文章编号:1671-1637(2013)02-0114-06

高速路交通流短时预测方法

许岩岩[1],翟 希[2],孔庆杰[1],刘允才[1]
(1. 上海交通大学 系统控制与信息处理教育部重点实验室,上海 200240;
2. 上海市城乡建设和交通发展研究院 交通信息中心研究部,上海 200032)

摘 要:针对短时交通流变化的复杂性与非线性特点,分析了分类回归树模型的建立,包括模型的生长、分裂与剪枝,研究了模型在高速路交通流短时预测中的应用,并对美国波特兰州高速路网的真实交通流量数据进行分析建模。采用 RMSE 与 MAPE 误差分析法,将试验结果与传统的交通流预测方法 ARIMA 模型与 Kalman 滤波预测模型进行比较。对比结果表明:分类回归树预测模型的 RMSE 比 ARIMA 模型与 Kalman 滤波预测模型分别降低了 42.1%、13.1%。

关键词:智能交通系统;交通流预测;数据挖掘;时间序列分析;分类回归树;Kalman 滤波器
中图分类号:U491.14 **文献标志码**:A

Short-term prediction method of freeway traffic flow

XU Yan-yan[1], ZHAI Xi[2], KONG Qing-jie[1], LIU Yun-cai[1]
(1. Key Laboratory of System Control and Information Processing of Ministry of Education, Shanghai Jiaotong University, Shanghai 200240, China; 2. Transportation Information Center, Shanghai Urban-Rural Construction and Transportation Development Research Institute, Shanghai 200032, China)

Abstract: According to the complexity and nonlinearity characteristics of short-term traffic flow, the application of classification and regression tree model in freeway traffic volume prediction was investigated, and its including growing, splitting and pruning of the model was studied. The real traffic volume data of the freeways in Portland State of US was tested and verified. Afterwards, the experimental result of model was compared with the traditional ARIMA model and Kalman filtering model by using the error analysis methods of RMSE and MAPE. Comparison result indicates that the RMSEs of tree model are 42.1% and 13.1% lower than ARIMA model and Kalman filtering model, respectively. 1 tab, 6 figs, 17 refs.

Key words: intelligent transportation system; traffic flow prediction; data mining; time series analysis; classification and regression tree; Kalman filter

Author resumes: XU Yan-yan(1987-), male, doctoral student, +86-21-34204028, xustone1987@gmail.com; LIU Yun-cai(1948-), male, professor, PhD, +86-21-34204340, whomliu@sjtu.edu.cn.

0 引 言

随着城市经济的飞速发展,传统交通模式在现代交通中遇到了越来越多的问题与挑战。为了缓解当前交通网络面临的巨大压力,如交通拥堵、交通事故等,智能交通系统(Intelligent Transportation

收稿日期:2012-11-09
基金项目:国家 863 计划项目(2012AA112307);上海市科委科技攻关项目(11231202801)
作者简介:许岩岩(1987-),男,山东泰安人,上海交通大学工学博士研究生,从事智能交通系统研究。
导师简介:刘允才(1948-),男,上海人,上海交通大学教授,工学博士。

System，ITS）被引入到动态交通管理中，并得到了快速发展。作为智能交通系统中的重要组成单元，短时交通流预测在城市交通管理系统（Urban Transportation Management，UTM）中始终占有重要地位，对于更好地分析路网交通状况，优化交通网络规划和控制策略都有十分重要的意义。

