引用格式：梁艳平,毛政元,邹为彬,等.基于相似数据聚合与变K值KNN的短时交通流量预测[J].地球信息科学学报,2018,20(10):1403-1411. [Liang Y P, Mao Z Y, Zou W B, et al. Short-term traffic flow prediction based on similar data aggregation and KNN with varying K-value[J]. Journal of Geo-information Science, 2018,20(10):1403-1411.] **DOI:**10.12082/dqxxkx.2018.180281

基于相似数据聚合与变K值KNN的短时交通流量预测

梁艳平[1,2,3],毛政元[1,2,3*],邹为彬[1,2,3,4],许　锐[5]

1. 福州大学福建省空间信息工程研究中心,福州 350002;2. 福州大学空间数据挖掘与信息共享教育部重点实验室,福州 350002;3. 福州大学地理空间信息技术国家地方联合工程研究中心,福州 350002;4. 福建工程学院交通运输学院,福州 350118;5. 福建工程学院信息科学与工程学院,福州 350118

Short-term Traffic Flow Prediction Based on Similar Data Aggregation and KNN with Varying K-value

LIANG Yanping[1,2,3], MAO Zhengyuan[1,2,3*], ZOU Weibin[1,2,3,4], XU Rui[5]

1. Provincial Spatial Information Engineering Research Center, Fuzhou University, Fuzhou 350002, China; 2. Key Laboratory of Spatial Data Mining and Information Sharing of Ministry of Education, Fuzhou University, Fuzhou 350002, China; 3. National Engineering Research Centre of Geospatial Space Information Technology, Fuzhou University, Fuzhou 350002, China; 4. School of Transportation, Fujian University of Technology, Fuzhou 350118, China; 5. School of Information Science and Engineering, Fujian University of Technology, Fuzhou 350002, China

Abstract: Real-time and accurate short-term traffic flow prediction, a critical technical problem in traffic control and guidance which is challenging and needs to be solved urgently in related research fields and engineering practice, still remains because of the hardship caused by the uncertainty and the temporal variability in traffic flow datasets acquired in different times. In order to improve the performance of the short-term traffic flow prediction, a new method based on similar data aggregation techniques and a modified KNN algorithm with varying K-value (KNN-SDA) was proposed and the related algorithm was also implemented and tested on actual measured datasets in this paper. Firstly state vectors were generated from the preprocessed traffic flow datasets by calculating the optimal time delay with the help of the mutual information theory. Each of our state vectors is composed of two parts, the first one of which is a regular state vector and the second one of which is a modified state vector which makes a contribution to a higher similarity between our state vectors and those in training datasets. Subsequently a historical traffic flow database of temporal series was constructed on the basis of results mentioned above for further experiments. After that, the proposed similar data aggregation techniques were applied to aggregate and clean data to obtain 144 training data sets in different times from historical traffic flow database, which would effectively improve the prediction accuracy and efficiency of the proposed algorithm. At last, the optimal K-values, each of which corresponded to a moment, were determined through the cross validation method. So far, the overall process of the KNN-SDA algorithm with varying K-value has been completed. In order to verify the performance of the proposed method, we compared the experimental results derived from our method with those from three other ones. It turns out that the KNN-SDA algorithm with

收稿日期：2018-06-11；**修回日期**：2018-07-19.

基金项目：国家自然科学基金项目(41471333)；福建省自然科学基金面上项目(2018J01619)。[**Foundation items**：National Natural Science Foundation of China, No.41471333; Project of Science and Technology of Fujian Province, No.2018J01619.]

作者简介：梁艳平(1993-)，男，硕士生，研究方向为短时交通流量预测、智能算法。E-mail: 497336236@qq.com

***通讯作者**：毛政元(1964-)，男，博士，教授，博士生导师，研究方向为时空序列分析、城市变化检测、信息化管理与信息服务。E-mail: zymao@fzu.edu.cn

varying K-value proposed in this article can improve the prediction accuracy significantly and ensure high execution efficiency as well.

Key words: short-term traffic prediction; mutual information method; similar data aggregation; KNN; cross validation

***Corresponding author:** MAO Zhengyuan, E-mail: zymao@fzu.edu.cn

摘要:短时交通流量预测是交通控制和诱导涉及的关键技术问题,由于短时交通流量存在不确定性和时变性,其预测难度较大,是相关研究领域与工程实践中亟待解决的难题。为提高短时交通流量预测的准确性,本文设计与实现了基于相似数据聚合和变K值KNN(KNN-SDA)的短时交通流量预测算法。该算法首先采用互信息法在经过预处理的交通流量数据集提取交通流量序列最佳延迟时间信息,生成状态向量,并构建交通流量历史数据库;然后以本文所提出的相似数据聚合方法完成历史数据的聚合与清洗得到训练数据集;最后通过交叉验证确定每个时刻的最优K近邻数,完成算法实现。实验结果表明,本文提出的变K值KNN-SDA算法在保证执行效率的同时能明显提高短时交通流量的预测精度。

关键词:短时交通流预测;互信息法;相似数据聚合;KNN;交叉验证

1 引言

随着经济的发展与城市化水平的提升,交通流量快速增加与相关基础设施建设滞后的矛盾越来越尖锐,交通拥堵已经成为国内大中城市管理中必须面对和解决的主要问题之一。智能交通控制和诱导是解决城市交通问题的有效途径,实时准确的交通流量预测、特别是时长在15 min以内的短时交通流量预测是智能交通控制和诱导的重要依据,对于解决城市交通问题更具实用价值。

