

基于 K-means 与 GRU 的短时交通流预测研究

凤少伟¹, 凤超², 申浩¹

(1. 长安大学 信息工程学院, 陕西 西安 710064;

2. 新疆理工学院 机电工程系, 新疆 阿克苏 843100)

摘要:随着神经网络的蓬勃发展,如今已在短时交通流预测领域得到了广泛的应用,并且有较高的预测准确度。针对训练集的选取对短时交通流预测结果影响显著的问题,从时间序列的角度出发,提出了一种基于 K-means 与门限循环单元(gated recurrent unit, GRU)神经网络相结合的短时交通流预测方法。利用 K-means 聚类算法建立交通流模式库,根据状态向量以及数据相似性确定训练集,并利用 GRU 神经网络预测短时交通流,通过美国交通研究数据实验室的真实数据验证了该方法的有效性。实验结果显示,与经典 GRU 神经网络相比,该方法预测结果的均方根误差(root mean square error, RMSE)降低了 2.28, 平均绝对百分比误差(mean absolute percent error, MAPE)降低了 2.54%,表明该方法与传统 GRU 神经网络预测模型相比,预测结果误差明显下降。因此,基于 K-means 与 GRU 神经网络结合的交通流预测方法能够更好地挖掘交通流时间序列的关联性,可以为交通控制提供可靠的依据。

关键词:短期交通流;神经网络;K-means 聚类;GRU;预测

中图分类号:U491.2

文献标识码:A

文章编号:1673-629X(2020)07-0125-05

doi:10.3969/j.issn.1673-629X.2020.07.027

Research on Short-term Traffic Flow Prediction Based on K-means and GRU

FENG Shao-wei¹, FENG Chao², SHEN Hao¹

(1. School of Information Engineering, Chang'an University, Xi'an 710064, China;

2. Department of Mechanical and Electrical Engineering, Xinjiang Institute of Technology, Aksu 843100, China)

Abstract: With the rapid development of neural network, it has been widely used in the field of short-term traffic flow prediction with high prediction accuracy. Aiming at the problem that the selection of training sets has a significant impact on short-term traffic flow prediction results, from the perspective of time series, we propose a short-term traffic flow prediction method based on the combination of K-means and gated recurrent unit (GRU) neural network. The K-means clustering algorithm is used to establish the traffic flow pattern database, and the training set is determined by the state vector and data similarity. The GRU neural network is used to predict the short-term traffic flow, and the proposed method is verified by real data from American Traffic Research Data Laboratory. The test results show that compared with the classical GRU neural network, the root mean square error (RMSE) and the mean absolute percent error (MAPE) of the proposed method decrease by 2.28 and 2.54% respectively, indicating that the proposed method has a significant decrease in the prediction result compared with the traditional GRU neural network prediction model. Therefore, the traffic flow prediction method based on K-means and GRU neural network can better mine the correlation of traffic flow time series, which can provide a reliable basis for traffic control.

Key words: short-term traffic flow; neural network; K-means clustering; GRU; prediction

0 引言

交通流预测是智能交通系统(ITS)不可或缺的重要组成部分^[1]。特别是在城市交通干道,准确的交通流预测,能提供及时有效的未来路况信息,帮助司机和乘客躲避拥堵。因此,交通流预测已成为未来建设智能城

市的基本方面之一^[2]。

国内外众多学者在短时交通流预测领域做了很多相关研究,如张晓利等使用非参数回归方法开展了对实际道路的交通流预测^[3]。运用时间序列预测的方法来对交通流量进行预测,是一个新的思路,Box 等基于

收稿日期:2019-10-14

修回日期:2020-02-17

基金项目:陕西省自然科学基金基础研究计划项目(2017ZDJC-23)

作者简介:凤少伟(1993-),男,硕士研究生,CCF 会员(F2264G),研究方向为交通大数据分析处理。

时间序列的深入分析研究,提出了一种全新的时序模型,即自回归滑动平均(ARIMA)^[4]。解小平等首次将ELMAN神经网络应用于短时交通流预测的问题上^[5]。

目前,短时交通流预测方法可基本分为三大类:第一类是基于数理统计模型的预测方法,例如ARIMA模型、卡尔曼滤波模型等,这些方法都是寻求一种准确的交通流数学模型进行预测。然而,由于交通流的随机性和非线性特征,难以建立固定的数学模型,而且这类方法无法摆脱随机干扰的影响,对于不同路况、不同天气、不同日期的预测结果的误差难以接受^[6-8];第二类是智能交通流预测方法,如非参数回归模型、支持向量机模型、神经网络模型等,这类方法的适应性较强,可以应对诸多随机因素的干扰^[9-11];为了提高预测精度,出现了前两类方法组合预测的第三类组合模型,如采用卡尔曼滤波模型与SVM相结合的方法,期望两种方法互补,最终得到更高的预测准确度^[12]。