近几十年来，很多研究人员致力于交通流数据的分析，并建立了多种交通流预测模型。一般来说，已有的交通流预测方法可以分为两类：基于时间序列的预测方法与基于时空联合的预测方法。基于时间序列的预测方法将交通流看作是单一的时间序列，利用信号处理的方法对交通流数据进行分析和预测。早在 20 世纪 90 年代，Voort 等就研究了 ARIMA 模型在交通流预测上的应用[1]；近年来，Williams 等研究了交通流的周期性，并利用季节性 ARIMA 模型进行短时预测[2]。此外，Kalman 滤波方法也在交通流的预测问题上得到了广泛应用[3]。陈相东等研究了基于局部多项式拟合的交通流预测方法[4]；Zhang 等将模糊逻辑系统理论应用于交通流预测[5]。随着智能计算的发展，神经网络也越来越多的被用于交通流预测。Zheng 等研究了使用贝叶斯模型与神经网络模型联合对高速路交通流进行短时预测的方法[6]。时空联合的预测方法大多是考虑了上游交通路段对当前路段交通流的影响，将上游路段与当前路段的历史交通数据作为模型的输入，来计算当前路段的交通流预测值，例如：矢量 ARIMA 的空间模型将路段上游观测点的交通流量利用速度关系和当前观测点联系起来[7]；基于贝叶斯网络的预测方法用贝叶斯图模型学习上下游交通流量的关系模型[8]；基于上游交通数据的多层感知（Multilayer Perceptron，MLP）预测方法[9]；基于高斯过程的交通流预测模型[10]等。此外，采用交通流的空间推理模型也应用于交通流的预测[11]。以上基于统计学习的方法模型复杂，需要优化的参数过多，计算量较大，不适合交通系统的实时计算。

目前的方法大多基于时空关系预测当前路段或者某个观测点的未来交通流，但是在实际应用中，往往会遇到这样一个问题，该路段或者观测点并没有上游观测值的存在，只能利用本观测点的历史信息进行时间序列上的预测。在 Xu 等的研究中[11]，就需要首先预测输入整个交通网络的流量，因此，针对时间序列上的交通流预测目前依旧是至关重要的。作为典型的非线性时间序列，分类学习的方法也被越来越多使用于交通流的预测。基于对交通流数据进行训练、预测的思想，本文使用分类回归树模型进行高速路交通流的短时预测。首先，对已有历史交通流数据进行分析学习，建立一个完备的模型树，使模型树中能够尽量完整地包含历史数据与预测数据之间的非线性关系；其次，把当前交通数据作为模型的输入值，预测未来 15 min 的交通流量。在试验过程中采用了美国波特兰州高速路网的真实数据对提出的模型进行测试。

1　分类回归树模型简介

在数据挖掘与机器学习中，回归算法是解决预测问题的一种有效的核心工具。通过比较不同算法在回归问题中建立的学习模型，可将这些算法分为基于线性模型的回归算法、基于 K 近邻（K-nearest Neighbor，KNN）模型的回归算法、基于树模型的回归算法和一些其他模型算法（人工神经网络 ANN、支持向量机 SVM 等）。其中 K 近邻模型[12-13]、人工神经网络[14]、支持向量机[15]等已在智能交通系统中得到了广泛应用，其中樊娜等将非参数回归模型和 BP 神经网络模型组合为混合预测模型对短时交通流进行预测[16]。作为机器学习中一种经典的基于决策树模型的数据处理方法，分类回归树（Classification and Regression Tree，CART）通过递归划分，可以用于对连续变量进行预测（回归）以及对不连续的变量类别进行预测（分类）。在本文交通流短时预测问题研究中，采用分类回归树模型对每 15 min 的交通流量数据进行预测。

经典的 CART 模型是由 Breiman 等于 1984 年提出并推广的一种决策树分类方法，是对数据空间进行递归划分并针对每一个节点建立一个回归预测模型。在回归树中，以序列的均方误差作为是否对序列继续划分的标准。CART 是基于下面 2 个关键的思想：第 1 个是关于递归划分自变量空间的想法；第 2 个是为了防止过拟合的发生，用验证数据进行剪枝。

1.1　分类回归树的结构

递归划分或者树的生长策略主要分为 3 个问题：首先是选择分类变量的标准；其次要找到被选择变量的分裂点的标准；最后是确定合适停止树生长过程的标准。交通流量的预测问题实质上是找到一组由输入变量 X 和输出变量 y 组成的对应关系：在参数模型中往往是一个固定的函数拟合；在非参数模型中是一系列的数学关系模型。本文中假定 X