道路交通是一个复杂的非线性系统,它受到多种自然和人为因素的影响,具有不确定性和时变性等特征,这些特征给交通流量的预测带来了较大困难。短时交通流量受随机干扰因素影响更大,不确定性和时变性更强,其预测也更具挑战性,是其中的难点。

现有文献中提出的短时交通流量预测模型(方法)可归纳为以下3种类型:① 经典数学模型,如贝叶斯网络[1]、卡尔曼滤波[2-3];② 人工智能模型,如神经网络[4]、极端学习机[5-6]、支持向量回归[7-8];③ 非参数回归方法,如K近邻算法[9-20]。其中,前两类模型是试图用数学表达式显式(经典数学模型)或隐式(人工智能模型)地描述影响变量X与预测变量Y之间的函数关系,这些方法侧重解决交通流量预测中的非线性函数关系问题,但对短时交通流量中的不确定性和时变性考虑较少;其次,当新观测数据加入到模型中时,需要重新计算模型参数,这个过程将非常耗时。因此,将其用于解决具有非线性、时变性与不确定性的短时交通流量预测问题时仍存在明显的局限性。

非参数回归是一类基于数据驱动的预测方法,预测过程中不需要建立影响变量与预测变量具体的函数关系,而是依据已有数据决定输入输出关系。新的观测数据可以随时加入到模型中作为历史数据样本;另外,非参数回归中不需对原始数据做平滑处理,其历史数据库中的数据保持了原始数据的特性,因此,其受不确定性与时变性影响更小,且在特殊路况出现时,非参数回归预测更为准确。因此,其被越来越多的研究者用于解决短时交通流量预测问题,大量通过实时检测器获取的交通流数据是该方法用于短时交通流量预测的前提[12]。

Davis等[11]最早将非参数回归方法应用于交通流量预测,指出KNN方法之所以适用于交通流量预测,源于交通流量数据本身所体现的非线性特性,但该算法运行时间相对较长。Smith等[12]在Davis等研究的基础上通过对状态向量进行改进,取得了比神经网络更好的效果,并指出聚类分析有助于提高预测精度。刘洋等[13]将K-means聚类分析与非参数回归结合,把交通流量分割为不同的交通流量模式,且每种模式都视为一种独立的训练数据集,提高了非参数回归的预测精度。张晓利等[16]采用聚类方法和平衡二叉树结构建立历史数据库,以提高KNN算法的搜索速度;孟梦等[17]在张晓利等的工作基础上做了进一步的改进,通过在KNN匹配中加入模式向量提高KNN算法的预测精度。Frank等[18]在KNN模式匹配中加入时间信息对算法进行改进,但该方法不能兼顾算法的运行效率。这些改进在一定程度上提高了预测精度和算法执行效率,但当数据库容量较大时,仍然难以满足实时预测的需求。现有文献中基于非参数回归方法预测短时交通流量的

研究成果主要存在以下2个方面的局限性：① 因历史数据库的不合理分割所导致的不相关数据匹配问题；② 因采用全局最优K值所导致的交通流量模式误判问题。

本文提出一种基于相似数据聚合和变K值KNN的短时交通流量预测方法（变K值KNN-SDA），试图通过克服上述2个方面的局限性提高预测精度。

2　变K值KNN-SDA算法框架

2.1　基于变K值KNN-SDA的非参数回归

模式识别是KNN非参数回归预测方法的基础，即利用数据库模式匹配的思想，找到一组与输入数据对应或相似的数据进行预测[12]。在KNN短时交通流量预测中，X 表示路段 t 时刻以及之前 m 个时刻的交通流量集合（状态向量），$\hat{y}$ 表示 X 经数据库匹配后得到的 $t+1$ 时刻交通流量预测值。

本文提出的变K值KNN-SDA短时交通流量预测的算法流程如图1所示，其中增加了一般KNN方法没有的相似数据聚合与变K值2个步骤，并针对状态向量做了改进。该算法的两项核心工作为确定变K值KNN-SDA算法所需的训练数据集与K值，算法具体过程为：