近年来,随着人工智能的逐步成熟,深度学习已广泛应用于语音识别、自然语言处理、图像分类等方面^[13-14]。已有学者将RNN(recurrent neural network)及其改良算法GRU网络应用于交通流预测领域,并取得了不错的预测结果,但是该方法的预测稳定性一般,对于不同天气、不同路况的道路交通流预测结果不尽如人意,而且预测精度还有待进一步提升^[15-17]。文中从时间序列的角度出发,提出一种K-means聚类与GRU网络相结合的预测方法,首先利用K-means聚类算法建立交通流模式库,然后根据状态向量及数据相似性确定训练集,最后利用GRU神经网络预测短时交通流。

1 K-means 聚类算法

K-means 算法是使用逐次迭代细化来产生最终的聚类结果。算法的输入是簇 K 和数据集的数量,数据集是每个数据点的功能集合, K 的值是随机生成或者从数据集中任意选定,在如下两个步骤之间迭代:

(1) 数据分配。

每个质心定义一个簇。在此步骤中,每个数据点分配到距其基于 2 范数的欧几里德距离最近的质心,如式(1)所示。如果 c_i 是集合 C 中的一个质心集合,则数据集中的点 x 都将被分配给一个基于质心 c_i 的类簇中。

$$\operatorname{argmin}_{c_i \in C} \operatorname{dist}(c_i, x)^2 \quad (1)$$

其中, $\operatorname{dist}()$ 是 2 范数下的欧几里德距离。

(2) 质心更新。

在此步骤中,通过计算所有数据点的均值更新质心,如式(2)所示。

$$c_i = \frac{1}{|S_i|} \sum_{x_i \in S_i} x_i \quad (2)$$

该算法不断对步骤 1 和步骤 2 进行迭代,直到没有数据点改变类簇,簇中每个数据点到质心距离的总和达到最小,或者达到最大迭代次数。

K-means 聚类必须预先设定簇的个数 K , 选取适当的 K 值才能得到理想的聚类效果,对于后续交通流预测的准确度有直接影响。文中聚类对象是不同日期的交通流时间序列数据,则选用这些数据间的欧氏距离来度量各个时间簇之间的相似度,如式(3)所示。

$$d = \sqrt{\sum_{i=1}^N (x_{1i} - x_{2i})^2} \quad (3)$$

其中, x_{1i} 、 x_{2i} 分别表示两个类簇的第 i 个值。

在评价聚类效果的优劣时,文中利用类内距离和类间距离进行评估。类内距离是指同一类各模式样本点间的均方距离,类间距离是指每类中的数据点和其他类中数据均值的欧氏距离之和。类内距离、类间距离越小,聚类的效果越优异。

2 神经网络

2.1 RNN 神经网络

RNN 神经网络,即循环神经网络,通过特点为环的连接实现存储神经元当前时刻输入与上一时刻输出间的联系。RNN 神经网络结构如图 1 所示。

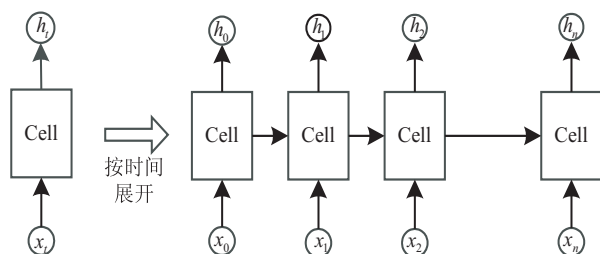


图 1 RNN 网络结构

RNN 中隐藏层节点的输入,包括上一隐藏层的输出和前一时刻该层的输出两部分,这种环结构可以存储历史信息,更有利于时序建模。RNN 网络中,每一时刻的输入,其输出又循环地在下一个时刻输入网络。理论上,网络当前输出决定于该时刻之前每一时刻的输出,但是,RNN 网络只能记忆一定时长的信息,此外,循环神经网络还会出现梯度爆炸和梯度消失的问题。