为模型中的自变量，即模型的输入变量，y 为目标变量(模型的输出，即需要预测的 $t+1$ 时刻的交通流量 v_{t+1})，在交通流预测问题上，本文将 X_t 定义为 t 时刻的交通流量

$$X_t = \left\{ v_{t-i} \mid i = 0,1,\cdots,d \right\} \tag{1}$$

式中：v_{t-i} 为 $t-i$ 时刻的交通流量。

对于包含有 N 个训练样本的训练样本库 Γ，可以定义为

$$\Gamma = \left\{ (X_t, y_t) \mid t = 1,2,\cdots,N \right\} \tag{2}$$

本文的目标是在给定一个新的自变量 X 后再预测 y 的值，这就需要不同的划分原则。理论上，一个很简单的做法就是把训练样本库 Γ 根据其不同数值特征划分为 K 个不同的子集 A_1、A_2、…、A_K，当 j 时刻新的交通状态 X_j 根据其数值特征被判定为属于集合 A_k 时，利用集合 A_k 所对应的规则来计算出 y_j，并将其作为预测值，这就是回归树用于序列预测的基本思想[17]。

分类回归树是通过依次把训练样本库 Γ 按照一定规则分裂成两部分而建立起来的，结构见图 1，其中：m_1 为树的根节点，包含全部训练样本 Γ；节点 m_2、m_3、m_5、m_7 为树模型中的中间节点；m_4、m_6、m_8、m_9、m_{10}、m_{11} 为叶节点，对应于最终分裂后的集合 A_1、A_2、…、A_K，即表示该节点的纯度已经达到预定要求，不再往下分裂。图 1 中子节点 m_2、m_3 由父节点 m_1 分裂而成，同时又作为父节点继续分裂，直至节点纯度达到预定要求。

1.2 分类回归树分裂准则

在分类回归树的生长过程中，节点的分裂准则决定了当前节点是否继续分裂为 2 个子节点。在模型建立之前，首先设定合适的决定节点是否分裂的纯度阈值 δ 与最小样本数量 n，使得树模型能够准确、完整的对训练数据进行分类。

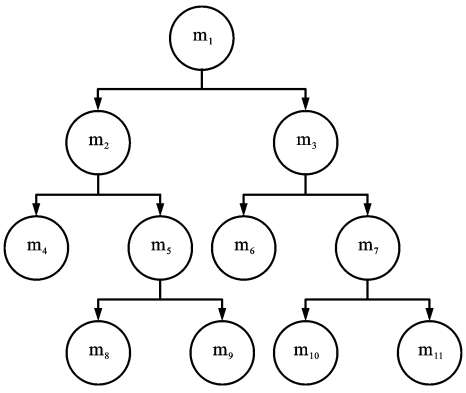

图 1 分类回归树模型结构

Fig. 1 Structure of classification and regression tree model

Step 1：选取特征值，从训练样本数据的 $d+1$ 个特征中选取合适的特征值，使在该特征值下的分裂能最大化地对节点上的样本进行分类。

Step 2：搜索样本空间，找到使下一代子节点中数据集的纯度最小的最优分裂变量，将本节点数据序列分为 2 个子集。

Step 3：在第 l 个特征值下，计算数据序列纯度是否小于 δ，或者样本数量是否小于 n，确定是否进行分裂。其中，使用序列在归一化之后的方差来计算纯度。

Step 4：如不能继续分裂，则将该节点记为叶节点，进而使用该叶节点样本集合的自变量 X 与目标变量 y 建立线性拟合模型 $f(X)$，如还需要继续分裂则转到 Step 1。

本文中的纯度阈值采用均方误差 MSE 来计算

$$\delta = (y - \hat{y})^2 \tag{3}$$

式中：$\hat{y}$ 为 y 的估计值。

1.3 分类回归树剪枝

当一个数据集在建立树模型时产生了大量的叶节点，就有可能发生过拟合现象，从而导致模型失效。本文采用交叉验证的思想来判定节点的数据集是否过拟合，这一判定并消除过拟合现象的方法称之为分类回归树的剪枝。

