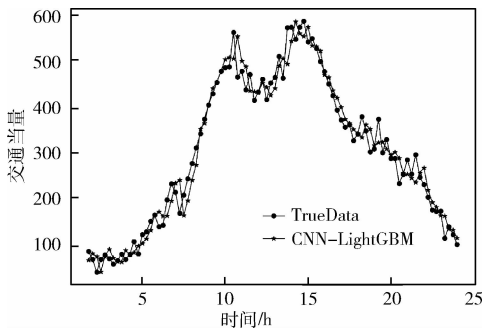

图 7　CNN-LightGBM 模型预测结果

同时,为了体现该模型与其他组合模型的优越性,本实验同时还利用了同类 CNN-Xgboost 模型与 CNN-LightGBM 模型作对比,如图 8 所示。由表 1 可以看出,同样挖掘交通量的时空特性,在本实验中,CNN-LightGBM 相比于 CNN-Xgboost,MAPE 降低了 2.89%,RMSE 降低了 2.26%,MAE 降低了 1.63%。同时在本实验中,以 CNN 提取的特征向量作为输入的 Xgboost 模型运行时间为 18.6 s,LightGBM 模型运行时间为 3.9 s,LightGBM 模型运行速度远高于 Xgboost 模型。在本实验中 CNN-LightGBM 的预测效果优于 CNN-Xgboost 模型。

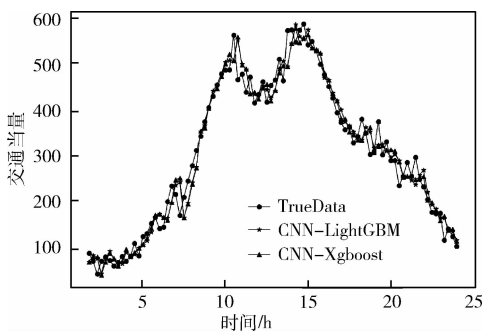

图 8　CNN-LightGBM 模型与 CNN-Xgboost 模型预测结果

为了更一步验证模型的有效性,本实验还与传统的 Xgboost 模型、KNN 模型、神经网络模型和 SVR 模型作对比,如图 9 所示。结合表 1 的预测误差,综合以上所有模型对比结果发现,考虑时空特性影响的 CNN-LightGBM 模型在本实验中可以明显降低预测误差。

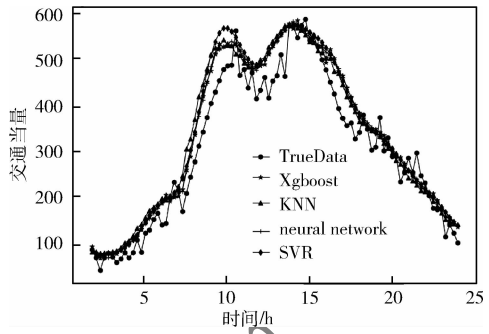

图 9　其他模型预测结果

表 1　实验结果对比

模型类型	MAPE/%	RMSE	MAE
CNN-LightGBM	11.08	29.78	23.58
CNN-Xgboost	11.41	30.47	23.97
LightGBM-1	13.48	36.10	29.49
CNN	14.52	35.32	30.05
LightGBM-2	15.45	43.08	35.49
Xgboost	15.42	43.44	34.98
KNN	15.75	43.60	35.57
神经网络	15.38	43.71	35.13
SVR	16.79	49.22	39.04

3　结束语

在高速公路系统中,相比于传统的车辆数量预测,交通当量的准确预测对交通控制具有重要的意义,可以反映出道路的实际占有情况,有助于居民出行和物流流通,对交通监管部门道路规划和交通判别具有很大参考价值。本文选取某高速公路某路段监测点数据,采用 CNN-LightGBM 算法对交通当量进行预测,结合上下游交通量与附近高速公路出入口交通量变化,使用卷积神经网络对交通当量提取空间维度数据特征,利用具有快速、低内存、高准确率特点的 LightGBM 模型对卷积神经网络提取的高维特征向量进行处理预测,可以获得较高的预测效果,相比于单独的机器学习模型和其他组合学习模型具有更低的预测误差。该模型具有很强的实用性与可靠性。但是由于采用的是高速公路某路段进行预测,并没有考虑到高速公路各级关联道

路路网的复杂性,下一步将结合路网情况进行预测分析,使得模型更加具有通用性。

参考文献

[1] 陈功. 数据挖掘技术在智慧交通中的应用[D]. 成都:电子科技大学,2016.