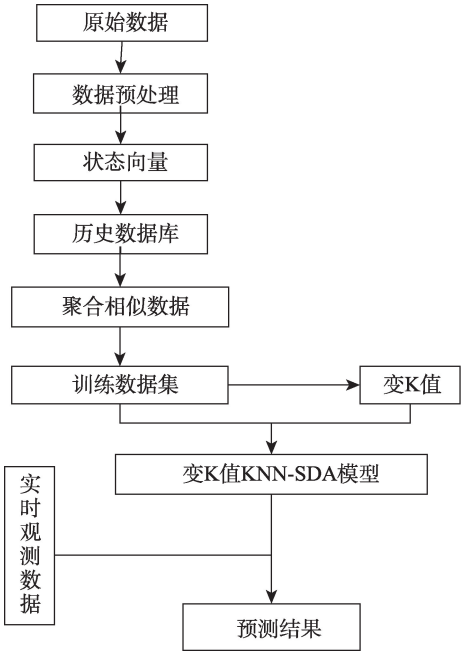

图1　基于KNN-SDA短时交通流量预测算法流程图

Fig. 1　The figure of short-term traffic flow prediction method based on KNN-SDA

（1）首先对原始交通流量数据做预处理，接着使用互信息发法求解交通流量序列的最佳延迟时间并生成状态向量，所有状态向量构成交通流量历史数据库；

（2）然后以本文提出的方法（见2.3节）聚合与清洗历史数据，得到KNN-SDA算法所需的训练数据集，不同时刻对应不同的训练数据集；

（3）再通过交叉验证确定每个时刻的最优K近邻数；

（4）最后利用距离倒数权重法预测短时交通流量，详细过程见见2.5节。

2.2　选择输入变量X

在KNN短时交通流量预测中，状态向量[15]（即输入变量）X 是模式匹配的基础，为提高模式匹配的准确性，本文在 X 中增加 $X(2)$ 部分，即 $X=[X(1), X(2)]$，其中：

$$X(1)=\left[v(t), v(t-1), \cdots, v(t-\tau)\right] \tag{1}$$

式中：$v(t)$、$v(t-1)$ 分别表示当前时刻和前一时刻的交通流量，以此类推，$v(t-\tau)$ 表示前 τ 时刻的交通流量；τ 表示交通流量的延迟时间，在连续的交通流量时间序列中，延迟时间 τ 决定的是哪些历史时刻的交通流量与当前时刻交通流量的状态最为匹配[20]。本文采用互信息法（MI）[21]确定最佳延迟时间 τ，该方法（MI）的定义如下：

考虑2个离散信息系统 $\{s_1, s_2, \cdots, s_n\}$ 和 $\{q_1, q_2, \cdots, q_m\}$ 构成的系统 S 和 Q。由信息论可知，从2个系统中所获得的平均信息量，即信息熵分别为：

$$H(S)=-\sum_{i=1}^{n}P_i(s_i)\log_2 P_s(s_i) \tag{2}$$

$$H(Q)=-\sum_{j=1}^{m}P_q(q_j)\log_2 P_q(q_j) \tag{3}$$

式中：$P_s(s_i)$ 和 $P_q(q_j)$ 分别为 S 和 Q 中事件 s_i 和 q_j 的概率。

在给定 S 的情况下，S 和 Q 的互信息为：

$$I(Q,S)=H(Q)-H(Q|S) \tag{4}$$

其中，

$$H(Q|s_i)=-\sum_{j}[P_{sq}(s_i,q_j)|P_s(s_i)]\log\left[P_{sq}(s_i,q_j)|P_s(s_i)\right] \tag{5}$$

因此有：

$$I(Q,S)=\sum_i\sum_j P_{sq}(s_i,q_j)\log_2\left[\frac{P_{sq}(s_i,q_j)}{P_s(s_i)P_q(q_j)}\right] \tag{6}$$

式中：$P_{sq}(s_i, q_j)$为事件s_i和事件q_j的联合分布概率。

定义$[s, q]=[x(t), x(t+\tau)]$，即s代表时间序列$x(t)$，q为其延时为τ的时间序列$x(t+\tau)$，则函数$I(Q, S)$与时间延迟有关，记为$I(\tau)$。$I(\tau)$的大小代表了在已知系统S即$x(t)$的情况下，系统Q也就是$x(t+\tau)$的确定性大小。$I(\tau)=0$，表示$x(t+\tau)$完全不可预测，即$x(t)$和$x(t+\tau)$完全不相关；而$I(\tau)$的极小值，则表示了$x(t)$和$x(t+\tau)$是最大可能的不相关，重构时使用$I(\tau)$的第一个极小值作为最优延迟时间。

$$X(2)=[\Delta(t), \Delta(t-1), \cdots, \Delta(t-\tau-1)] \quad (7)$$

式中：$\Delta(t)=v(t)-v(t-1)$，以此类推。

基于KNN预测短时交通流量最重要的环节是使用距离度量准则准确地匹配与当前状态最为相似的邻居。传统方法仅以2个状态向量之间的“距离”为依据，没有考虑到两者所对应的交通流量变化趋势是否相同。本文通过在状态向量X中加入$X(2)$使其能准确地匹配到与当前时刻交通流量最为相似且变化趋势最相近的历史状态向量。记状态向量：

$$X=[X(1), X(2)]=[v(t), v(t-1), \cdots, v(t-\tau), \Delta(t), \Delta(t-1), \cdots, \Delta(t-\tau-1)] \quad (8)$$

2.3 聚合相似数据

文献[10]中指出对当前时刻预测有用的邻居可在附近时段内找到，为此将历史数据集中的状态向量按照0:00−24:00点每10 min为间隔分为144个不同的子集。聚合相似数据的目的是确定可用于待预测时刻的历史数据子集，该方法按照同样的流程分别在144个历史数据子集中划分用于聚合的相似数据，因此，不同时刻间聚合后的数据集可能存在重叠。相似数据聚合完成后，最近邻搜索过程中只需要考虑聚合后的相似数据集而不是整个历史数据库，可提高搜索匹配的效率与准确性。