树模型的剪枝可以在训练过程中进行也可以在树生长停止后进行，例如 Chi-平方自动交互检测就是通过显著性去评估对某一节点分裂是否能提高分类纯度，如果不能显著提高，就不进行划分，然而，在树训练过度前就停止并不一定是最好的方法。在 CART 算法中：首先将所有样本数据划分为训练集与测试集，在训练集基础上根据预设参数纯度阈值与最小样本数量生成一个树模型；其次，利用测试集数据测试建好的模型树，依次消除叶节点并计算误差，如果消除该节点可以降低树模型在测试集上的误差，就将该叶节点剪枝。具体算法流程如下。

Step 1：在测试集上执行树模型。

Step 2：如果任何一个子节点是树，在树节点上执行 Step 1，否则执行 Step 3。

Step 3：消除 2 个子节点，并在测试集上进行测试，计算预测误差 e_s。

Step 4：计算消除节点之前的误差 e_o，如若 e_o 不小于 e_s，则消除 2 个子节点，即对该节点进行剪枝，否则保留。

2　数据分析与模型建立

为了测试分类回归树模型在交通流预测问题上的应用效果，本文采用了波特兰州立大学 Portal 平台发布的美国波特兰州高速路网交通数据来测试。该交通流量数据是通过在高速路的每个车道上安装检测线圈来获取的，每 20 s 计算一次车流量与平均车速。本文采集了波特兰州高速公路 I-84 上编号为 33 的交通信息检测站所获取的车流量信息，见图 2。计算每 15 min 内的车流量，训练样本数据持续 4 周，从 2011-05-01 的 0:00 至 2011-05-29 的 0:00，共计 28 天 2 688 个样本数据，见图 3。测试样本数据持续一周，从 2011-05-29 的 0:00 至 2011-06-05 的 0:00，共计 7 d 672 个样本数据。为了真实反映道路交通拥挤状况，模型输入的所有流量数据均转换为每小时的平均车流量(veh·h^{-1})。

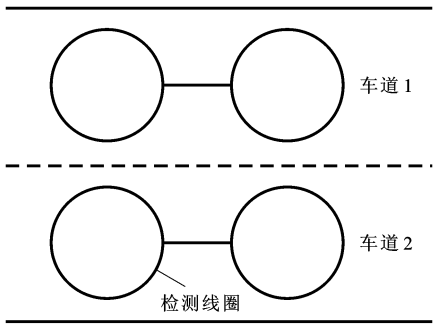

图 2　交通流量检测线圈

Fig. 2　Detecting coils of traffic flow

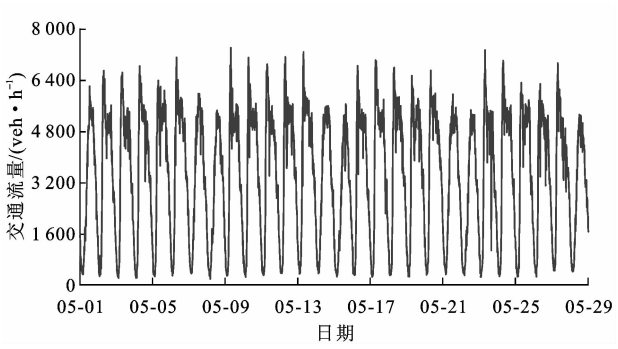

图 3　训练样本数据

Fig. 3　Training sample data

在建立模型之前，将时间序列数据变换成交通状态向量空间，使得 2 688 个训练数据样本就转换为一个(p+2)×(2 688−p−1)的样本库矩阵。矩阵的每一行(v_t，v_{t-1}，…，v_{t-p}，v_{t+1})由 X_t 与 y_t 组成，即样本库矩阵的前 p+1 列是自变量 X，最后一列表示因变量 y。而测试集则是(p+2)×(672−p−1)的矩阵，在模型测试中，只有前 p+1 列的矩阵被输入到训练好的模型中，最后一列作为真实值来测试分类回归树预测模型的准确度。