2.2 GRU 神经网络

门限循环单元(gated recurrent unit,GRU)神经网络,是基于 RNN 网络的优秀变体,它既克服了 RNN 的中长期依赖问题,而且结构简洁,训练时间和收敛速度更为优异。

GRU 结构示意图如图 2 所示。

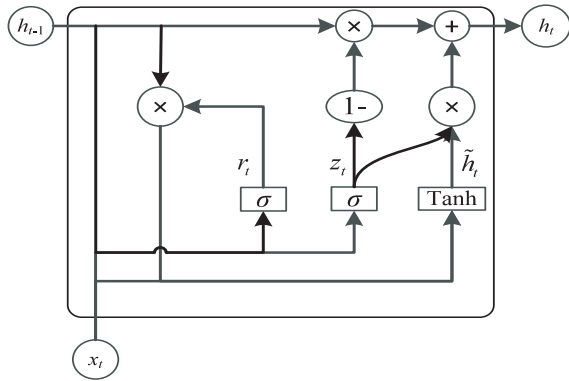


图2 GRU 结构示意图

图2中的 z_t 和 r_t 分别表示更新门和重置门。 z_t 决定前一时刻的状态信息传输到当前状态的多少, z_t 越大表明前一时刻的状态信息传入到当前状态越多。 r_t 决定前一状态有多少信息被写入到当前的候选集 h_t 上, r_t 越小,前一状态的信息被写入的越少。

单个门控单元在 t 时刻的计算过程如式(4)~式(8)所示。

$$r_t = \sigma(\omega_r \cdot [h_{t-1}, x_t]) \quad (4)$$

$$z_t = \sigma(\omega_z \cdot [h_{t-1}, x_t]) \quad (5)$$

$$\tilde{h}_t = \tanh(\omega_h \cdot [r_t * h_{t-1}, x_t]) \quad (6)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t \quad (7)$$

$$y_t = \sigma(\omega_o \cdot h_t) \quad (8)$$

其中, $[\]$ 表示两个向量相连, $*$ 表示矩阵的乘积, h_{t-1} 表示上一个神经元的输出, x_t 表示当前细胞的输入, σ 表示sigmoid函数, $\tanh(\)$ 表示双曲正切函数, ω_{\odot} 表示相应的连接权值矩阵,其余的 r_t 、 z_t 、 \tilde{h}_t 均是便于信息运算而设置的中间变量。

3 基于 K-means 与 GRU 网络的交通流预测方法

GRU 网络可以有效学习时间序列的特征,根据历史序列预测短时交通流,但是不同路况、不同天气、不同日期等因素对交通流时间序列分布影响很大。例如,由于假期旅游出行的人数激增,此时道路交通模式与工作日截然不同。因此,文中首先通过聚类分析建立多类交通流模式库,然后根据交通流数据选择交通流模式,最后利用 GRU 网络进行交通流预测。

基于聚类与 GRU 网络的短时交通流预测架构如图3所示。

依据图3架构,文中预测短时交通流的具体流程如下:

(1)利用 K-means 聚类方法建立交通流预测模式库。交通流模式库的构建对 GRU 预测结果的准确度及效率有显著影响。模式库所包含的种类需适当,既要

涵盖该条道路交通流的所有状态,又不存在过多的冗余数据。文中将以天为单位的交通流数据作为聚类对象,把历史数据划分为多类,并分别建立模式库。如此,可以区分所研究道路不同日期的交通流模式。

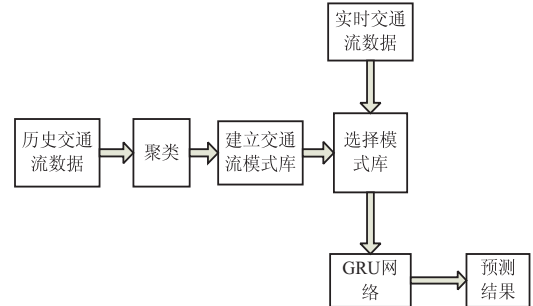


图3 基于聚类与 GRU 网络的短时交通流预测架构

(2)定义状态向量及数据相似性判别。选择所研究路段的前 24 个时刻的交通流时间序列作为状态向量,选择欧氏距离来衡量数据相似性的指标。

(3)确定用于训练的样本数据。计算待预测时间段前 24 个时刻的实时交通流数据序列与模式库中各类时间序列的相似性,选择欧氏距离最小,即相似性最大的库作为样本数据库。

