Advances in Applied Mathematics 应用数学进展, 2023, 12(10), 4350-4356
Published Online October 2023 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2023.1210428

Hans汉斯

基于K-Means交通流量空间特征分析方法

王佳程，苏庆华*，田宝平，顾敏杭，宫瑞英，李　泳

北京物资学院信息学院，北京

收稿日期：2023年9月17日；录用日期：2023年10月11日；发布日期：2023年10月18日

摘　要

随着汽车保有量的上升，交通流量预测已经成为当前研究的一个重点。针对交通中空间特征利用率点的问题，本文针对道路具有的空间特征，提出将空间路段分成四种上下路段关系，并对不同路段上的路段ID、路段长度宽度特征，利用K-means对空间交通流量存在相互的内在影响进行空间特征分析。本方法充分考虑交通流量数据在空间上的特性，利用K-means聚类深度分析空间特征因素及其之间的关系，完成交通流量空间特征分析，为更精准的时空联合预测处理提供更有效的数据。

关键词

K-Means，交通流量，影响因素，空间特征

K-Means Traffic Flow Prediction Method Based on Spatial Features

Jiacheng Wang, Qinghua Su*, Baoping Tian, Minhang Gu, Ruiying Gong, Yong Li

School of Information, Beijing Wuzi University, Beijing

Received: Sep. 17th, 2023; accepted: Oct. 11th, 2023; published: Oct. 18th, 2023

Abstract

With the increase in car ownership, traffic flow forecasting has become a focus of current research. Aiming at the problem of utilization points of spatial features in traffic, this paper proposes to divide road sections into four kinds of relations between upper and lower sections and analyze the spatial characteristics of road ID, road length, and width on different sections by using K-means to analyze the internal influence of spatial traffic flow. This method fully considers the spatial characteristics of traffic flow data, uses K-means clustering to analyze the spatial feature factors and

*通讯作者。

文章引用: 王佳程, 苏庆华, 田宝平, 顾敏杭, 宫瑞英, 李泳. 基于 K-Means 交通流量空间特征分析方法[J]. 应用数学进展, 2023, 12(10): 4350-4356. DOI: 10.12677/aam.2023.1210428

the relationship between them in-depth, completes the spatial feature analysis of traffic flow, and provides more effective data for more accurate spatial-temporal joint prediction processing.

Keywords

K-Means Algorithm, Traffic Flow, Influencing Factors, Spatial Features

1. 引言

道路的交通流量预测问题是一个在空间上具有相关性的经典数据预测问题[1]，准确的交通预测能推进城市智慧交通的发展，利于交通管理部门管理，便于出行者及时调整路线[2]。交通流量数据特性在空间上体现为静态。在复杂的城市道路交通中，交通流量路网实际状态往往呈现拓扑结构，而这种结构也就决定了节点间的交通情况会存在互相作用。通常，上游路段通过传导机制作用于下游路段，但遇拥堵状况，下游路段通过反馈机制对上游路段进行作用[3]。所以，某一时刻的交通流量与道路状况结构有很大关系。而整个系统中有很多路段情况相似，因此可以对相似的路段进行聚类分析。

K-means 算法是实现便捷、运算快、在分析较大样本时效率高的一种可调聚类方法[4] [5] [6] [7]。已有学者采用 K-means 算法进行处理分析，提升处理精度[8] [9]。本文采用 K-means 算法来对交通流量预测中的空间特征进行预测。

2. 交通数据空间特征

在空间分析上，已知与路段有关系的特征有上下路段，路段的长度，宽度，路段 ID 和路的上下游属性，而这些特征对空间交通流量预测均有影响[10]。

2.1. 上下路段特征

道路的每一路段都和相关路段有关，通过固定时间点来分析空间上的特点，道路路段可分为四类，其中 ni 表示当前可进入道路的个数，n_o 表示当前可以出的道路的个数。一对一道路：指一条道路进入，且只有一条道路出去，即 $n_i = n_o = 1$。一对多道路：指一条道路进入，可选择多条道路中的任意一条出去，$n_i = 1$，$n_o > 1$。多对一道路：指可选择多条道路中的任意一条进入，只可选择一条道路出去，$n_i > 1$，$n_o = 1$。多对多道路：是指可选择多条道路中的任意一条进入，可选择多条道路中的任意一条出去，$n_i > 1$，$n_o > 1$。

