一种基于非参数回归的交通速度预测方法

史殿习 丁涛杰 丁 博 刘 惠
(国防科学技术大学计算机学院 长沙 410073)

摘 要 非参数回归模型是近年来提出的一种交通状态预测模型。为进一步提高预测精度,基于非参数回归模型的特点,针对近邻状态的选取问题,提出了基于速度变化趋势和密集度的变K近邻精确搜索策略,对原有模型的近邻匹配方式进行了改进和优化,进而提出了一种短时交通平均速度预测模型。利用北京市浮动车系统数据对算法精度进行了验证,结果表明,该模型的预测精度优于基础的非参数回归和BP神经网络模型,并能为短时交通速度预测提供可行的结果。

关键词 非参数回归,速度预测,短时交通状态
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.2.047

Traffic Speed Forecasting Method Based on Nonparametric Regression

SHI Dian-xi DING Tao-jie DING Bo LIU Hui
(College of Computer, National University of Defense Technology, Changsha 410073, China)

Abstract Non-parametric regression model is a traffic forecasting model proposed in recent years. Based on the characteristics of the model, in order to improve the forecasting precision on the issue of neighboring states selection, the original neighbor matching was optimized by the classification of the trend of speed and varying K neighbors precise search strategy based on intensity, and then a short-term traffic speed forecasting model was proposed. Floating car data in Beijing was used in the experiments. Results show that the optimized model is better than normal non-parametric regression model and BP neural network model, and can provide practical speed for short-term traffic prediction.

Keywords Non-parametric regression, Speed forecasting, Short-term traffic state

1 引言

智能交通系统通过对城市道路进行交通控制以及实时交通诱导,有效缓解了城市交通拥堵现象[1]。而实现该功能系统的关键是对交通状态进行准确可靠的短时预测,即利用历史和实时的交通信息来对未来短时间(一般小于15分钟)内的交通状态进行预测。

从20世纪90年代开始到现在,学者们利用各自领域的学科方法提出了多种短时交通状态预测模型[2],根据使用方法的不同,其主要可分为基于交通仿真模型预测、基于统计分析与信息处理预测以及基于数据挖掘预测这3类预测方法。基于交通仿真模型的预测方法[3],通过对交通系统的交通流量、道路网络以及信号控制规律进行模拟仿真,从而得到道路网络可能性较高的状态与变化趋势,一般用于大型仿真系统。基于统计分析与信息处理的预测方法,根据统计学原理来选择合适的模型,并对模型的参数进行调节、优化,利用训练好的模型来对交通状态进行预测,主要有ARIMA模型[4]、马尔可夫链模型[5]以及卡尔曼滤波模型[6,7]等。基于数据挖掘的预测方法主要利用数据挖掘领域相关知识对历史数据进行分析推理,根据交通系统的历史回归性对未来交通状态进行预测,目前得到广泛使用的有支持向量机回归模型[8]、神经网络模型[9]以及非参数回归模型[10,11]等。

上述3类方法中,基于数据挖掘的预测方法设计较为简单,不需要精确建模,并且预测精度相对较高,是目前短时交通预测领域的研究热点[12]。神经网络模型预测精度较高,但存在收敛速度慢且易收敛于局部最优值、没有确定隐层节点数的一般方法等缺点,通常与其它的预测理论或模型相结合组成混合预测模型[13-15];支持向量机回归模型对小样本情况下非线性问题具有很好的拟合效果,但是参数设置复杂;非参数回归直接从历史数据中挖掘信息,保持了数据变化的随机性和不确定性,非常适合用于短时交通状态的预测[16],同时又不需设置、训练参数,简洁有效。

本文基于非参数回归模型的特点,为进一步提高预测精度,针对近邻状态的选取问题,对原有模型的近邻匹配方式进行了改进和优化,提出了一种短时交通平均速度预测模型,最后利用北京市浮动车系统数据对算法精度进行了验证,验证结果表明改进的模型在预测精度上有了一定的提高。

2 非参数回归预测模型及其改进

2.1 非参数回归预测模型

非参数回归的预测过程包括历史数据准备、状态向量和

到稿日期:2015-01-22 返修日期:2015-05-18 本文受国家自然科学基金(91118008)资助。
史殿习(1966-),男,博士,研究员,硕士生导师,主要研究方向为分布计算技术、普适计算、移动感知,E-mail:dxshi@nudt.edu.cn;**丁涛杰**(1990-),男,硕士,主要研究方向为分布计算技术;**丁 博** 男,助理研究员,主要研究方向为分布计算技术、普适计算、移动感知;**刘 惠** 女,副研究员,主要研究方向为分布计算技术、普适计算、移动感知。

相似机制定义(用于度量计算)、近邻匹配机制以及预测函数的选取[17]。

2.1.1 历史数据库的准备

历史数据库准备是整个模型非常重要的一步,历史数据质量的好坏直接影响整个模型的预测效果。由于非参数回归方法是从历史数据中寻找当前的近邻状态,历史数据越丰富,包含的交通状态越全面,在近邻匹配时就能找到越多可靠的近邻信息,从而得到更准确的预测结果。同时当有新的状态数据来临时,历史数据库需要实时更新。