为确定某时刻的数据集对于预测当前时刻交通流量的贡献，引入预测误差指标作为数据选择与聚合的标准。

2.3.1 选取相似数据

将交通流量历史数据库V按照时间顺序排序，得到：

$$V=\{v(0), v(1), \cdots, v(t), \cdots, v(144)\} \quad (9)$$

式中：$v(0)$表示零点时刻的交通流量数据集；$v(t)$表示当前时刻的交通流量数据集。

相似数据聚合的过程是以当前时刻的交通流量数据集为基准从左右两侧分别聚合下一时刻的数据集，即

$$V_0'=\{v(t)\},$$
$$V_1'=\{v(t)、v(t-1)\},$$
$$V_2'=\{v(t)、v(t-1)、v(t+1)\},$$
$$V_3'=\{v(t)、v(t-1)、v(t+1)、v(t-2)\},$$
$$V_4'=\{v(t)、v(t-1)、v(t+1)、v(t-2)、v(t+2)\},$$

依次类推。分别使用$\{V_i'(t)|(i=0, 1, \cdots, 143)\}$所对应数据集来预测$v(t)$时刻的交通流量，以MAE衡量预测误差，按照相似数据聚合的顺序选择第一个极小值所对应的数据集作为该时刻的训练数据集，并记为$V'(t)$。

2.3.2 剔除异常数据

聚合后的数据集$V_i'(t)$中可能存在某些杂质数据，本文将利用矩阵相似度[22]指标去除这类数据。设$C^{m\times n}$表示$m\times n$的矩阵，若$A, B\in C^{m\times n}$，则矩阵内积$\langle\cdot,\cdot\rangle$定义为：

$$\langle A, B\rangle=tr(B^T A) \quad (10)$$

式中：tr(·)表示矩阵主对角线元素之和。由矩阵内积可导出其范数$\|\cdot\|$为：

$$\|A\|=\sqrt{\langle A, A\rangle} \quad (11)$$

矩阵相似度r定义为：

$$r=\cos\theta=\frac{\langle A, B\rangle}{\|A\|\|B\|} \quad (12)$$

式中：θ为2个矩阵之间的夹角，r的值域为[−1, 1]。$r=0$，表示2个矩阵不相似；r=1，此时2个矩阵相似性最好。

设式(12)中A为$V_i'(t)$中的$v(t)$，B为$V_i'(t)$中的其他元素，分别求$v(t)$与$V'(t)$中其他元素的相似度，通过预测精度选定阈值剔除$V'(t)$中的杂质数据。

2.4 选取K值

K值表示从训练数据集中选取的近邻个数，其大小对KNN预测算法的精度有直接影响，故如何选取合理的K值成为与KNN预测算法相关的重要研究内容[17]。本文摒弃传统的全局最优K值(即所有预测时刻为一个固定的K值)的方法，而分别为每个预测时刻对应的训练数据集选取最合适的K值。

对于每个时刻，本文先在区间[2,20]上为其分配一组K值，对每个K值都做十折交叉验证（即十次交叉验证）得到一个预测误差，选取预测误差最小的K值作为当前时刻的近邻数，重复此过程，得到所有时刻的近邻数。

交叉验证过程中，对2.3节得到的每个时刻的训练数据集 $V'(t)$，首先将 $v(t)$ 从 $V'(t)$ 中分离出来，并记剩余数据为 $V''(t)$，然后将 $v(t)$ 等分为10部分，即 $v(t)=\{v(t_1), v(t_2), \cdots, v(t_{10})\}$，每次取 $v(t)$ 中的一个部分作为验证数据集，剩余九个部分与 $V''(t)$ 共同组成训练样本集，得到一个预测误差，如此重复10次，并取10个预测误差的平均值作为当前K值的预测误差。

2.5 选取预测算法

将经过上述匹配机制在训练数据集中获取的K个近邻点用于预测下一时刻的交通流量。本文选用带权重的预测算法[15]，其计算公式为：

$$\tilde{y}=\sum_{i=1}^{K}\beta_i v_{hi}(t+1) \tag{13}$$

$$\beta_i=\frac{d_i^{-1}}{\sum_{i=1}^{K} d_i^{-1}} \tag{14}$$

式中：$\tilde{y}$ 表示 $t+1$ 时刻的交通流量预测值；K 为近邻点个数；$v_{hi}(t+1)$ 为近邻点中下一时刻的交通流量；d_i 为当前状态向量与近邻点状态向量之间的欧氏距离。

3 实验及结果分析

本文数据源通过微波传感器实时采集，采集地点为福州市五四路东大路口，采集时间为2016年4月1日至2016年6月30日，采集频率为每10 min一次。原始数据经预处理后用于本文实验。本文算法全部采用Python编程语言实现，算法均在Intel i7−6700 CPU、8G内存及Windows 10环境下运行。

3.1 建立训练数据集

本节实验为KNN-SDA算法建立所需训练数据集，其中3.1.1节为构建状态向量，3.1.2节则以3.1.1节构建的状态向量为基础建立训练数据集。

3.1.1 确定状态向量X

原始交通流量数据经过数据筛选、异常数据替换后即可用于建立状态向量，本小节实验使用MI算法计算交通流量序列的最佳延迟时间 τ，并通过2.2节所述方法建立状态向量 X。MI算法实验结果如图2所示，其中 t 表示向前延迟时刻的个数，$I(t)$ 表示互信息值。当 t=5时（图中直线标注位置），$I(t)$ 取得第一个最小值，即为最佳延迟时间 $\tau=5$。