2.2. 路段的 ID

路段 ID 由于其唯一性，被用于识别路段，可用于进行空间特征分析。但由于路段 ID 属于分类变量，无法直接被模型识别。需要先对其做编码操作，将其转换成数字进行输入，做类别变量的处理。本采用的是 label encoder 方式，通过将 n 类不同的标签名编码为[0, n − 1]之间的整数，构建一对一的映射关系。

2.3. 路段的长度

路段长度一定会对交通流量数据产生影响。路段的长度越大，通过这条路所花费的时间就会越多。

2.4. 路段的宽度

路段宽度会对交通流量数据产生影响，但影响系数要依据分析来判断。相同条件下，路段的宽度越大，允许通过的车辆多，通过这条路所花费的时间就会越小，不易堵车。

3. K-Means 空间特征分析

3.1. K-Means

K-means 作为经典的一种聚类算法，可应用于道路交通中的空间特征分析。K-means 聚类的策略是通过最小化损失函数来进行最优的划分，直至满足划分停止条件，即：划分后的每个样本到其所归类中心的距离总是要小于其到其他类中心的距离。使类内的数据样本尽可能地归在一起，并且每类间的数据样本距离尽量大利用距离或者相似度将数据划分为 K 类[11]。

3.2. K-Means 空间特征分析

K-means 的核心目标是将给定的交通数据集划分成 K 个簇，K 为超参，并给出每个样本数据对应的中心点。利用 K-means 进行交通流量空间分析算法如下：

① 初始化。在给定的数据集中，任选 K 个交通数据样本数据作起始的聚类中心，此时 K 表示初始要聚类的数目。

② 空间特征样本聚类。计算交通数据集中每个样本到所有起始聚类中心的欧氏距离，让每个样本分配到最近距离的中心的类去，形成 K 类。欧氏距离计算公式为：

$$D=\left(\sum_{i=1}^{K}\left\|X-U_i\right\|^2\right)^{\frac{1}{2}} \tag{1}$$

式中，$X\in$类K_i 中心的样本，U_i 为当前样本，下同。

③ 计算新的聚类类中心。按已有的聚类结果，把现在每个聚类中样本的均值算出，更新为新的聚类中心。计算公式为：

$$U_i=\frac{1}{K_i}\sum_{X\in K_i}X \tag{2}$$

④ 如果迭代收敛或者满足停止条件，算法结束。否则令，重复进行步骤② ③。迭代收敛停止条件为：

$$\left|U_{n+1}-U_n\right|\le\alpha \tag{3}$$

式中，α 为某一具体收敛数值。

损失函数最小的停止条件。其中损失函数是将样本与其所划分类的中心之间的距离的总和，即：

$$W=\sum_{i=1}^{K}\sum_{X\in K_i}\left\|X-U_i\right\|^2 \tag{4}$$

为将所有路段按照通过该路段的时间进行划分，在相同时间段的条件下，以该时间段下路段通过时间差的欧氏距离为指标进行 K-means 聚类。划分出几个类，代表不同的通过时间范围。

3.3. K 值的选择

选择用肘部法进行判断 K 值。肘部法是一种用来确定聚类数量的方法，它的原理是计算不同聚类数量下的类别畸变程度之和，也就是每个样本到其所属类别中心的距离平方和。当聚类数量增加时，类别畸变程度之和会逐渐减小，但是当达到某个临界点时，减小的幅度会变得很小，这个临界点就可以看作

是一个合适的聚类数量。如下图 1 所示的折线图，横坐标表示的是 k 的取值，纵坐标表示的是误差平方和(用 SSE 表示)，SSE 用于计算拟合数据和原始数据对应点的误差平方和，找到拐点位置，作为合适的聚类数量。