CART 模型建立的预设纯度阈值为 2，叶节点最小样本数量为 20。即当节点的误差下降小于 2 或者该节点含有状态向量数目不大于 20 时，将此节点标记为叶节点并停止分裂。在结束树模型剪枝之后，对每个叶节点的自变量 X 与目标变量 y 进行线性拟合，得到线性函数关系式

$$y = f(X) \tag{4}$$

这样就形成了以线性模型为叶节点的模型树。在预测过程中，首先将输入的新的测试数据使用树模型进行分类，得到与之匹配的叶节点。然后将该叶节点对应的线性模型应用于输入数据 X_t，产生预测数值 $\hat{y}_t$。

3　试验结果分析

为了评估分类回归树模型在交通流短时预测上的效果，本文同时采用了传统的时间序列预测方法：差分自回归移动平均模型(Autoregressive Integrated Moving Average，ARIMA)和 Kalman 滤波预测模型对相同的测试序列进行预测。并采用均方根误差(RMSE)与平方绝对百分比误差(MAPE)2 种误差评估方法。RMSE、MAPE 计算如下

$$E_{\mathrm{RMSE}} = \left[\frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_t)\right]^{1/2} \tag{5}$$

$$E_{\mathrm{MAPE}} = \frac{1}{T}\sum_{t=1}^{T}\frac{|y_t - \hat{y}_t|}{y_t} \tag{6}$$

式中：T 为测试样本数量，即需要预测的交通流的数目；y_t 为交通流量的真实值。

使用 CART 模型对一周的交通流量进行预测，结果见图 4。根据文献[12]中非参数方法的试验结果，在取当前交通状态的 20 个历史状态邻域，可以得到满意的试验结果。同样，将树模型的

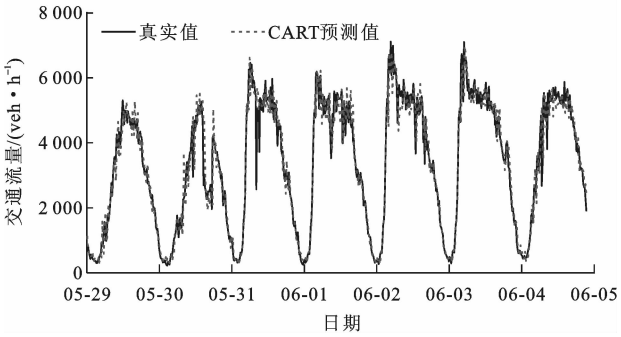

图 4　CART 模型预测结果

Fig. 4　Prediction result of CART model

每个叶节点中含有状态向量的个数限制在至少20个，以便进行线性回归。同时，在预剪枝中，限制每个节点当前误差与在分裂之后的最小误差的差小于100时停止剪枝，将该节点标记为叶节点。图5为使用另外2个经典的时间序列模型来对同样的交通流量数据进行预测的结果，可以看出，基于CART模型的预测方法能够很好匹配真实值，虽然在某些极端情况下不能很好的做出正确预测。

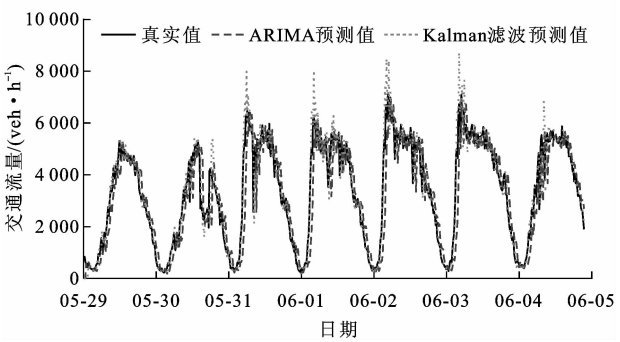

图5 两种模型的预测结果比较

Fig. 5 Prediction results of 2 models

为了对比模型对交通流复杂变化的预测性能，将6月1日(周三)的预测结果单独提取出来，见图6。可以看出，由于是固定参数模型，ARIMA模型鲁棒性不好，预测结果滞后严重，特别是在4:00～6:00这一段流量上升期。Kalman滤波模型在大部分时段表现良好，但是在一些流量较大时段，如6:00～6:30期间，预测结果不稳定，超出真实值约30%。相比较而言，CART模型的预测结果较为稳定，不论是在早高峰的上升期还是在晚高峰的下降期，都能很好地估计真实值的变化。