(4)利用 GRU 神经网络预测选定时段的交通流量。首先利用步骤(3)中选出的样本库作为训练集训练 GRU 网络,然后将待预测时间段的前 N 个数据序列输入训练好的 GRU 网络中得到预测结果,并与真实交通流序列进行对比。

4 实例验证

4.1 数据来源

为了验证文中提出的基于聚类与 GRU 网络相结合的交通流预测效果,选取美国加州 Performance Measurement System(PeMS)数据为研究对象进行预测分析,并将预测结果与传统预测方法进行对比分析。PeMS 数据集在加利福尼亚全州范围内部署了超过 15 000 个传感器,文中随机选择 PeMS 路网中任意一个检测站的交通流进行试验,该数据集位于美国奥克兰的 Alameda,原始数据每 30 秒收集一次,根据以往的研究,5 分钟的交通流量更适合该预测。虽然随机选择的数据集存在缺失数据,但仅占整个数据集的小部分,因此使用历史平均值来估算缺失的数据点进行数据填充。文中均是基于填补以后的数据集进行的实验。

4.2 模式库构建

选取 2016 年 1 月 4 日到 2016 年 3 月 30 日的交通流数据,对其进行聚类并构建交通流模式库。从交通流的实际特征以及预测的目的出发, K 值不宜取得过大。文中分别对 K 取 3、4、5 三个数值进行 K-means 聚类分析,并且利用轮廓系数 S_i 作为评价 K 值选取优劣

的标准。轮廓系数结合了聚类的凝聚度和分离度,用于评估聚类的效果,该值处于 $-1 \sim 1$ 之间,轮廓系数值越大,表示聚类效果越好,如式(9)和式(10)所示。

$$s_i = \frac{b_i - a_i}{\max\{a_i, b_i\}} \quad (9)$$

$$S_i = \frac{1}{N} \sum_{i=1}^N s_i \quad (10)$$

其中, s_i 表示 i 向量的轮廓系数, a_i 表示向量 i 到同一簇内其他点不相似程度的平均值, b_i 表示向量 i 到其他簇的平均不相似程度的最小值, S_i 表示该聚类结果总的轮廓系数。文中所选不同 K 值下的交通流量聚类轮廓系数如图 4 所示。

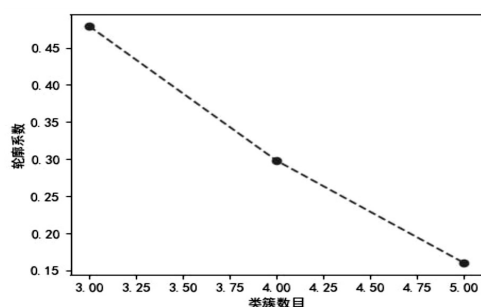
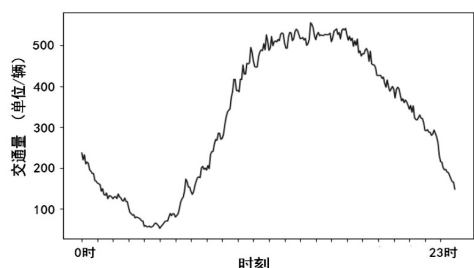
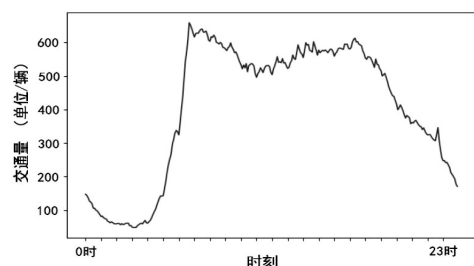


图 4 不同 K 值的轮廓系数

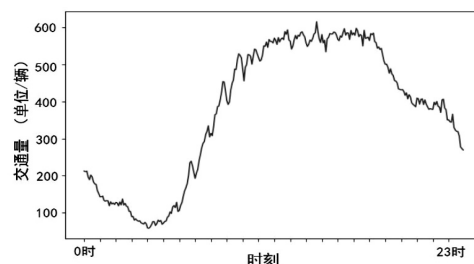
观察图 4 可知,当 $K=3$ 时,轮廓系数最大,因此将这 81 天的交通流分为 3 类模式,如图 5 所示。



