

一种修复交通流异常数据的改进 KNN 算法^{*}

李翠, 黄侃, 李霞

(江西交通职业技术学院 信息工程系, 江西 南昌 330013)

摘要: 现场采集的交通流数据经常出现错误与缺失等异常现象。为有效修复交通流异常数据, 文中提出一种改进的 KNN 算法——相关系数—调幅权重法。该算法采用每天采集的所有数据构造固定长度状态向量, 以适应异常数据随机分布的特点; 采用相关系数筛选近邻指标, 以满足近邻须为相似交通流的要求; 采用相关系数和调幅系数对传统距离倒数权重进行修正, 以克服相关系数过小和近邻距离过远带来的不利影响。基于实测交通流数据进行蒙特卡洛分析, 结果表明该算法具有良好的数据修复能力, 能适应近邻相关系数与欧式距离严重背离的情况。

关键词: 公路交通; 交通流; 异常数据; 数据修复; 改进 KNN 算法

中图分类号: U491.1

文献标志码: A

文章编号: 1671-2668(2022)04-0039-05

交通流数据为智能交通系统的正常运行提供数据支撑, 其中蕴含的交通时空分布规律对现代交通的科学管理与决策及交通流的基础理论研究具有重要价值。但由于设备故障、传输问题和处理错误等原因, 现场采集的交通流数据经常出现错误与缺失等异常现象, 需对异常数据进行修复处理。

交通流具有很强的周期性和随机性, 因而常基于历史数据建立交通流异常数据修复模型。常见修复方法可分为传统数学方法、智能方法和混合方法, 其中智能方法包括非参数回归法、深度学习法等。作为一种非参数回归方法, K 近邻算法 (K-Nearest Neighbors, KNN) 完全由数据驱动而无需假设数据的分布模式, 原理简单且易于扩展。在交通领域, KNN 法主要用于交通流预测。近年来, 将 KNN 算法应用于交通流异常数据修复的研究也受到关注, 如文献[8]通过优选近邻数量和调整近邻权重来修复交通流异常数据, 但仅讨论了异常数据非连续分布的情况。实际上, 异常数据的位置是随机的, 也存在连续分布的情况。为适应异常数据随机出现的特点, 该文提出改进 KNN 算法来实现交通流异常数据的修复。

1 KNN 法的基本原理

KNN 法完全由数据驱动, 通过搜索历史数据库寻找具有合适数量、与当前数据(包含异常数据)

相似的多个近邻, 并对近邻数据进行加权组合以实现异常数据的修复。基本流程为构造具有较大容量且有代表性的历史数据库→设定合适的状态向量→计算异常数据状态向量与历史数据状态向量之间的距离→设定距离阈值后选取近邻个数 k →选取近邻权重计算方式→对 k 个近邻进行加权计算, 获得异常数据修复值。

根据以上流程, KNN 法的修复性能同时取决于状态向量的构造方式、状态向量之间距离的度量方式及近邻个数和近邻权重的选取方法。状态向量作为当前数据与历史数据相似程度的匹配模式, 表征历史数据库中的数据特征。距离度量方式用于计算异常数据状态与历史数据库中各状态向量之间的距离。近邻个数 k 表示从历史数据库中选取与当前数据相似的数据组数, k 值的确定与历史数据库的容量及设定的距离阈值相关。各近邻的权重系数一般按经验指定, 决定 k 个近邻对修复异常数据的贡献。

2 改进 KNN 算法

筛选近邻之前需构造状态向量。状态向量的长度一般较小, 如文献[8]仅利用 5 个连续采样点来构造状态向量, 状态向量中间的元素为异常数据。这么短的状态向量很难满足随机出现且连续分布的异常数据修复需求, 利用较多的采样点来构造较长的

^{*} 基金项目: 江西省交通运输厅科技项目(2020X0012); 江西省教育厅科学技术研究项目(191325)