2.1.2 状态向量和相似机制定义

在交通预测领域,状态向量是指能描述当前交通状态的相关因素集合,也是相似机制具体判断的对象。在现实交通系统中,可以使用多种相关因素来描述当前状态,比如路段速度、天气情况以及相邻路段的交通状况等。相似机制是用来判断当前状态与历史状态的相似度的度量方法,一般根据状态向量采用状态与状态之间的空间距离来度量。

2.1.3 近邻搜索机制

近邻搜索是整个模型的核心问题,它需要从历史数据库中快速地搜索并获取与当前状态相近邻的状态。当历史数据规模比较庞大时,如何高效获取准确近邻状态是非参数回归模型面临的最严重问题,直接影响到整个算法的预测精度和实时性处理能力。常见近邻搜索机制有 K 近邻法和核近邻法等[18,19]。核近邻法是指在以当前状态点为核心、以 R 为半径的球形范围内选取状态点作为当前状态的近邻;K 近邻法是指将历史数据库中的历史状态与当前状态做比较,从中选取离当前状态最相近的 K 个状态作为近邻状态。

2.1.4 预测方法选择

预测方法是影响最终预测精度的重要因素。它根据近邻匹配获得的最近邻状态采用一定的预测方法对未来时段的交通状态进行预测。

2.2 非参数回归预测模型的应用与改进

2.2.1 模型改进思路

由于基本非参数回归模型存在效率不高以及预测精度提升困难的问题,在实际应用中,一般通过加入辅助搜索库和误差反馈机制对模型进行改进[11,16],具体框架如图 1 所示。本文也基于这个思路对模型进行改进,在近邻搜索阶段根据改进方法加入索引信息库以提高搜索精度,根据预测结果与下一时段真实值之间的误差情况对相似度量中的参数进行调整。

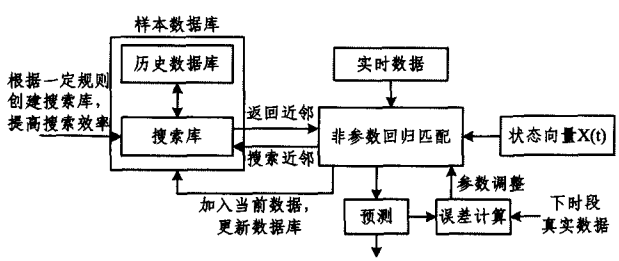


图 1 非参数回归模型应用改进框架

2.2.2 历史数据库设置

由于相邻路段之间的速度变化没有必然联系,因此根据目标路段当前时段的速度变化情况来从历史数据库中选取近邻状态,这里选取连续 4 个时段(包含当前时段)的速度序列作为状态向量,即 $X(t)=[v(t),v(t-1),v(t-2),v(t-3)]$,并根据上述状态向量建立历史样本数据库。选取加权修正空间距离作为相似度量标准,则相似度量计算方法如下:

$$dis=\sqrt{\begin{array}{l}a_1(v(t)-v_h(t))^2+a_2(v(t-1)-v_h(t-1))^2+\\ a_3(v(t-2)-v_h(t-2))^2+a_4(v(t-3)-v_h(t-3))^2\end{array}} \quad (1)$$

其中,$v_h(t)$表示历史数据中 t 时段的路段交通速度,a_1、a_2、a_3、a_4 分别代表状态向量中各部分因素所占的权重,也是后面误差反馈进行参数调节的对象。

2.2.3 对近邻搜索的改进

目前,近邻搜索时采用较多的是 K 近邻法,其具有较好的近邻搜索结果。本文同样采用 K 近邻法,并在近邻搜索阶段做如下两个方面的改进。

(1)基于交通速度变化趋势的 K 近邻搜索策略

在进行 K 近邻搜索时,状态向量中一般都存在过去相邻时段的状态信息,而在选取近邻状态时大多只考虑了当前状态与历史状态的相似度量大小。这种方式在一定程度上忽视了过去时段与当前时段状态之间的内在联系,可能搜索到的近邻状态在相似度量上差异较小,但实则与当前状态相似度并不高。

为此,引入了基于交通速度变化趋势的近邻搜索思想,目的在于提高近邻搜索结果的准确性,同时在一定程度上也对近邻搜索范围进行缩小。目前在具体的分类过程中,本文只考虑速度的变化趋势,不考虑变化趋势的剧烈程度。假定相邻两个时段的路段交通速度不可能相同,那么根据设定的状态向量,速度变化趋势可分为如图 2 所示的 8 种情况,例如,图 2 中的箭头分别表示了相邻 4 个时段的速度变化情况,上升箭头表示速度升高,下降箭头表示速度降低。

在进行近邻搜索时,先根据当前时段的状态向量进行速度变化趋势分析,然后在根据分类情况到相应的历史数据范围内(实现根据速度变化趋势对历史数据库进行划分)进行近邻搜索。