(a) 第一类交通流模式



(b) 第二类交通流模式



(c) 第三类交通流模式

图 5 3 类交通流量模式图

通过比较 2016 年 3 月 31 日前 24 个时间点的交通流量与图 5 所示的 3 类交通流模式的欧氏距离,得到其与第 2 类流量模式最为接近,因此,选定第二类交通流模式中所有天数的数据作为 GRU 网络的训练集,以 2016 年 3 月 31 日的数据作为测试数据。

4.3 评价指标

均方根误差 (root mean square error, RMSE)、平均绝对百分比误差 (mean absolute percent error, MAPE) 及 R^2 _Score 是评价预测效果的 3 个重要指标。其中, RMSE、MAPE 评估预测结果的误差,其值越小表明预测结果越准确; R^2 _Score 评估预测模型的拟合优度,其值越接近于 1 表明该模型的拟合效果越好。其具体定义分别如式(11)~式(13)所示。

$$RMSE = \frac{1}{n} \sqrt{\sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (11)$$

$$MAPE = \sum_{i=1}^n \left| \frac{\text{observed}_i - \text{predicted}_i}{\text{observed}_i} \right| \times \frac{100}{n} \quad (12)$$

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=0}^{n_{\text{train}}-1} (y_i - \hat{y}_i)^2}{\sum_{i=0}^{n_{\text{sample}}-1} (y_i - \bar{y})^2} \quad (13)$$

其中, y_i 表示实际交通量, \hat{y} 表示预测交通量, n 表示预测时间点数。

4.4 预测结果分析

运用 K-means 与 GRU 网络结合的方法进行预测,其预测值与真实值的对比结果如图 6 所示,其中,实线表示 s183-E 检测站 2016 年 3 月 31 日的实际流量,虚线表示使用文中方法得到的预测流量值。

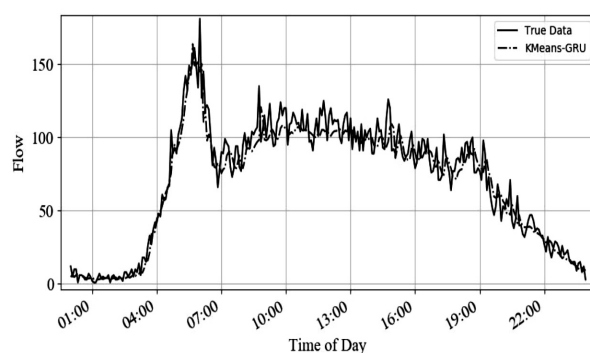


图 6 预测值与真实值对比

观察图 6 可知,使用文中方法对交通流预测的结果与实际流量吻合度很高,证明基于 K-means 与 GRU 网络的交通流预测方法具有可行性。

为了评价基于 K-means 与 GRU 网络的交通流预测方法的准确性和有效性,与传统 GRU 模型、SAEs 模型的预测结果进行对比,结果如图 7 所示。其中,实线表示真实流量,点状虚线表示 K-means-GRU 方法的预测流量,虚线表示传统 GRU 网络的预测流量, + 形实

线表示 SAEs 模型的预测流量。

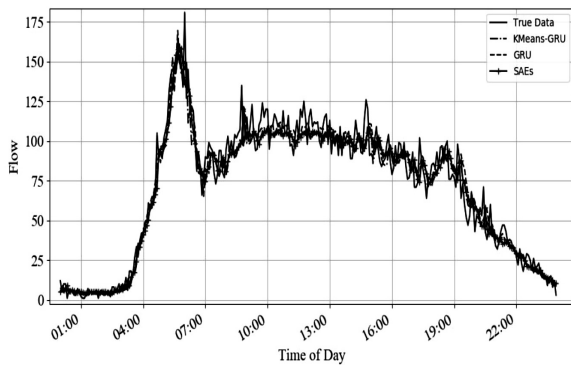


图7 K-means-GRU 与传统 GRU、SAEs 网络的预测对比

观察图7可知,与传统 GRU 网络和 SAEs 相比,基于 K-means 与 GRU 网络相结合的交通流预测方法在流量波动的细节上与实际情况更为相符。

利用评价指标分别计算 K-means-GRU、GRU 与 SAEs 三种预测模型的 RMSE、MAPE、R2_Score,如表1所示。

表1 K-means-GRU、GRU、SAEs 模型的 RMSE、MAPE、R2_Score

预测模型	RMSE	MAPE	R2_Score
K-means-GRU	7.685 6	14.24%	0.942 1
GRU	9.965 7	16.78%	0.938 9
SAEs	9.595 6	17.8%	0.923 3