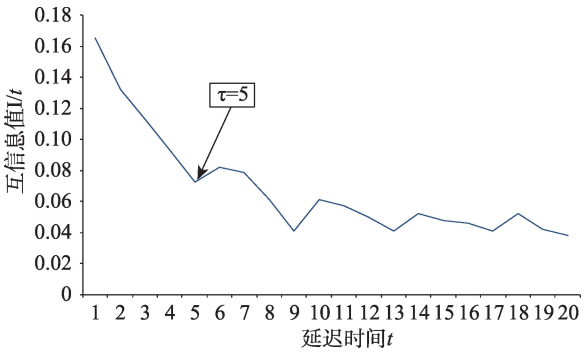

图2　互信息法求解延迟时间

Fig. 2　Solving time delay based on MI

3.1.2 确定训练数据集

本节实验按照2.2节中所述方法确定样本数据集。首先将3.11节中确定的状态向量按照2.3节所述方法分割为144个历史数据子集；然后按照2.3.1节所述方法为每个子集确定相似数据集，即各对应时刻的训练数据集。

本节实验以12:20历史数据子集为例为其确定相似数据集，$MAE(i)$ 与 $V_i'(t)$ 的关系曲线如图3所示，其中横轴代表第 i 次聚合所对应的数据集 $V_i'(t)$，纵轴代表MAE误差值。当 i=24时（图中直线标注位置），$MAE(t)$ 取得第一个极小值，即聚合边界为右侧14:20、左侧10:30（即第23次聚合边界），即得到该时刻的训练数据集 $V'(t)$。

确定好聚合边界后，再按照2.2.2节所述方法剔除上述实验得到的训练数据集 $V'(t)$ 中的异常数据，相似度 r 与 $v(t)$ 的关系曲线如图4所示，其中横轴代

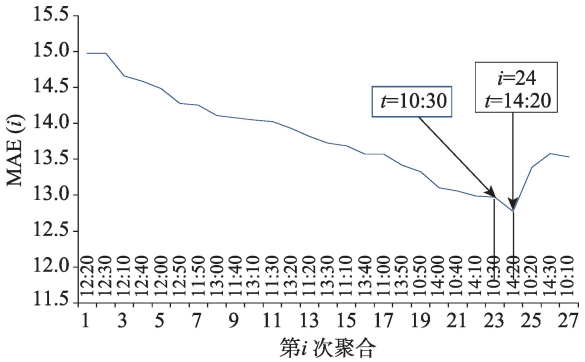

图3　聚合数据误差曲线图

Fig. 3　Error curve of data aggregation

表对应时刻的历史数据子集，纵轴代表矩阵相似度。本实验选取矩阵相似度阈值为0.9，并将低于该阈值的数据视为杂质数据。如图4中13:10和13:50所对应的MS值小于0.9，则将这2个时刻的历史数据子集从相似数据集中剔除。

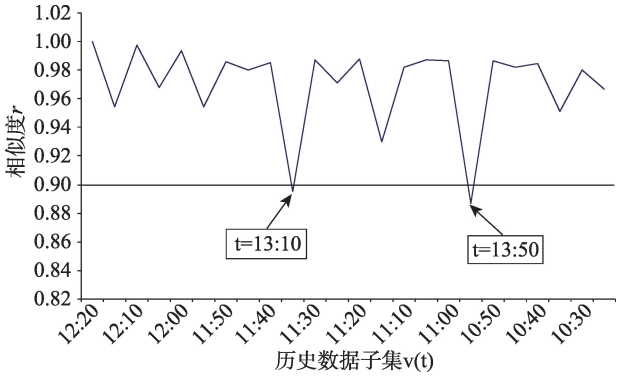

图4 相似度曲线图

Fig. 4 Similarity curve of aggregation data

3.2 近邻点个数

本节实验按照2.3节所述方法采用3.1节实验确定的训练数据集为每个时刻选取不同的K值，K值的分布如图5所示。观察该图可知，K值集中分布在3、4、5、6、7这5个值上，只有少数几个时刻对应的K值大于10，K值整体上偏小的主要原因是本实验对数据进行了精细的划分，使训练数据集中的数据相比历史数据库中的数据分布更集中。

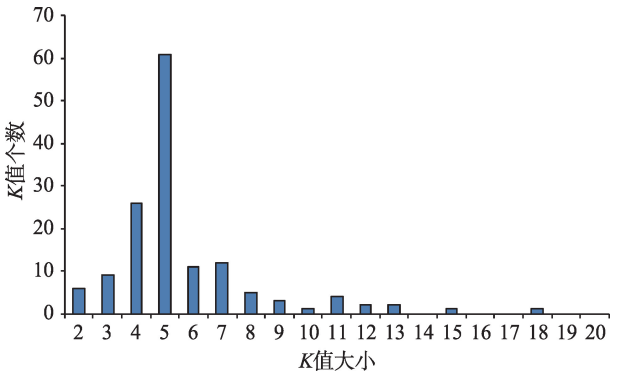

图5 K值分布图

Fig. 5 Distribution of K value

3.3 精度评价指标

为精确地评价本算法及比较算法的预测效果，本文引入4种精度评价指标[20]，分别是：平均绝对误差（MAE）、平均绝对百分误差（MAPE）、均方根误差（RMSE）、均等系数（EC）。其中，MAE、MAPE、RMSE越小，说明预测误差越小，即预测效果越好；EC越接近1，说明预测值与真实值之间的拟合度越好，EC值大于0.9时说明模型性能较好[6]。4种评价指标公式如下：

$$MAE=\frac{1}{n}\sum_{i=1}^{n}|y_i-\tilde{y}_i| \tag{15}$$

$$MAPE=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i-\tilde{y}_i}{y_i}\right| \tag{16}$$