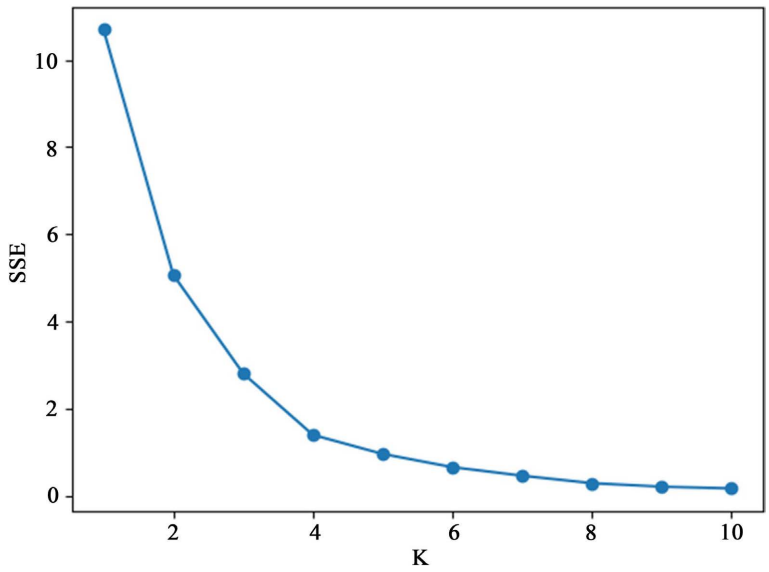


Figure 1. The relationship between k value and SSE
图 1. k 取值与 SSE 关系图

从图 1 中可以明显看出，当拐点位置在 k = 4 时，减小的幅度变得很小，可以选择 k = 4 作为合适的聚类数量。

4. 实验分析

4.1. 实验环境

本实验硬件采用 AMD Ryzen 7 4800H 作为 CPU，主频为 2.9 GHz，内存为 16 GB；显卡使用 GeForce RTX 3060，显存为 12 G；操作系统为 Windows 10 (64bit)。编程语言使用 Python3.9，开发程序选择 Pycharm。

4.2. 实验数据

数据集：智慧交通预测赛(阿里天池平台)，共 9,126,031 条记录，大小为 735 MB。首先对数据集进行异常值裁剪，以减少异常值在后续分析中带来的敏感性。

在分析和处理过程中，以固定时间来分析空间，在固定时间段中对选取的所有路段进行 K-means 空间特征分析。考虑到不同日期的交通流量不同，如工作日、周末、节假日等，通过路段的时间也不相同，将时间固定为工作日、周末和节假日三类进行分析。同时，将时间段聚焦在早上八点到九点，下午十五点到十六点，晚上十八点到十九点这三个时间段，以达到最大的实验效果。