状态向量更合适。此外,近邻的筛选方式及近邻权重的取值都应进行改进,以提高修复精度。

2.1 状态向量的构造

历史数据状态向量为:

$$H_i = [h_{i1}, h_{i2}, \dots, h_{in}]^T; i=1, 2, 3, \dots, p \quad (1)$$

式中: h_{ij} 为第*i*个历史数据状态向量的第*j*个采样点数据; n 为每个状态向量包含的采样点数量,可取较大的值(如100左右); p 为历史数据状态向量的数量。

相应地,某待修复的异常数据状态向量为:

$$A = [a_1, a_2, \dots, a_n]^T \quad (2)$$

式中: a_i 为异常数据状态向量的第*i*个采样点。

需指出的是, A 中同时包含正常数据和异常数据。

2.2 近邻的筛选指标

常用欧式距离衡量2个状态向量之间的匹配程度(相似程度),以便从历史数据状态向量库中筛选出异常数据状态向量的近邻。欧式距离的表达式为:

$$l_i = \sum_{m \notin E}^n (h_{im} - a_m)^2 \quad (3)$$

式中: l_i 为异常数据状态向量和第*i*个历史数据状态向量之间的欧式距离; E 为异常数据对应采样点的集合。

考虑到相关系数更能反映状态向量的相似性,有利于判断相似交通流,采用相关系数代替欧式距离进行近邻筛选。相关系数的表达式为:

$$c_i = \frac{\sum_{m \notin E}^n (h_{im} - \bar{h}_i)(a_m - \bar{a})}{\sqrt{\left(\sum_{m \notin E}^n (h_{im} - \bar{h}_i)^2\right) \left(\sum_{m \notin E}^n (a_m - \bar{a})^2\right)}} \quad (4)$$

式中: c_i 为异常数据状态向量和第*i*个历史数据状态向量之间的相关系数; $\bar{h}_i$ 、 $\bar{a}$ 分别为 H_i 和 A 剔除异常数据后的均值。

根据式(3)或式(4)可将欧式距离最小或相关系数最大的*k*个历史数据状态向量视作异常数据状态向量的近邻。

2.3 异常数据的修复

经典的异常数据修复公式为:

$$a_m = \sum_i^k \alpha_i h_{im}, m \in E \quad (5)$$

式中: α_i 为第*i*个近邻的权重; k 为近邻数量。

近邻权重有多种取值方法,常见的有历史均值法和距离权重法,计算公式分别为:

$$\alpha_i = 1/k \quad (6)$$

$$\alpha_i = \frac{l_i^{-1}}{\sum_m^k l_m^{-1}} \quad (7)$$

近邻与异常数据具有良好的相似性是式(5)成立的前提,这种相似性同时体现在相关系数和欧氏距离这2个指标上。显然,较大的相关系数和较小的欧氏距离有利于异常数据的修复。但挑选出的近邻可能存在如下情况:1)欧氏距离相当但相关系数差别较大;2)欧式距离过大造成近邻与异常数据的曲线明显偏离。因此,同时采用相关系数和调幅系数对式(7)进行修正,得到新的近邻权重:

$$\alpha_i = c_i \gamma_i \frac{l_i^{-1}}{\sum_m^k l_m^{-1}} \quad (8)$$

式中: γ_i 为调幅系数, $\gamma_i = \sum_{m \notin E}^n a_m / \sum_{m \notin E}^n h_{im}$; $\sum_{m \notin E}^n a_m$ 、 $\sum_{m \notin E}^n h_{im}$ 分别为 A 、 H_i 剔除异常数据后的交通流总量。

采用 c_i 进行修正的目的是希望相关系数更大的近邻能获得更大的近邻权重。采用 γ_i 进行修正的目的是将 H_i 的幅值拉回到 A 的幅值水平,最终目的是对式(5)中 h_{im} 进行调整。因此,式(8)给出的近邻权重计算方法可称为调幅权重法。