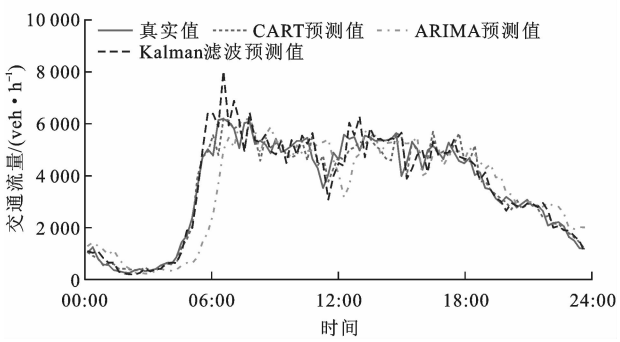

图6 6月1日预测结果比较

Fig. 6 Prediction results comparison of June 1

表1为分别对3种模型预测值与真实值进行误差计算的结果，可以精确的说明CART模型预测的效果。从对比结果中可以看出基于CART模型的预测方法在测试的一周数据中有4天的RMSE和MAPE均低于ARIMA与Kalman滤波模型。通过一周平均RMSE的比较来看，CART预测模型的精度较ARIMA模型提高了42.1%，较Kalman滤波模型提高了13.1%。由此看出基于CART的预测模型是一种有效的交通流预测方法。

表1 三种模型的预测误差对比

Tab. 1 Prediction error comparison of 3 models

日期	RMSE			MAPE		
	CART模型	ARIMA模型	Kalman滤波模型	CART模型	ARIMA模型	Kalman滤波模型
5月29日	169.76	235.11	133.54	0.118	0.176	0.094
5月30日	273.16	381.20	266.98	0.167	0.307	0.151
5月31日	253.50	474.94	347.96	0.110	0.266	0.145
6月1日	229.86	458.04	278.51	0.111	0.271	0.131
6月2日	255.58	469.94	309.38	0.105	0.234	0.125
6月3日	195.67	427.81	262.34	0.083	0.215	0.103
6月4日	244.47	312.53	267.73	0.124	0.203	0.107
平均值	231.71	394.22	266.63	0.117	0.234	0.123

4 结 语

本文提出了基于分类回归树模型的道路交通流量预测方法，使用大量的训练样本数据建立了分类回归树模型。在初始模型建立之后，为了防止模型的过拟合，采用了树模型剪枝的方法进行修正。在树模型建立之后，在每个叶节点上利用训练样本数据中的自变量与因变量，拟合得到线性模型，从而达到在输入新的自变量后可以对其进行预测的目的。为了验证本文模型的有效性，采用美国波特兰州某高速路的交通流量数据进行验证，预测15 min后的交通流量，并与经典的时间序列方法，ARIMA与Kalman滤波预测模型进行了比较。最终结果表明，基于CART的预测模型

可以有效地对交通流量进行短时预测，尽管基于CART的模型在高速路网交通流量预测上取得了不错的效果，但本文在叶节点上采用的是最简单的线性模型，下一步将探索其他非线性模型，使其更好地预测更为复杂的交通状态，如城市路网的交通流量。

参考文献：

References：

[1] VOORT M, DOUGHERTY M, WATSON S. Combining kohonen maps with ARIMA time series models to forecast traffic flow[J]. Transportation Research Part C: Emerging Technologies, 1996, 4(5): 307-318.