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\tilde{y}_i)^2} \tag{17}$$

$$EC=1-\frac{\sqrt{\sum_{i=1}^{n}(y_i-\tilde{y}_i)^2}}{\sqrt{\sum_{i=1}^{n}(y_i)^2}+\sqrt{\sum_{i=1}^{n}(\tilde{y}_i)^2}} \tag{18}$$

式中：$\tilde{y}_i$ 为第 i 个预测值；y_i 为第 i 个真实值；n 为样本大小。

3.4 实验设计及结果分析

本文共包含3组实验，分别是KNN算法参数分析及比较实验、KNN算法与其他算法比较实验以及分离工作日与非工作日数据的对比实验。

3.4.1 KNN算法参数分析及比较实验

如表1所示，本小节共包含5种算法，其中算法F1为原始KNN算法，F5为本文所提出的KNN算法，F2、F3、F4为3种采用不同参数的同类算法；F2算法所使用状态向量未包含 $X(2)$ 部分；F3算法中所使用数据是全部历史数据集；F4算法中所有时刻的K值均为5。表中√表示该算法是否包含对相应列的改进。

表1 算法参数选择表

Tab. 1 Selection of algorithm parameters

算法	改进状态向量	数据聚合	变K值
F1			
F2		√	√
F3	√		√
F4	√	√	
F5	√	√	√

本节实验以表1所述5种算法为标准，数据通过预处理使其适用于F5算法，通过参数选取分别为F1-F5算法分配所需数据及参数。本节实验采用交叉验证（十折交叉验证）的思想，得到如表2所示的预测结果误差比较表。

表2中EC列的值表明，除F1之外的其他4种算法均为可靠的预测算法；从F2、F3、F4算法与F5算法的预测误差比较来看，状态向量、数据聚合方

表2　预测结果误差比较表

Tab. 2　Error comparison of prediction results

算法	MAE	MAPE/%	RMSE	EC
F1	15.64	7.92	21.47	0.899
F2	13.25	6.78	18.93	0.945
F3	13.67	7.0	19.35	0.944
F4	11.61	5.93	16.84	0.952
F5	11.07	5.69	16.29	0.969

式、K值的改变对预测结果均有影响。其中，K值的改变（F4）对算法预测精度的影响最小，这主要是因为F4中选取的K值为5，而F5算法中的K值主要集中在5附近（参见3.2.3节）；状态向量的改进（F2），数据聚合的改进（F3）均对预测精度有显著影响，说明本文针对这3种参数所做的改进都有助于提高KNN模型的预测精度。

3.4.2　全部数据、工作日及非工作日数据分离比较实验

本节的3组实验均采用本文提出的算法完成，区别在于历史数据集不同。全部数据与3.3.1节F5算法相同；工作日数据与节假日数据分离表示从历史数据库分别提取工作日数据与非工作日数据，生成两个独立的子数据库，分别用于实验。实验采用交叉验证思想，预测结果误差如表3所示。

表3　预测结果误差表

Tab. 3　Error comparison of prediction results

数据类型	MAE	MAPE/%	RMSE	EC
全部数据	11.07	5.69	16.29	0.969
工作日数据	10.53	5.62	15.6	0.973
非工作日数据	11.81	5.91	17.38	0.952

由表3评价指标，3组实验误差相差很小，说明本文设计的实验可以同时适用于工作日和非工作日。其中，非工作日的预测结果最差，这主要是因为非工作日数据量少且非工作日的交通模式差别大；工作日取得了较好的结果，说明在数据充沛的情况下对数据集进行精细的分割有助于提高模型的预测精度。结合非参数模型的特性（对历史数据库质量要求较高），说明本文所设计的数据聚合方法能够选择到合适的训练数据集，据此下文实验所用数据不再区分工作日与非工作日。

3.4.3　与其他预测算法的比较实验

本节实验选择目前在短时交通流量预测领域被广泛应用的支持向量回归（SVR）、神经网络（BP）以及KNN（3.4.1节中F1算法）作为比较算法。

本节采用的SVR算法参数设置为：核函数rbf、惩罚系数le4gamma系数0.001。

本节采用的BP算法为三层BP神经网络，其中输入层单元个数为4，输出层单元个数为1，隐含层单元个数为5，相关参数设置为最大循环次数10000，目标误差0.001，动量常数0.001，学习速率0.001。

本节实验中将最后一周的数据作为测试数据，剩余数据作为训练数据，预测结果误差如表4所示。

表4　预测结果误差表

Tab. 4　Error comparison of prediction results

算法	MAE	MAPE/%	RMSE	EC	时间/s
KNN	16.01	8.41	641.64	0.898	14.59
SVR	14.84	7.72	536.68	0.935	5.86
BP	13.08	6.70	378.79	0.943	6.31
KNN-SDA	11.75	6.22	332.55	0.949	6.17