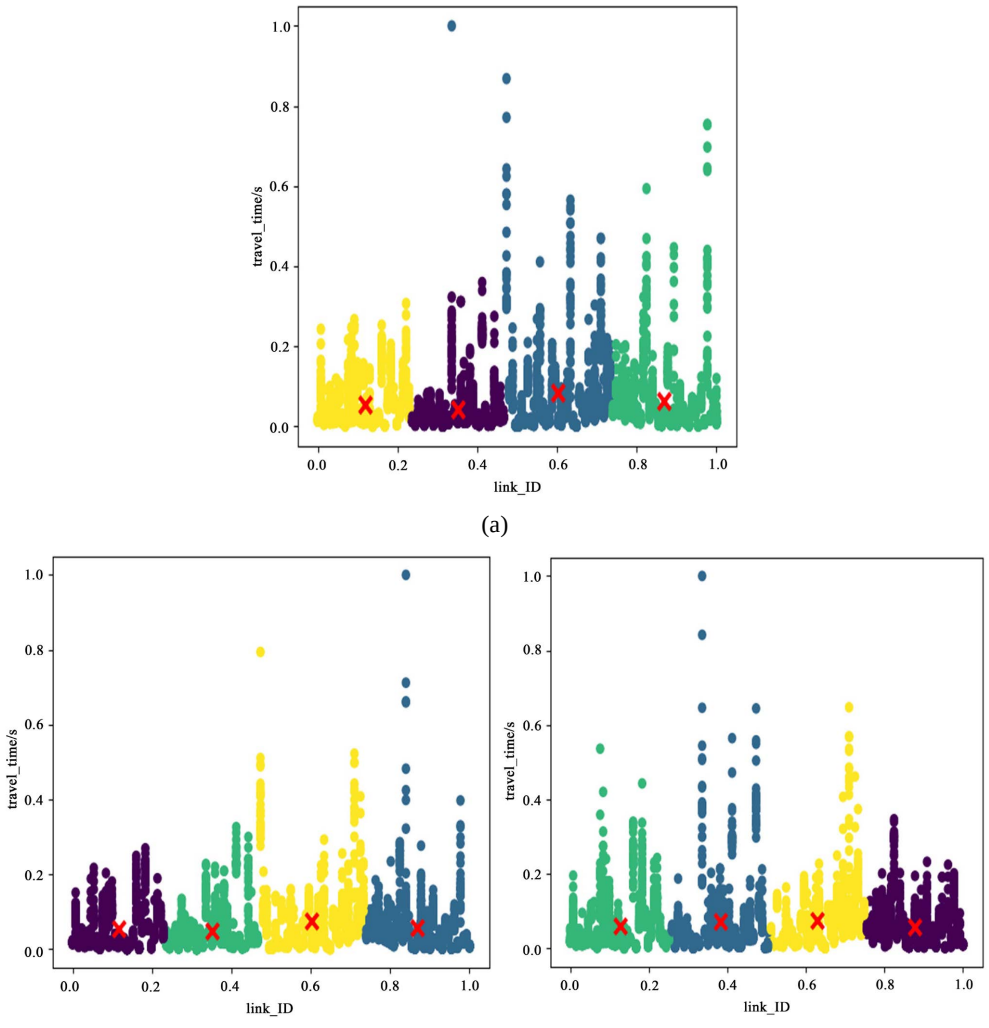
4.3. 实验结果及其分析

因为空间特征中的不同属性往往具有不同的取值范围，每个量之间的差别可能很大，为了消除每个指标之间由于取值范围带来的差异，先进行规范化处理。采用的是最小－最大规范化方法，将特征数据按照比例进行缩放，使不同属性的数值统一到[0, 1]区间，以方便综合分析。图 2 为规范化后的路段聚类图，从图中可以看出有四类聚类结果。

	link_ID	length	width	link_class	聚类结果
0	4377906289869500514	57	3	1	3
1	4377906284594800514	247	9	1	1
2	4377906289425800514	194	3	1	3
3	4377906284525800514	839	3	1	0
4	4377906284422600514	55	12	1	2
..	...	...	...	...	...
127	4377906288421600514	128	9	1	1
128	4377906289041600514	208	9	1	1
129	4377906286032600514	257	12	1	2
130	4377906281234600514	16	3	1	3
131	4377906286334600514	33	15	1	2

Figure 2. Spatial feature analysis diagram
图 2. 空间特征分析图

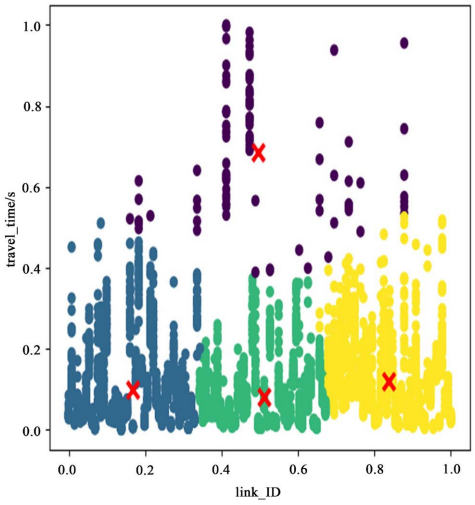
图 3、图 4、图 5 是选的 k = 4 时，在工作日、周末和节假日的 7~8 点、13~14 点、17~18 点的聚类结果图，分别用四种颜色代表一对一(黄色)，一对多(紫色)，多对一(蓝色)，多对多(绿色)路况下的道路空间特征分析结果。从图中可以明显看出时间不同路段的空间特征影响特征不同。