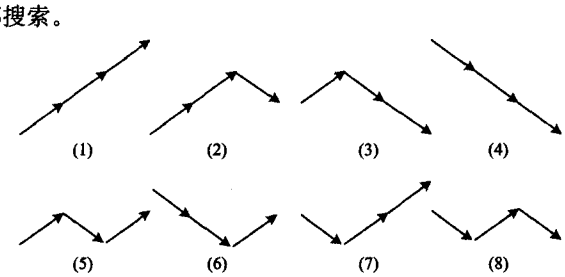


图 2 速度变化趋势分类情况

(2)基于密集度的变 K 近邻状态搜索方法

K 近邻搜索首先要确定合适的 K 值,K 值过大或过小都会影响模型最后的预测精度。目前常用的方法是先通过历史数据计算出预测值与实际值之间的误差随 K 值增大的变化情况,然后再从中选取最优 K 值。例如,假设交通速度预测平均误差随 K 值的变化情况如图 3 所示,依图可知当 $K=9$ 时预测平均误差达到最小值,则该路段预测模型的最优 K 值为 9。

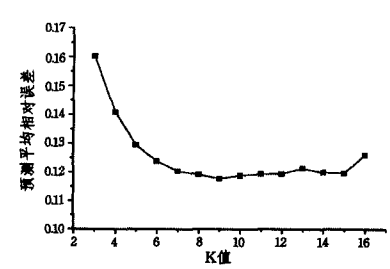


图 3 交通速度预测平均误差随 K 值的变化曲线

但对于同一路段，不同时刻下最优 K 值往往是不同的。为进一步提高预测精度，需要动态地改变 K 值，快速获取当前时段下的最优 K 值大小。本文通过引入状态密集度的概念来对历史数据库中的历史状态进行密集度计算，在近邻搜索时只搜索与当前状态最相似的历史状态，最后根据最相似历史状态的密集度来确定 K 值的选择。

状态向量一般包含多个要素，那么可将状态向量映射为由这些要素组成的多维空间中的点。数据库中历史状态映射至多维空间后，本文把以各状态映射点为中心、R 为半径的范围内的其余状态映射点的个数定义为该历史状态的密集度；同时，将这些状态定义为该状态的密集度状态。例如某时段的空间映射结果如图 4 所示，则目标状态的密集度为 5。

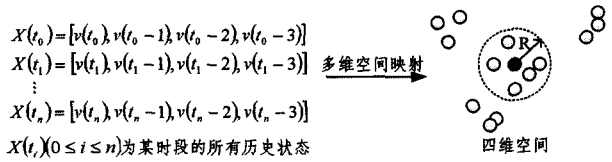


图 4　状态密集度示意图

由于历史状态分布和密集度参数设置的不同，最相似状态的密集度有可能存在极小或极大的问题，如果直接将最相似状态的密集度作为 K 值，则会在很大程度上影响模型预测的精度和稳定性。为此，根据非参数回归预测值与实际值之间的误差随 K 值增大的变化情况对 K 值的取值范围[K_{min}，K_{max}]作一个事先的设定。例如，图 3 中根据变化曲线可见 K 值从 7 增加到 15 时，预测误差变化较缓且存在最小值，由此可将 K 的取值范围定为[7,15]。确定了 K 值的取值区间之后，将最相似状态的密集度(记为 M)与之相比较以选择最终的 K 值，具体如下：

$$\begin{cases} M<K_{min}\text{时}, & \text{取 } K=K_{min} \\ K_{min}\leqslant M<K_{max}\text{时}, & \text{取 } K=M+1 \\ M\geqslant K_{max}\text{时}, & \text{取 } K=K_{max} \end{cases} \tag{2}$$

同时在近邻状态选取时，不再与数据库中的历史状态逐一比较，而是从最相似状态周边获取其余的近邻状态。这里根据 K 值选择的不同对近邻状态选择采取不同的策略。当 $M\geqslant K_{max}$ 时，选取最相似状态以及其密集度状态中与当前状态最相近的 $K-1$ 个历史状态作为 K 近邻状态；当 $K_{min}\leqslant M<K_{max}$ 时，选取最相似状态以及其密集度状态作为 K 近邻状态；当 $M<K_{min}$ 时，从最相似状态的密集度状态及其各密集度状态中(如果近邻状态数量仍不够，则对剩余点进行广泛搜索)选取最相近的 K 个历史状态作为 K 近邻状态，如图 5 所示。