由表4可知，KNN-SDA算法的4种评价指标均优于其他3种算法，说明KNN-SDA算法的预测结果更准确，且KNN-SDA算法的EC值比其他2种算法的EC值更接近1，说明该模型的预测值与实际值拟合更佳。同时与SVR、BP算法相比，本文算法用时适中，但明显小于KNN算法，说明本文对KNN算法的改进可更好地满足实时预测的需求。

为了更直观地反映3种算法预测结果与真实交通流量之间的差异、对比三者的预测结果，本文以2016年6月24日全天的交通数据为例，在图6中分别展示真实交通流量及3种算法预测的交通流量。

从图5整体预测效果及左上、右上两幅子图的曲线拟合效果来看，KNN-SDA算法的预测结果与真实交通流量拟合度最好，在整个交通流量曲线上突刺（偏差较大处）偏少，而其他3种算法尤其是SVR算法在整个曲线上存在更多突刺，而KNN算法预测值则与真实交通流量值之间存在明显偏差。KNN-SDA算法优于其他3种算法的原因为：① KNN算法比较适用解决短时交通流量预测这类具有时变、不确定的非线性问题；② 本文为全天144个时刻分别分配合理的训练数据集与K值，参数设置的优化使KNN-SDA算法能够更加准确地搜索到所需历史数据。

4　结论

非参数回归摒弃了传统求解数学模型的方法，仅在历史数据库的支持下，就可较好地适应短时交通流量预测问题的不确定性、时变性和非线性。只

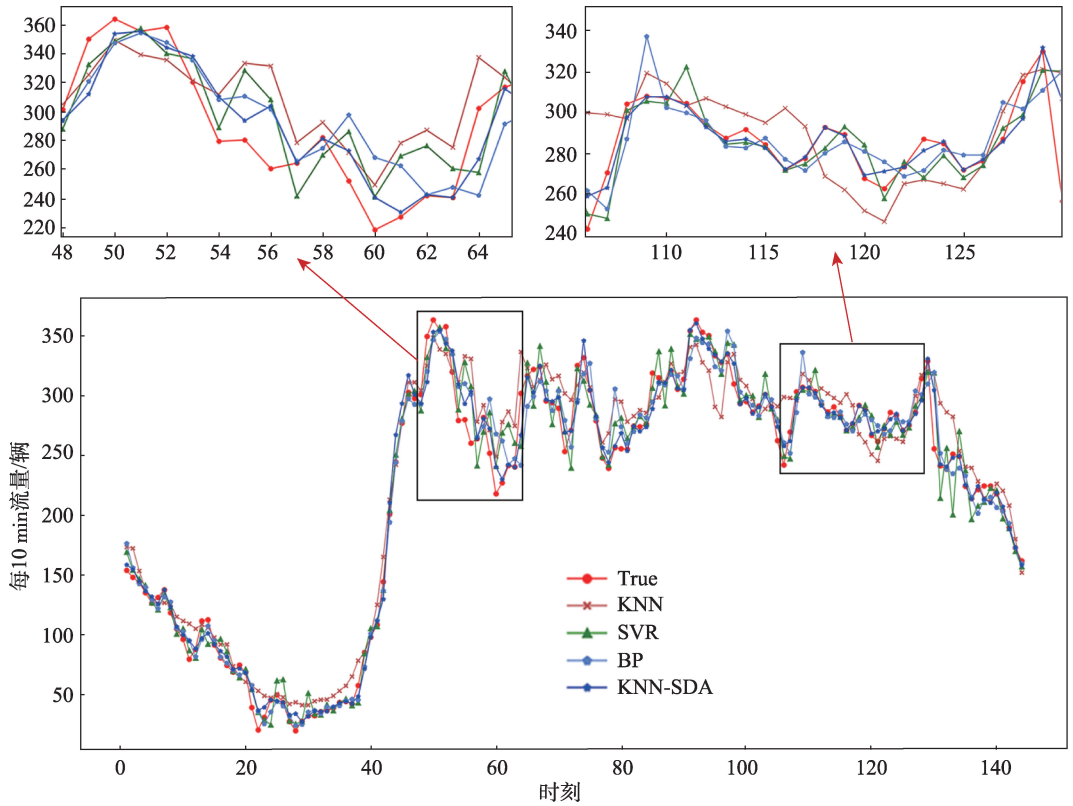

图6 不同算法预测结果

Fig. 6 Prediction results of different models

要有完备的历史数据,并制定适宜的交通流量模式匹配机制,即可取得满意的预测结果。本文中对KNN方法进行了若干改进,包括与预测密切相关的输入变量X、训练数据集与K值等,这些改进优化了KNN模式匹配精度。研究表明,与在短时交通流量预测研究领域常用的支持向量机(SVR)与BP神经网络等模型相比,本文提出的KNN-SDA算法具有训练参数少、预测精度高、时效性好等相对优势。但本文算法(仅)适用于单路口短时交通流量预测,对于不同的路口,仍需要构造对应的训练数据集与K值。如何改进该算法使其适用于大规模城市路网短时交通流量预测是后续研究努力的方向。

参考文献(References):

[1] Sun S, Zhang C, Yu G. A bayesian network approach to traffic flow forecasting[J]. IEEE Transactions on Intelligent Transportation Systems, 2006,7(1):124-132.

[2] Guo J, 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.