2.4 改进 KNN 算法流程

改进 KNN 算法流程为按式(1)、式(2)构造较长的状态向量→按式(4)计算相关系数→挑选相关系数最大的前*k*个历史数据作为近邻→按式(8)计算近邻权重→按式(5)进行异常数据修复。改进 KNN 算法也可称为相关系数—调幅权重法。

3 评价指标

均方根误差 RMSE 是最常见的用来评价异常数据修复效果的指标, RMSE 越小,修复效果越好。其表达式为:

$$RMSE = \sqrt{\frac{1}{q} \sum_{j=1}^q (a_j^r - a_j^*)^2} \times 100 \quad (9)$$

式中: q 为异常数据个数; a_j^r 为异常数据的修复值; a_j^* 为异常数据的真实值。

另一个常见评价指标是平均相对误差 MRE, MRE 越小,修复效果越好。其表达式为:

$$MRE = \frac{1}{q} \sum_{j=1}^q \left| \frac{a_j^* - a_j^r}{a_j^*} \right| \times 100 \quad (10)$$

为避免 a_j^* 过小(如为 0、1 等)导致 MRE 指标评价失真,采用 $RMSE$ 评价改进 KNN 算法的修复精度。

4 算例分析

4.1 数据来源与处理

采用明尼苏达大学交通数据研究实验室在 2018 年 1 月 12 日—2019 年 1 月 12 日(共 366 d)采集的交通流量数据(第 25 号测点)展开分析。原始数据的采样间隔为 30 s,每天共有 2 880 个采样点。经统计,共有 146 d 未出现异常数据。在出现异常数据的 220 d 中,有 218 d 的数据异常率小于 6.32%,另外 2 d 的数据异常率分别为 15.4% 和 46.5%。此外,异常数据呈现随机分布的特点。

为检验上述异常数据修复算法的性能,将 146 d 的正常数据按 15 min 采样间隔重新进行整理(96 个采样点),形成实验数据库。以前面 142 d 的数据作为历史数据,后面 4 d 的数据作为测试数据(分别称为测试数据 1、2、3、4)。

4.2 算法的修复性能比较

分别采用表 1 所示 6 种不同修复算法进行处理,比较其修复性能。这些算法均以天为时间单位构造较长的状态向量,每个状态向量均包含 96 个采样数据。算法 1~3 采用相关系数[式(4)]筛选近邻,算法 4~6 采用欧式距离[式(3)]筛选近邻。

表 1 6 种异常数据修复算法

算法编号	算法名称	公式组合
算法 1	相关系数—调幅权重法	式(4)+式(8)
算法 2	相关系数—距离权重法	式(4)+式(7)
算法 3	相关系数—历史均值法	式(4)+式(6)
算法 4	欧式距离—调幅权重法	式(3)+式(8)
算法 5	欧式距离—距离权重法	式(3)+式(7)
算法 6	欧式距离—历史均值法	式(3)+式(6)

由于异常数据的出现位置具有很强的随机性,需从统计意义上比较各种修复算法的优劣。采用蒙特卡洛抽样方法进行统计分析。每次抽样时,先根据指定的数据异常率生成异常数据数量,再随机生成它们的位置。需指出的是,在筛选近邻时,最优 k 值难以预先确定,故采用动态方法确定 k 值。每次抽样均按式(4)计算相关系数,对于算法 1~3,将相关系数大于 0.95 的历史数据视作

近邻,且 k 值至少取 10,最多取 20;对于算法 4~6, k 值与算法 1~3 相同,将欧式距离最小的 k 个历史数据视作近邻。

考虑 20% 的数据异常率,对每个测试数据进行 5 000 次随机抽样,随机生成异常数据的位置。对计算结果进行统计分析,得到 6 种算法的 $RMSE$ 值(见图 1)。