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

[3] XIE Yuan-chang, ZHANG Yun-long, YE Zhi-rui. Short-term traffic volume forecasting using Kalman filter with discrete wavelet decomposition[J]. Computer-Aided Civil and Infrastructure Engineering, 2007, 22(5): 326-334.

[4] 陈相东，张　勇. 基于局部多项式拟合的交通流预测[J]. 计算机工程与应用，2012，48(19)：238-242.
CHEN Xiang-dong, ZHANG Yong. Predication of traffic flow based on local polynomial fitting[J]. Computer Engineering and Applications, 2012, 48(19): 238-242. (in Chinese)

[5] ZHANG Yun-long, YE Zhi-rui. Short-term traffic flow forecasting using fuzzy logic system methods[J]. Journal of Intelligent Transportation Systems, 2008, 12(3): 102-112.

[6] ZHENG Wei-zhong, LEE D H, SHI Qi-xin. Short-term freeway traffic flow prediction: Bayesian combined neural network approach[J]. Journal of Transportation Engineering, 2006, 132(2): 114-121.

[7] MIN W, WYNTER L. Real-time road traffic prediction with spatio-temporal correlations[J]. Transportation Research Part C: Emerging Technologies, 2011, 19(4): 606-616.

[8] SUN Shi-liang, ZHANG Chang-shui, YU Guo-qiang. A Bayesian network approach to traffic flow forecasting[J]. IEEE Transactions on Intelligent Transportation Systems, 2006, 7(1): 124-132.

[9] VLAHOGIANNI E I, KARLAFTIS M G, GOLIAS J C. Spatio-temporal short-term urban traffic volume forecasting using genetically optimized modular networks[J]. Computer-Aided Civil and Infrastructure Engineering, 2007, 22(5): 317-325.

[10] XIE Yuan-chang, ZHAO Kai-guang, SUN Ying, et al. Gaussian processes for short-term traffic volume forecasting[J]. Transportation Research Record, 2010(2165): 69-78.

[11] XU Yan-yan, KONG Qing-jie, LIN Shu, et al. Urban traffic flow prediction based on road network model[C]// IEEE. The 9th IEEE International Conference on Networking, Sensing and Control. Beijing: IEEE, 2012: 334-339.

[12] SMITH B L, WILLIAMS B M, OSWALD R K. Comparison of parametric and nonparametric models for traffic flow forecasting[J]. Transportation Research Part C: Emerging Technologies, 2002, 10(4): 303-321.

[13] 于　滨，邬珊华，王明华，等. K 近邻短时交通流预测模型[J]. 交通运输工程学报，2012，12(2)：105-111.
YU Bin, WU Shan-hua, WANG Ming-hua, et al. K-nearest neighbor model of short-term traffic flow forecast[J]. Journal of Traffic and Transportation Engineering, 2012, 12(2): 105-111. (in Chinese)

[14] DOUGHERTY M S, COBBETT M R. Short-term inter-urban traffic forecasts using neural networks[J]. International Journal of Forecasting, 1997, 13(1): 21-31.

[15] MANOEL C N, JEONG Y S, JEONG M K, et al. Online-SVR for short-term traffic flow prediction under typical and atypical traffic conditions[J]. Expert Systems with Applications, 2009, 36(3): 6164-6173.

[16] 樊　娜，赵祥模，戴　明，等. 短时交通流预测模型[J]. 交通运输工程学报，2012，12(4)：114-119.
FAN Na, ZHAO Xiang-mo, DAI Ming, et al. Short-term traffic flow prediction model[J]. Journal of Traffic and Transportation Engineering, 2012, 12(4): 114-119. (in Chinese)

[17] 张立彬，张其前，胥　芳，等. 基于分类回归树(CART)方法的统计解析模型的应用与研究[J]. 浙江工业大学学报，2002，30(4)：315-318.
ZHANG Li-bin, ZHANG Qi-qian, XU Fang, et al. Research and application of the statistical models based on CART[J]. Journal of Zhejiang University of Technology, 2002, 30(4): 315-318. (in Chinese)