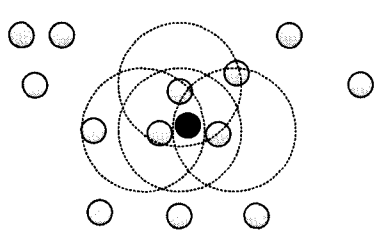


图 5　K 个近邻状态的选择

基于上述两点改进，在一定程度上提高了 K 近邻搜索的搜索精度和效率。改进的 K 近邻搜索算法的步骤具体如下：

Step1：根据空间距离计算方法 $distance(X[i],X[j])$ 计算历史数据库中各状态的密集度 M，并标注各状态的密集度状态 Xm；

Step2：根据状态向量中速度时间序列 $v(t-3)$，$v(t-2)$，$v(t-1)$，$v(t)$ 的变化情况选择相应的趋势分类类型 T；

Step3：根据当前时段 t 从类型 T 的历史状态数据库中计算得到最相似状态 X；

Step4：获取最相似状态 X 的密集度 M，根据取值范围[K_{min}，K_{max}]确定最终 K 值；

Step5：根据 K 值以及最相似状态的具体情况，从样本数据库中找出 K 个近邻状态；

Step6：返回最后结果 $X_i(t)$，dis_i，$1\leqslant i\leqslant K$。

2.2.4　预测方法与误差反馈

在实际情况中，下一时段的交通状态或多或少受当前状态的影响，除突发状况外不会发生突然性的剧烈变化。因而本文考虑根据最相似状态下一时段速度的变化趋势来预测下一时段的交通状态。同时，本文采用了加权平均的思想，根据与当前状态的相似程度来确定相似状态的权值，下一时段交通速度的最终预测模型如式(3)所示：

$$v(t+1)=\sum_{i=1}^{k}w_i\cdot(v(t)+\Delta v_i) \tag{3}$$

其中，w_i 为权值，计算方法为 $w_i=\dfrac{1/dis_i}{\sum_{i=1}^{k}\dfrac{1}{dis_i}}$，$dis_i$ 为与第 i 个 K 近邻状态的距离；$\Delta v_i=v_{hi}(t+1)-v_{hi}(t)$，即为第 i 个 K 近邻状态的下一时段速度 $v_{hi}(t+1)$ 与本时段速度 $v_{hi}(t)$ 之差。

为了提高预测精度，本文采用一般的误差反馈机制对相似度量的计算方法进行调整，即对式(1)中的 a_1、a_2、a_3、a_4 权值参数进行调节。选取 $E=\dfrac{1}{2}e^2$ 作为误差能量函数，其中 $e=\dfrac{v_{真}-v_{预}}{v_{真}}$ 为预测值与真实值之间的相对误差。当 $|e|<10\%$ 时，则认为误差在允许范围之内，不进行权值修正；当 $|e|>100\%$ 时，则认为当前发生了突发事件，影响了交通速度，不进行权值修正；当 $10\%\leqslant|e|\leqslant100\%$ 时，对权值进行调整 $a_i=a_i+\Delta a_i(i=1,2,3,4)$，具体如下：

$$\begin{cases} \Delta a_1=\dfrac{\partial E}{\partial e}\cdot\dfrac{\partial e}{\partial a_1}=e\cdot\dfrac{1}{2v_{真}\sum_{i=1}^{k}\dfrac{1}{dis_i}}\cdot\sum_{i=1}^{k}\dfrac{(v_i(t)-v_{hi}(t))^2}{(v_{hi}(t+1)-v_{hi}(t))dis_i^2\sqrt{dis_i}} \\ \Delta a_2=\dfrac{\partial E}{\partial e}\cdot\dfrac{\partial e}{\partial a_2}=e\cdot\dfrac{1}{2v_{真}\sum_{i=1}^{k}\dfrac{1}{dis_i}}\cdot\sum_{i=1}^{k}\dfrac{(v_i(t-1)-v_{hi}(t-1))^2}{(v_{hi}(t+1)-v_{hi}(t))dis_i^2\sqrt{dis_i}} \\ \Delta a_3=\dfrac{\partial E}{\partial e}\cdot\dfrac{\partial e}{\partial a_3}=e\cdot\dfrac{1}{2v_{真}\sum_{i=1}^{k}\dfrac{1}{dis_i}}\cdot\sum_{i=1}^{k}\dfrac{(v_i(t-2)-v_{hi}(t-2))^2}{(v_{hi}(t+1)-v_{hi}(t))dis_i^2\sqrt{dis_i}} \\ \Delta a_4=\dfrac{\partial E}{\partial e}\cdot\dfrac{\partial e}{\partial a_4}=e\cdot\dfrac{1}{2v_{真}\sum_{i=1}^{k}\dfrac{1}{dis_i}}\cdot\sum_{i=1}^{k}\dfrac{(v_i(t-3)-v_{hi}(t-3))^2}{(v_{hi}(t+1)-v_{hi}(t))dis_i^2\sqrt{dis_i}} \end{cases} \tag{4}$$

2.3 样本数据库设计

根据 K 近邻搜索的改进方案来设计样本数据库。可见，样本数据库主要需要向 K 近邻搜索阶段提供历史状态信息、速度变化趋势分类信息、历史状态的密集度信息、历史状态的密集度状态信息。为此，本主要设计以下 3 类数据库。