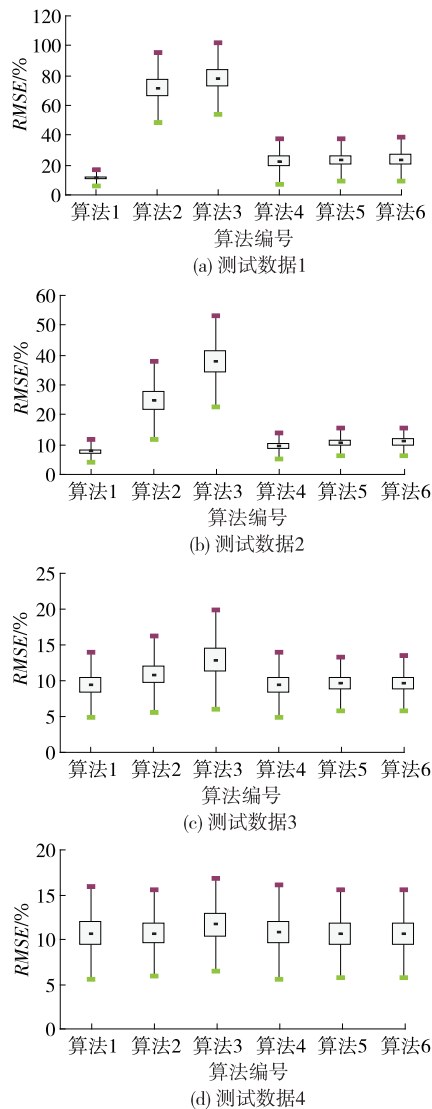


图 1 不同算法的 $RMSE$ 值(数据异常率=20%)

由图 1 可知:对于测试数据 1、2,算法 1 远好于其他算法,算法 2、3 最差,算法 4~6 相当;对于测试数据 3,从中值判断,算法 1 远好于算法 2、3,但仅比算法 4~6 稍好;对于测试数据 4,从中值判断,算法 1 稍好于其他 5 种算法。总的来说,对于测试数据 1、2,算法 1 表现最好;对于测试数据 3、4,算法 1 与算法 4~6 相当。可见,算法 1 对异常数据的修复能力具有更广泛的适应性。

4.3 近邻的相关系数与欧式距离

将不同算法在某次抽样分析时筛选出的近邻挑选出来,绘制图2所示相关系数与欧式距离之间的关系图,分析算法1具有更广泛适应性的原因。

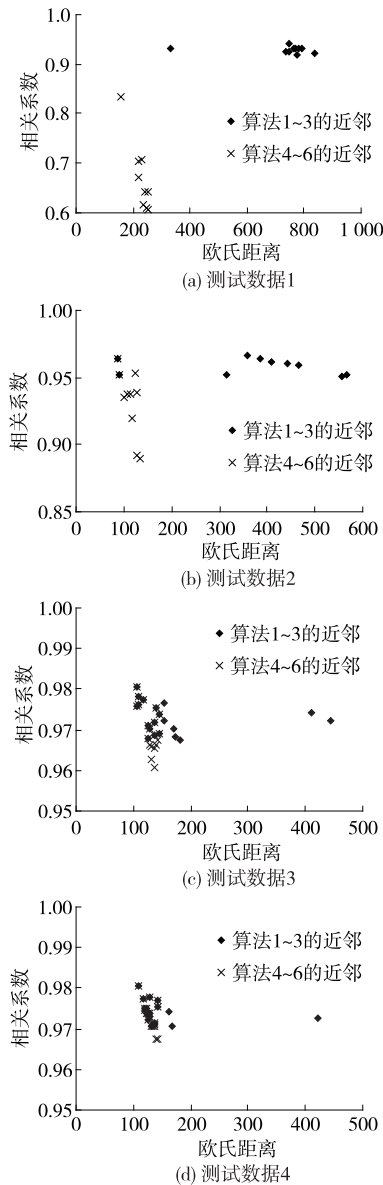


图2 不同算法 k 个近邻的相关系数—欧式距离关系

由图2可知:测试数据1和2、测试数据3和4的散点分布规律分别较为相似。对于测试数据1、2,算法1~3挑选出的近邻大多具有较大的相关系数和欧式距离,而算法4~6挑选出的近邻大多具有较小的相关系数和欧式距离。在筛选近邻时,总是希望近邻具有较大的相关系数和较小的欧式距离。对于测试数据1、2,两类算法(算法1~3为一类,算法4~6为另一类)挑选出的近邻的相关系数和欧式距离背离严重;对于测试数据3、4,算法1~3和算