由表1可知,与 GRU 网络相比,文中提出的 K-means 聚类与 GRU 神经网络结合的交通流预测方法的 RMSE 降低了 2.280 1,MAPE 降低了 2.54%,R2_Score 的值更趋向于 1,表明该方法对于交通流的预测拟合优度较好;与 SAEs 方法相比,预测误差也更小。综上所述,K-means 聚类与 GRU 神经网络结合的交通流预测方法可以更为准确、有效地预测交通流。

5 结束语

由于天气、节假日、大型事件等因素的影响,使得不同日期的交通流有着不同的规律,神经网络研究交通流分布特性时,笼统地将过去的时间序列作为训练集来训练网络,没有深入地考虑交通流的现实特点。文中提出基于 K-means 聚类与 GRU 神经网络相结合的方法,充分考虑了交通流的实际分布特性,通过 K-means 方法建立模式库,利用状态向量及数据相似性确定与历史相似性更高的数据作为训练集,有效地降

低了预测结果的 RMSE 和 MAE 值,是一种性能更好的预测方法。

参考文献:

- [1] 胡波. 数据挖掘在高速公路联网运营管理及决策上的应用探讨[J]. 中国交通信息业,2004(12):86-88.
- [2] 罗文慧,董宝田,王泽胜. 基于 CNN-SVR 混合深度学习模型的短时交通流预测[J]. 交通运输系统工程与信息,2017,17(5):68-74.
- [3] 张晓利,陆化普. 非参数回归方法在短时交通流预测中的应用[J]. 清华大学学报:自然科学版,2009,49(9):1471-1475.
- [4] BOX G E P, JENKINS G M. Time series analysis: forecasting and control[J]. Journal of Time,2010,31(4):303.
- [5] PENG K, LEUNG V C M, HUANG Q. Clustering approach based on mini batch kmeans for intrusion detection system over big data[J]. IEEE Access,2018,6(99):11897-11906.
- [6] 解小平. 基于 Elman 神经网络的短时交通流预测及应用研究[D]. 兰州:兰州交通大学,2017.
- [7] DUBEY A K, GUPTA U, JAIN S. Analysis of k-means clustering approach on the breast cancer Wisconsin dataset[J]. International Journal of Computer Assisted Radiology & Surgery,2016,11(11):2033-2047.
- [8] LIU Y, DONG S, LU M, et al. LSTM based reserve prediction for bank outlets[J]. Tsinghua Science and Technology,2019,24(1):77-85.
- [9] 陈佳维. 基于 K 近邻非参数回归方法的短时交通流预测[D]. 成都:西南交通大学,2017.
- [10] 余涛. 基于 SVM 和 BP 神经网络的短时交通流预测与实现[D]. 南京:南京邮电大学,2018.
- [11] 石睿. 基于粒子滤波与神经网络的短时交通流预测[D]. 北京:北京交通大学,2018.
- [12] 王祥雪,许伦辉. 基于深度学习的短时交通流预测研究[J]. 交通运输系统工程与信息,2018,18(1):81-88.
- [13] 崔翔鹏,黄洪琼. 基于 GA 优化 ELM 的船舶交通流预测模型[J]. 微型机与应用,2017,36(9):15-17.
- [14] 钱伟,杨慧慧,孙玉娟. 相空间重构的卡尔曼滤波交通流预测研究[J]. 计算机工程与应用,2016,52(14):37-41.
- [15] 钟足峰. 联网收费系统数据分析与挖掘的理论和实现[D]. 长沙:长沙理工大学,2007.
- [16] 刘田. ADF 与 PP 单位根检验法对非线性趋势平稳序列的伪检验[J]. 数量经济技术经济研究,2008,25(6):137-145.
- [17] GERS F A, SCHMIDHUBER J, CUMMINS F. Learning to forget: continual prediction with LSTM[J]. Neural Computation,2000,12(10):2451-2471.