(1)原始数据库

原始数据库主要用于保存不同时段下的状态信息，其余的各种操作都是基于原始数据进行的，其数据结构如表 1 所列。

表 1 原始数据库数据结构

序号 id_x	v(t)	v(t−1)	v(t−2)	v(t−3)	时段 t	下时段速度 v
1982	36.000	40.623	41.358	48.741	1140	45.870
1983	35.505	34.078	30.271	27.377	1140	37.993
1984	30.048	39.994	35.221	50.926	1140	37.272
…	…	…	…	…	…	…

(2)趋势分类数据库

趋势分类数据库的主要作用是根据速度的变化趋势对原始数据库进行状态分类保存，一个原始数据库分别由 8 个趋势分类数据库分开保存。数据库在上述基础上还保存历史状态的密集度以及密集度状态信息，为了降低冗余度，这里的信息只保存状态在原始数据库中的编号。趋势分类数据库的数据结构如表 2 所列(以图 2 中的趋势(1)为例)。其中密集度状态 *set(t)*一项存储类型为文本，用于保存密集度状态在原始数据库中的序号。

表 2 趋势分类数据库数据结构

序号 id_m	密集度 M	密集度状态 set(t)	时段 t	序号 id_x
162	3	3064,3085,3089	1440	3086
163	4	3064,3067,3085,3086	1440	3089
164	1	3116	1445	3094
…	…	…	…	…

(3)区域索引库

区域索引库主要用于快速查找与当前状态最相似的状态。由于根据状态向量可以将历史状态映射到四维空间的点，且这些点在该空间上的分布范围为每一维取值在 5～85 之间的区域。对分布范围进行 $N\times N\times N\times N$ 的区域划分，每块区域保存各自区域内的状态信息。

当要进行最相似状态查找时，根据每一维的偏移计算可将当前状态快速定位到所在的小区域。若该区域存在当前时段的历史状态，则通过对比获得最相似状态；若该区域不存在当前时段的历史状态，则比较范围向外扩散一个索引范围直至扩散区域存在当前时段的历史状态为止。区域索引库的数据结构如表 3 所列(以图 2 中的趋势(1)为例)。其中历史状态集存储类型为文本，用于保存历史状态在趋势分类数据库中的序号。

表 3 区域索引库数据结构

序号 id_s	历史状态集	时段 t
6544	226,228	1720
6654	231,230,233	1725
b876	344	2025
…	…	…

有关数据库的数据处理机制主要涉及数据库的初始化、K 近邻的搜索以及数据库的更新，具体步骤如图 6 所示。

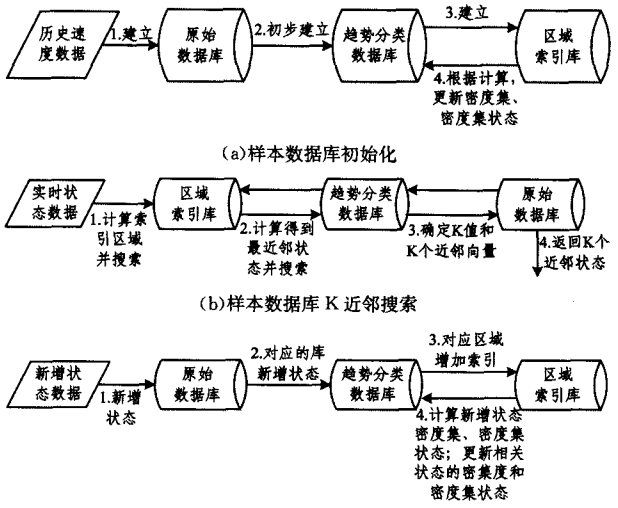


图 6 数据库主要操作流程

3 实验结果与分析

选取北京二环内的某一主干道路段为研究对象进行实验，预测间隔为 5 分钟，考虑该路段早上 6:20 到晚上 22:55 之间近 200 个时段的交通速度信息，其余时间则认为出行车辆不多，路段畅通，不予计算。

在实验过程中，交通速度数据来自北京市浮动车系统 2012 年 11 月的出租车浮动车数据，根据一定算法对这些数据进行有效性判断、预处理，以及对交通速度的估计，数据准确性可以得到保证，能满足需求，最终一共得到该路段连续 30 天各时段的交通速度。本文根据第 2 节中所提到的策略方法建立原始数据库、趋势分类数据库以及区域索引库(以 5 为间隔建立区域块，并用十六进制进行记录，如“b876”)。

为了验证本算法的有效性，将实验结果分别与基本非参数回归方法、BP 神经网络方法进行对比，其中 BP 神经网络采用单隐层结构，隐层中包含 3 个神经元。选取交通速度数据中前 28 天的数据作为样本历史数据，最后 2 天数据为测试数据进行实验。实验主要用到的比较指标为平均绝对误差 *MAE* 和平均相对误差 *MAPE*，即

$$MAE=\frac{1}{N}\sum_{i=1}^{N}|y_i-y_{i预测}|,MAPE=\frac{1}{N}\sum_{i=1}^{N}\frac{|y_i-y_{i预测}|}{y_i} \quad (5)$$