法4~6挑选出的近邻均大多具有较大的相关系数和较小的欧式距离,两类算法挑选出的近邻的相关系数和欧式距离背离程度较轻。

图3为测试数据与所挑选近邻之间的关系。由3(a)可知:测试数据与该近邻之间的距离明显较远($l_i=411.98$),而实际上它们之间的相关系数较大($c_i=0.974$)。由图3(b)可知:测试数据与该近邻之间的距离明显较近($l_i=105.82$),而实际上它们之间的相关系数也较大($c_i=0.980$)。对于图3(a)所示欧式距离很大的情况,若仍采用历史均值法和距离权重法计算近邻权重,则异常数据的修复值会发生严重失真,采用经调幅系数修正的调幅权重法计算近邻权重更合理。图1表明,调幅权重法只有和相关系数筛选近邻法相结合时(即算法1)才具有最佳的修复效果。

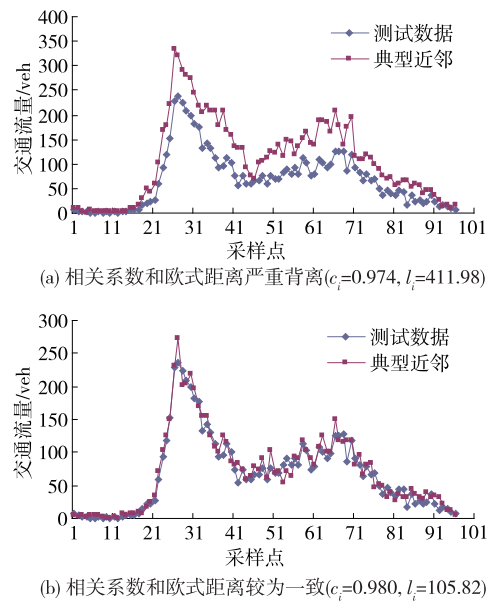


图3 测试数据与所挑选近邻之间的典型关系

综合图1~3分析,算法1比其他算法具有更广泛的适应性,即使对于近邻相关系数与欧式距离严重背离的情况,算法1也具有较好的修复性能。而其他算法只有在近邻相关系数与欧式距离较为一致时才表现出较好的修复性能。

4.4 数据异常率的影响

将数据异常率以10%步长从10%增加至90%,对每个数据异常率进行5000次抽样,分析算法1的修复性能与数据异常率的关系。不同数据异常率下算法1的RMSE值见图4。

由图4可知:对于所有测试数据,随着数据异常率的增大,箱形图的中值和下限上升,表明算法1的

修复性能下降; $RMSE$ 指标的变化幅度先减小后增大,在数据异常率约为 50% 时最小。这是由于数据异常率较小时,异常数据容易集中在数值较大或较小的采样点上, $RMSE$ 指标出现较大值或较小值。而当数据异常率较大时,由于正常数据较少,算法 1 的修复性能变差,进一步导致 $RMSE$ 指标的变化幅度增大。