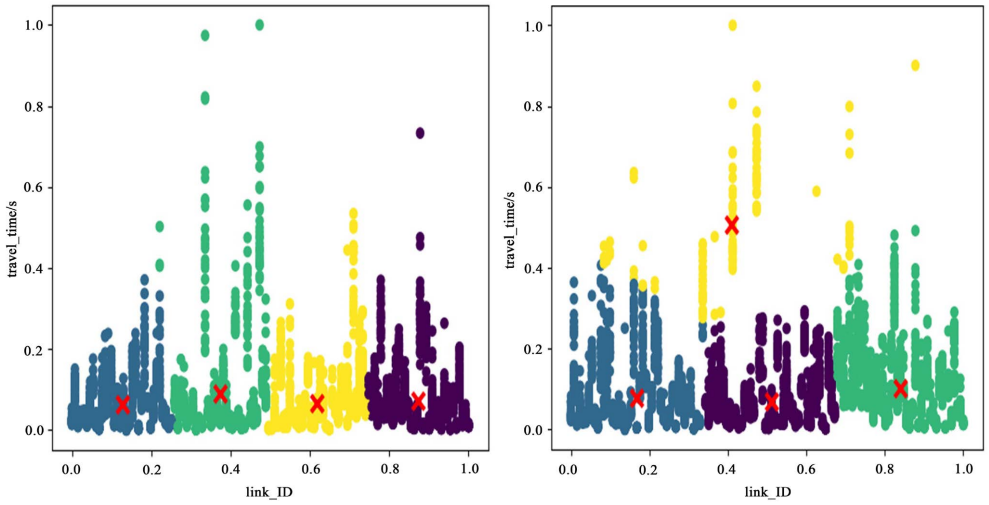
(a)

(b) (c)

Figure 3. Clustering results at 7~8 o‘clock (left), 13~14 o‘clock (middle), 17~18 o’clock (right) results on Workday
图 3. 工作日 7~8 点(左)、13~14 点(中)、17~18 点(右)结果图

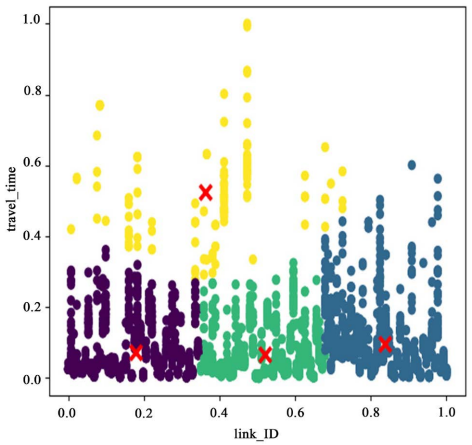


(a)

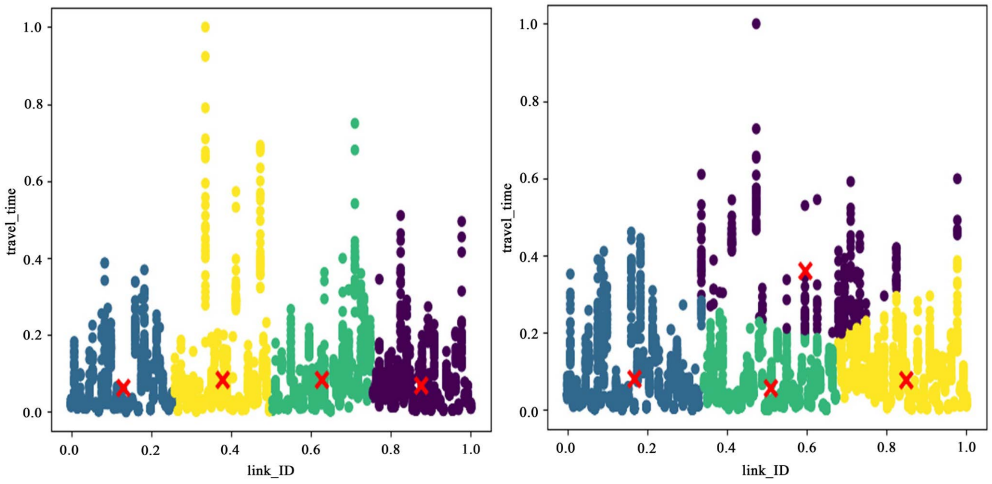


(b) (c)