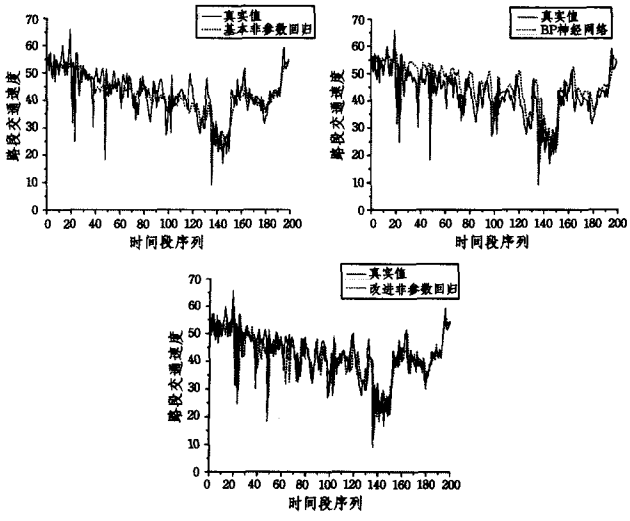


图 7 3 种算法预测数据与真实数据比较

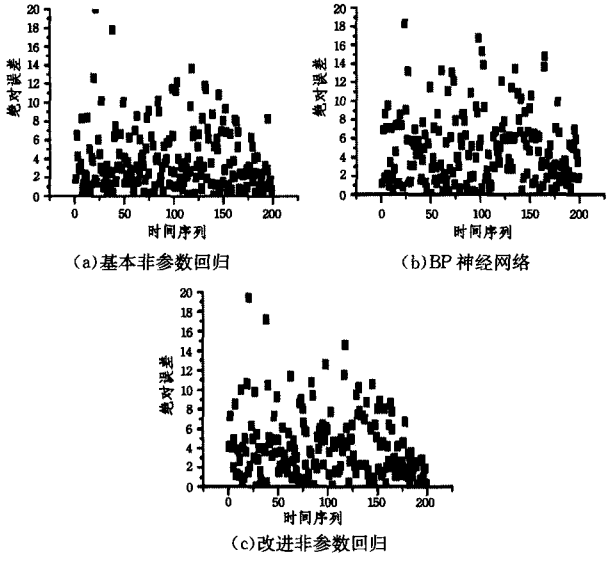
首先，本文对 3 种预测方法的整体拟合情况进行了对比

分析，得到的预测值与实测值的对比结果如图 7 所示。可以看到，3 种预测算法都有较好的拟合效果，但基本非参数回归预测算法对发生突发情况的交通速度预测结果并不好，它更倾向于交通速度的一个整体变化趋势；而 BP 神经网络和非参数回归改进算法能很好地应对突发情况下的交通速度预测。

对比了整体拟合效果之后，本文对 3 种预测方法的具体预测误差情况进行了对比，表 4 为 3 种预测方法的误差指标对比结果，图 8 为 3 种预测算法的绝对误差随时间序列的具体分布图。

表 4　3 种预测算法误差指标对比

	基本非参数回归	BP 神经网络	改进非参数回归
MAE	4.21	5.52	3.90
MAPE	10.48%	16.45%	9.98%



(a)基本非参数回归　(b)BP 神经网络

(c)改进非参数回归

图 8　绝对误差随时间序列的分布图

从图表中可以发现，非参数回归算法的预测结果均好于 BP 神经网络方法，而本文改进的非参数回归算法在基本非参数回归算法的预测结果上又有一定的预测精度提高，但是这个提升并不明显。通过分析可知，改进方案对实验中的历史交通速度数据按变化趋势进行了分类，使得每一类下的历史状态数变得不充裕。因而在大多情况下获取的状态与基本方法相近，当少部分分类情况下历史状态充裕时，预测精度才有一定提高，就一天的平均预测精度而言，这种提高并不明显。但随着历史数据量的不断丰富，预测精度提高会变得明显。

由上述对比可得知，本文改进的非参数回归预测方法无论是在路段整体的交通速度拟合情况方面还是在具体的预测结果上都具有很好的预测能力，其结果令人满意。

结束语　非参数回归是一种非常适合于交通状态预测的预测模型，在拥有丰富历史数据的情况下具有相对较高的预测精度。本文提出的改进方案旨在进一步提高模型在单一因素预测情况下的预测精度，其思想简单直观：通过速度变化趋势分类查询提高了近邻搜索的精度，很好地反映了状态的变化趋势；通过变 K 近邻搜索策略提高了各时段模型预测时的灵活性以及近邻状态数量的合理性。实验表明，改进方案具有一定的借鉴意义，下一步工作需要丰富的历史数据进行进一步实验。

参 考 文 献

[1] Zhu Yin, Wang Jun-li, Zhou Tong-mei. Introduction to Intelligent Transportation Systems[D]. Beijing: People's Public Security University of China Press, 2007(in Chinese)
朱茵，王军利，周彤梅. 智能交通系统导论[M]. 北京：中国人民公安大学出版社，2007