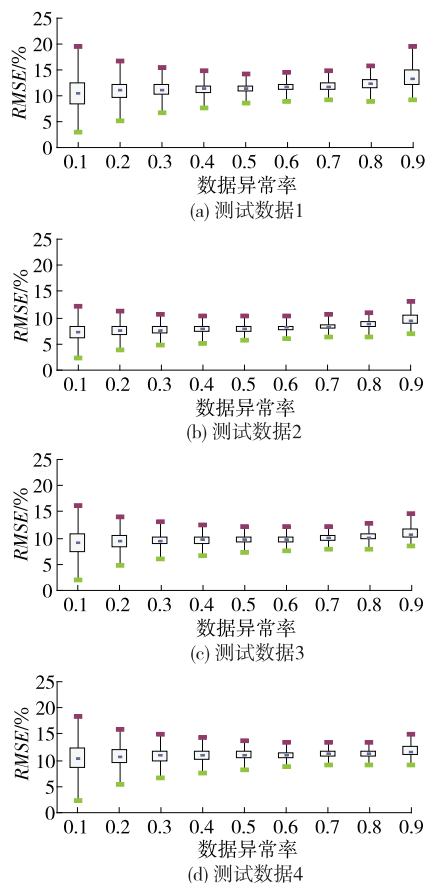


图4 算法1在不同数据异常率下的 $RMSE$ 值

5 结论

(1) 构造较长且长度固定的状态向量可很好地适应异常数据随机分布的特点。

(2) 按式(3)或式(4)挑选出的近邻有时不能同时满足相关系数较大且欧式距离较小的要求。对于欧式距离较大的情况,采用经调幅系数修正的调幅权重法来计算近邻权重更合理。

(3) 相关系数—调幅权重法具有广泛的适应性,即使在近邻相关系数与欧式距离严重背离时,该算法也具有较好的修复性能。

参考文献:

[1] 李翠,李雪.基于车辆积压长度的高速公路交通事件检

测算法[J].公路与汽运,2021(1):21—28.

- [2] LUO X L, MENG X, GAN W J, et al. Traffic data imputation algorithm based on improved low-rank matrix decomposition[J]. Journal of Sensors, 2019(12): 1—11.
- [3] 姜桂艳,江龙晖,张晓东,等.动态交通数据故障识别与修复方法[J].交通运输工程学报,2004,4(1):121—125.
- [4] 胡伟超,刘君,丛浩哲,等.实时交通流数据缺失修复方法对比研究[C]//中国智能交通协会.第八届中国智能交通年会论文集.2013:205—209.
- [5] 韩卫国,王劲峰,胡建军.交通流量数据缺失值的插补方法[J].交通与计算机,2005(1):39—42.
- [6] 郭敏,蓝金辉,李娟娟,等.基于灰色残差 $GM(1, N)$ 模型的交通流数据恢复算法[J].交通运输系统工程与信息,2012,12(1):42—47.
- [7] 王薇,程泽阳,刘梦依,等.基于时空相关性的交通流故障数据修复方法[J].浙江大学学报(工学版),2017,51(9):1727—1734.
- [8] 秦一菲,马明辉,王岩松,等.基于改进 KNN 算法的交通流异常数据修复方法[J].计算机测量与控制,2018,26(12):180—184.
- [9] ZHAO J, NIE Y, NI S, et al. Traffic data imputation and prediction: An efficient realization of deep learning[J]. IEEE Access, 2020(8): 46713—46722.
- [10] AYDILEK I B, ARSIAN A. A novel hybrid approach to estimating missing values in databases using K-nearest neighbors and neural networks[J]. Innovative Computing Information & Control, 2012, 8(7): 4705—4717.
- [11] ZHENG Z, SU D. Short-term traffic volume forecasting: A k-nearest neighbor approach enhanced by constrained linearly sewing principle component algorithm[J]. Transportation Research Part C: Emerging Technologies, 2014, 43(1): 143—157.
- [12] 罗向龙,李丹阳,杨彧,等.基于 KNN-LSTM 的短时交通流预测[J].北京工业大学学报,2018,44(12): 1521—1527.
- [13] 谢海红,戴许吴,齐远.短时交通流预测的改进 K 近邻算法[J].交通运输工程学报,2014,14(3): 87—94.
- [14] HABTEMICHAEL F G, CETIN M. Short-term traffic flow rate forecasting based on identifying similar traffic patterns[J]. Transportation Research Part C: Emerging Technologies, 2016, 66: 61—78.
- [15] LUO X, LI D, YANG Y, et al. Spatiotemporal traffic flow prediction with KNN and LSTM[J]. Journal of Advanced Transportation, 2019(5): 1—10.

收稿日期:2021—11—16