Figure 4. Clustering results at 7~8 o'clock (left), 13~14 o'clock (middle) and 17~18 o'clock (right) on weekends
图 4. 周末 7~8 点(左)、13~14 点(中)、17~18 点(右)聚类结果图



(a)



(b) (c)

Figure 5. Clustering results of 7~8 o'clock (left), 13~14 o'clock (middle) and 17~18 o'clock (right) holidays
图 5. 节假日 7~8 点(左)、13~14 点(中)、17~18 点(右)聚类结果图

5. 总结

本文针对预测过程中数据空间特征利用不充分的问题，采用 K-means 进行空间特征的分析，提出了结合少样本、类型可调的四种典型空间道路融合路段的 ID、长度和宽度的基于 K-means 的空间特征进行分析的方法。展现 K-means 算法在交通流量等空间预测方面中的可使用性与可行性。实验表明，本文提出的处理和分析方法能有效处理空间预测方面问题中多因素的影响及内在相互影响问题，具有一定的实际意义。

参考文献

[1] 刘静, 关伟. 交通流预测方法综述[J]. 公路交通科技, 2004, 21(3): 82-85.

[2] Guo, S., Lin, Y., Li, S., Chen, Z. and Wan, H. (2019) Deep Spatial-Temporal 3D Convolutional Neural Networks for Traffic Data Forecasting. *IEEE Transactions on Intelligent Transportation Systems*, **20**, 3913-3926. https://doi.org/10.1109/TITS.2019.2906365

[3] 陆文琦, 谷远利, 陈伦. 基于时空融合的城市快速路短时交通流预测[J]. 计算机仿真, 2018, 35(9): 136-140+206.

[4] 孙斌. 基于全局时空图注意力网络的交通流量预测[D]: [硕士学位论文]. 北京: 中国矿业大学, 2021.

[5] 彭海驹, 严科文, 林松, 赖浩源, 张泽鑫. 融合 kmeans 聚类与 Hausdorff 距离的点云精简算法改进[J]. 地理空间信息, 2022, 20(8): 59-63.

[6] 王伟. Kmeans 聚类与多波谱阈值相结合的烟检测算法研究[J]. 工业加热, 2022, 51(4): 45-47+57.

[7] 史新颖, 夏元平, 毛曦, 殷红梅. DBSCAN 与 Kmeans 相结合的手机大数据聚类方法研究[J]. 北京测绘, 2019, 33(2): 132-137. https://doi.org/10.19580/j.cnki.1007-3000.2019.02.002

[8] 程万里. 超密集网络中基于聚类的资源高效分配技术研究[D]: [硕士学位论文]. 南京: 南京邮电大学, 2020.

[9] Lin, Y., Dai, X., Li, L., *et al*. (2019) Pattern Sensitive Prediction of Traffic Flow Based on Generative Adversarial Framework. *IEEE Transactions on Intelligent Transportation Systems*, **20**, 2395-2400. https://doi.org/10.1109/TITS.2018.2857224

[10] Chen, Y., Li, K., Yeo, C.K. and Li, K. (2023) Traffic Forecasting with Graph Spatial-Temporal Position Recurrent Network. *Neural Networks*, **162**, 340-349. https://doi.org/10.1016/j.neunet.2023.03.009

[11] 廖一迁, 岳显昌, 吴雄斌, 张兰. 基于 AIS 和 Canopy + Kmeans 算法的高频雷达阵列幅相校准[J/OL]. 现代雷达. https://kns.cnki.net/kcms/detail/32.1353.TN.20220411.1525.002.html