[2] Liu Le-min. The Short-term Traffic Flow Prediction of Main Thoroughfares of City[D]. Beijing: North China University of Technology, 2013(in Chinese)
刘乐敏. 城市主干道短时交通流预测研究[D]. 北京：北方工业大学，2013

[3] Van Lint H. Reliable travel time prediction for freeways[D]. Netherlands: TRAIL Research School, 2004

[4] Han C, Song S, Wang C H. A Real-time Short-term Traffic Flow Adaptive Forecasting Method Based on ARIMA Model[J]. Acta Simulata Systematica Sinica, 2004, 16(7): 1530-1456(in Chinese)
韩超，宋苏，王成红. 基于 ARIMA 模型的短时交通流实时自适应预测[J]. 系统仿真学报，2004，16(7)：1530-1532

[5] Sun S, Yu G, Zhang C. Short-term traffic flow forecasting using sampling Markov Chain method with incomplete data[C]// Intelligent Vehicles Symposium, 2004 IEEE. IEEE, 2004: 437-441

[6] Xie Y, Zhang Y, Ye Z. Short-Term Traffic Volume Forecasting Using Kalman Filter with Discrete Wavelet Decomposition[J]. Computer-Aided Civil and Infrastructure Engineering, 2007, 22(5): 326-334

[7] Nie P L, Yu Z, He Z C. Constrained Kalman filter combined predictor for short—term traffic flow[J]. Journal of Traffic & Transportation Engineering, 2008, 8(5): 86-90(in Chinese)
聂佩林，余志，何兆成. 基于约束卡尔曼滤波的短时交通流量组合预测模型[J]. 交通运输工程学报，2008，8(5)：86-90

[8] Jun G, Lin Q, Ming-yue L, et al. Forecasting urban traffic flow by SVR[C]// 2013 25th Chinese Control and Decision Conference (CCDC). IEEE, 2013: 981-984

[9] Chen Xue-ping, Zeng Sheng, Hu Gang. Short-time Traffic Flow Prediction Based on BP Neural Network[J]. Technology of Highway and Transport, 2008(3): 115-117(in Chinese)
陈雪平，曾盛，胡刚. 基于 BP 神经网络的短时交通流预测[J]. 公路交通技术，2008(3)：115-117

[10] Zhang L, Rao Q, Yang W, et al. An Improved k-NN Nonparametric Regression-Based Short-Term Traffic Flow Forecasting Model for Urban Expressways[C]// ICTE 2013: Safety, Speediness, Intelligence, Low-Carbon, Innovation. ASCE, 2014: 1214-1223

[11] Qu L, Lan S Y, Zhang J W. Short-term traffic forecasting based on nonparametric regression and floating car data[J]. Computer Engineering & Design, 2013, 34(9): 3298-3301(in Chinese)
屈莉，兰时勇，张建伟. 基于浮动车数据非参数回归短时交通速度预测[J]. 计算机工程与设计，2013，34(9)：3298-3301

[12] Gong Xiao-yan. Short-term Traffic Flow Forecasting and Routing based on Data Mining[D]. Beijing: Institute of Automation of Chinese Academy of Sciences, 2003(in Chinese)
宫晓燕. 基于数据挖掘的短时交通流预测及辅助诱导[D]. 北京：中国科学院自动化研究所，2003

[13] Li S, Liu L J, Xie Y L. Chaotic prediction for short-term traffic flow of optimized BP neural network based on genetic algorithm [J]. Control & Decision, 2011, 26(10): 1581-1585(in Chinese)
李松，刘力军，解永乐. 遗传算法优化 BP 神经网络的短时交通流混沌预测[J]. 控制与决策，2011，26(10)：1581-1585

[14] Zheng W, Lee D H, Shi Q. Short-term freeway traffic flow pre-

diction:Bayesian combined neural network approach[J]. Journal of Transportation Engineering,2006,132(2):114-121

[15] He Wei. Research on Prediction of Traffic Flow Using Fuzzy Neural Networks[D]. Lanzhou:Lanzhou Jiaotong University, 2012(in Chinese)
何伟. 模糊神经网络在交通流量预测中的应用研究[D]. 兰州:兰州交通大学,2012

[16] Zhang Xiao-li. A Study on Short-term Traffic Volume Forecasting Based on Non-Parametric Regression[D]. Tianjin:Tianjin University,2007(in Chinese)
张晓利. 基于非参数回归的短时交通流量预测方法研究[D]. 天津:天津大学,2007

[17] Li Zhen-long,Zhang Li-guo,Qian Hai-feng. Review of the Short-term Traffic Flow Forecasting Based on the Non-parametric Regression[J]. Journal of Transportation Engineering and Information,2009,6(4):34-39(in Chinese)
李振龙,张利国,钱海峰. 基于非参数回归的短时交通流预测研究综述[J]. 交通运输工程与信息学报,2009,6(4):34-39

[18] Guo F,Krishnan R,Polak J W. Short-term traffic prediction under normal and incident conditions using singular spectrum analysis and the k-nearest neighbour method[C]// IET and ITS Conference on Road Transport Information and Control (RTIC 2012). IET,2012:1-6