[2] YE Z R, ZHANG Y L, MIDDLETON D R. Unscented Kalman filter method for speed estimation using single loop detector data[J]. Transportation Research Record: Journal of the Transportation Research Board, 2006, 1968 (1): 117-125.

[3] GUO J H, HUANG W, WILLIAMS B M. Adaptive Kalman filter approach for stochastic short-term traffic flow rate prediction and uncertainty quantification[J]. Transportation Research Part C, 2014, 43: 50-64.

[4] 陈婧敏,基于 KNN 回归的短时交通流预测[J]. 微型电脑应用, 2015,31(9):25-29.

[5] WILLIAMS B M, HOEL L A. Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: heoretical basis and tempirical results[J]. Journal of Transportation Engineering, 2003, 129(6): 664-672.

[6] GUO J Y, WANG Z J, CHEN H W. On-line multi-step prediction of short term traffic flow based on GRU neural network[C]. Proceedings of the 2nd International Conference on Intelligent Information Processing, 2017.

[7] ZHAO Z, CHEN W H, WU X M, et al. LSTM network: a deep learning approach for short-term traffic forecast[J]. IET Intelligent Transport Systems, 2017, 11(2): 68-75.

[8] LEE Y J, MIN O G. Long short-term memory recurrent neural network for urban traffic prediction: a case study of seoul[C]. 2018 IEEE International Conference on Intelligent Transportation System (ITSC), 2018: 1279-1284.

[9] XIA Y, CHEN J G. Traffic flow forecasting method based on gradient boosting decision tree[C]. Proceedings of the 5th International Conference on Frontiers of Manufacturing Science and Measuring Technology, 2017: 413-419.

[10] DONG X C, LEI T, JIN S T, et al. Short-term traffic flow prediction based on XGBoost[C]. Proceedings of 2018 IEEE 7th Data Driven Control and Learning Systems Conference. New York: IEEE, 2019: 854-859.

[11] 程政,陈贤富,基于随机森林模型的短时交通流预测方法[J]. 微型机与应用,2016,35(10):46-49.

[12] 王青松,谢兴生,佘颢. 基于 CNN-XGBoost 混合模型的短时交通流预测[J]. 测控技术, 2019, 38(4): 37-40,67.

[13] 王章章. 基于机器学习的价格预测模型研究与实现[D]. 西安:长安大学,2018.

[14] FUKUSHIMA K, MIYAKE S. Neocognitron: a self-organizing neural network model for a mechanism of visual pattern recognition[M]. Cometition and Cooperation in Neural Nets. Brlin: Springer, 1982: 267-285.

[15] LE CUN Y, BOSER B, DENKER J S, et al. Backpropagation applied to handwritten zip code recognition[J]. Neural computation, 1989, 1(4): 541-551.

[16] 梁雪剑,曲海成. 不同池化模型的卷积神经网络学习性能研究[J]. 中国图象图形学报, 2016, 21(9): 1178-1190.

[17] 赵蕾. 基于卷积神经网络的快速路交通流预测研究[D]. 北京:北京交通大学,2019.

[18] KE G L, MENG Q, FINLEY T, et al. LightGBM: a highly efficient gradient boosting decision tree[C]. Proceedings of the 31st Annual Conference on Neural Informaiton Processing Systems (NIPS), LongBeach, CA, 2017.

[19] 交通运输部公路局,中国工程建设标准化协会公路分会. JTG B01-2017 公路工程技术标准[S]. 2017.

[20] 林志坚,鲁迪,林锐涛,等. 基于 k-means 聚类和变分位鲁棒极限学习机的短期负荷预测方法[J]. 智慧电力, 2019,47(3):46-53.

(收稿日期:2019-12-17)

作者简介:

张振(1995 -),通信作者,男,硕士研究生,主要研究方向:数据挖掘、大数据分析。E-mail:1903625390@qq.com。

曾献辉(1974 -),男,博士,副教授,主要研究方向:大数据挖掘、智能优化问题、决策与分析。

版权声明