[3] 赵建东,王浩,刘文辉.高速公路旅行时间的自适应插值卡尔曼滤波预测[J].华南理工大学学报(自然科学版),2014,42(2):109-115. [Zhao J D, Wang H, Liu W H. Prediction of expressway travel time based on adaptive interpolation Kalman filtering[J]. Journal of South China University of Technology (Natural Science Edition), 2014,42(2):109-115.]

[4] Oh S D, Kim Y J, Hong J S. Urban traffic flow prediction system using a multifactor pattern recognition model[J]. IEEE Transactions on Intelligent Transportation Systems, 2015,16(5):2744-2755.

[5] Xing Y, Ban X, Guo C. Probabilistic forecasting of traffic flow using multikernel based extreme learning machine [J]. Scientific Programming, 2017(2):1-12.

[6] 商强,杨兆升,李志林,等.基于相空间重构和RELM的短时交通流量预测[J].华南理工大学学报(自然科学版),2016,44(4):109-114. [Shang Q, Yang Z S, Li Z L, et al. Short-term traffic flow prediction based on phase space reconstruction and RELM[J]. Journal of South China University of Technology (Natural Science Edition), 2016,44(4):109-114.]

[7] 杨兆升,王媛,管青.基于支持向量机方法的短时交通流量预测方法[J].吉林大学学报(工学版),2006,36(6):881-884. [Yang Z S, Wang Y, Guan Q. Short-term traffic

flow prediction method based on SVM[J]. Journal of Jilin University (Engineering and Technology Edition), 2016, 36(6):881-884.]

[8] 姚卫红,方仁孝,张旭东.基于混合人工鱼群优化SVR的交通流预测模型[J].大连理工大学学报,2015,55(6):632-637. [Yao W H, Fang R X, Zhang X D. Traffic flow forecasting based on optimized SVR with hybrid artificial fish swarm algorithm[J]. Journal of Dalian University of Technology, 2015,55(6):632-637.]

[9] Zhang L, Liu Q, Yang W, et al. An improved K - nearest neighbor model for short-term traffic flow prediction[J]. Procedia - Social and Behavioral Sciences, 2013,96:653-662.

[10] Bernaś M, Płaczek B, Porwik P, et al. Segmentation of vehicle detector data for improved k- nearest neighbours-based traffic flow prediction[J]. IET Intelligent Transport Systems, 2014,9(3):264-274.

[11] Davis G A, Nihan N L. Nonparametric regression and short-term freeway traffic forecasting[J]. Journal of Transportation Engineering, 1991,117(2):178-188.

[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, 2002,10 (4):303-321.

[13] 刘洋,马寿峰.基于聚类分析的非参数回归短时交通流预测方法[J].交通信息与安全,2013,31(2):27-31. [Liu Y, Ma S F. Non-parametric regression for short-term traffic flow forecasting based on cluster analysis[J]. Journal of Transport Information and Safety, 2013,31(2):27-31.]

[14] 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. [Zhang X L, Lu H P. Non-parametric regression and application for short-term traffic flow forecasting[J]. Journal of Tsinghua University (Science and Technology), 2009,49(9):1471-1475.]

[15] 张涛,陈先,谢美萍,等.基于K近邻非参数回归的短时交通流预测方法[J].系统工程理论与实践,2010,30(2):376-384. [Zhang T, Chen X, Xie M P, et al. K-NN based non-parametric regression method for short-term traffic flow forecasting[J]. Systems Engineering- Theory & Practice, 2010,30(2):376-384.]

[16] 张晓利,贺国光,陆化普.基于K-邻域非参数回归短时交通流预测方法[J].系统工程学报,2009,24(2):178-183. [Zhang X L, He G G, Lu H P, Short-term traffic flow forecasting based on K-nearest neighbors non-parametric regression[J]. Journal of Systems Engineering, 2009,24 (2):178-183.]

[17] Meng M, Wang B B, Shao C F, et al.A two-stage short-term traffic flow prediction method based on AVL and AKNN techniques[J]. Journal of Central South University, 2015,22(2):779-786.

[18] Frank B. Time-Aware Multivariate Nearest Neighbor Regression Methods for Traffic Flow Prediction[J]. IEEE Transactions on Intelligent Transportation Systems, 2015, 16(6):3393-3402.

[19] Xia D, Li H, Wang B, et al. A map reduce-based nearest neighbor approach for big-data-driven traffic flow prediction[J]. IEEE Access, 2016,4:2920-2934.

[20] 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. [Zhang X L. A study on short-term traffic volume forecasting based on non-parametric regression[D]. Tianjin: Tianjin University, 2007.]

[21] 吕小青,曹彪,曾敏,等.确定延迟时间互信息法的一种算法[J].计算物理,2006,23(2):184-188. [Lv X Q, Cao B, Zeng M, et al. An algorithm of selecting delay time in the mutual information method[J]. Chinese Journal of Computational Physics, 2006,23(2):184-188.]

[22] 翟东海,李同亮,段维夏,等.基于矩阵相似度的最佳样本块匹配算法及其在图像修复中的应用[J].计算机科学,2014,41(1):307-310. [Zhai D H, Li T L, Duan W X, et al. Optimal exemplar metching algorithm based on matrix similarity and its application in image inpainting[J]. Computer Science, 2014,41(1):307-310.]