[19] Elfaouzi N E. Nonparametric traffic flow prediction using kernel estimator[C]// Internaional symposium on transportation and traffic theory. 1996:41-54

(上接第 203 页)

表 4 应用 MEO 和 TEO 减少的 MTP 执行次数百分比

程序包	MEO			MEO 和 TEO		
	Aver	Max	Min	Aver	Max	Min
printtokens	15.9%	17.8%	14.8%	22.8%	27.8%	18.3%
schedule	19.5%	21.8%	17.9%	30.3%	37.0%	26.1%
tcas	60.2%	85.0%	40.5%	77.9%	89.3%	68.0%
totinfo	20.3%	61.0%	13.5%	36.0%	62.9%	25.7%
replace	39.3%	82.3%	0.3%	59.8%	89.6%	28.6%
space	45.6%	66.4%	3.0%	56.4%	75.2%	3.0%

结束语 基于变异的错误定位方法 MBFL 作为一种具有高定位精确度的错误定位技术,存在着执行开销过大的问题。本文从优化变异执行策略的角度出发,提出了动态变异执行策略 DMES,避免在变异体执行过程中执行不必要的变异-测试对。通过对 6 个程序包的 127 个错误版本进行实验分析,一方面验证了应用 DMES 并不会对 MBFL 的错误定位准确度产生影响;另一方面,验证了应用 DMES 可有效减少变异-测试对的执行开销。实验结果表明本文提出的 DMES 在不降低 MBFL 的定位准确度的条件下,可以有效减少计算语句怀疑度所需执行的变异-测试对次数,提高了 MBFL 方法的执行效率。

此外,现有针对 MBFL 的执行开销减少方法主要从减少变异体数量方面进行考虑,而本文方法则关注尽量避免变异体执行过程中不必要的执行开销。因此,后续研究可考虑将二者结合,进一步提高 MBFL 的执行效率。

参 考 文 献

[1] Masri W,Abou-Assi R,El-Ghali M,et al. An empirical study of the factors that reduce the effectiveness of coverage-based fault localization[C]//Proceedings of the 2nd International Workshop on Defects in Large Software Systems:Held in conjunction with the ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA 2009). ACM,2009:1-5

[2] Yu K,Lin M X. Advances in Automatic Software Fault Localization Techniques[J]. Chinese Journal of Computers,2011,34(8):1411-1422(in Chinese)
虞凯,林梦香. 自动化软件错误定位技术研究进展[J]. 计算机学报,2011,34(8):1411-1422

[3] Jones J A,Harrold M J. Empirical evaluation of the tarantula automatic fault-localization technique[C]// Proceedings of the 20th IEEE/ACM international Conference on Automated software engineering. ACM,2005:273-282

[4] Abreu R, Zoeteweij P, Van Gemund A J C. An evaluation of similarity coefficients for software fault localization[C]// 12th Pacific Rim International Symposium on Dependable Computing,2006(PRDC'06). IEEE,2006:39-46

[5] Naish L,Lee H J,Ramamohanarao K. A model for spectra-based software diagnosis[J]. ACM Transactions on software engineering and methodology (TOSEM),2011,20(3):1-32

[6] Masri W. Fault localization based on information flow coverage [J]. Software: Testing, Verification and Reliability, 2010, 20 (2):121-147

[7] Santelices R,Jones J A,Yu Y,et al. Lightweight fault-localization using multiple coverage types[C]// IEEE 31st International Conference on Software Engineering,2009(ICSE 2009). IEEE, 2009:56-66

[8] Papadakis M,Le Traon Y. Using mutants to locate "Unknown" faults[C]// 2012 IEEE Fifth International Conference on Software:Testing,Verification and Validation (ICST). IEEE,2012: 691-700

[9] Papadakis M,Le Traon Y. Metallaxis-FL:Mutation-based Fault Localization[J]. Software:Testing,Verification and Reliability, 2015,25(5-7):605-628

[10] Papadakis M,Le Traon Y. Effective Fault Localization via Mutation Analysis:A Selective Mutation Approach[C]// Proceedings of the 29th Annual ACM Symposium on Applied Computing. 2014:1293-1300

[11] Jia Y,Harman M. An analysis and survey of the development of mutation testing[J]. Software Engineering, 2011, 37(5): 649-678

[12] Zhang L, Marinov D, Khurshid S. Faster mutation testing inspired by test prioritization and reduction[C]// Proceedings of the 2013 International Symposium on Software Testing and Analysis. ACM,2013:235-245

[13] Do H, Elbaum S, Rothermel G. Supporting controlled experimentation with testing techniques:An infrastructure and its potential impact[J]. Empirical Software Engineering,2005,10(4): 405-435

[14] Delamaro M E,Maldonado J C,Vincenzi A M R. Proteum/IM 2.0:An integrated mutation testing environment[M]// Mutation testing for the new century. Springer US,2001:91-101