基于数据挖掘的成都市交通轨迹模式识别与预测研究

李佩儒[1]

[1]（北京航空航天大学 计算机学院）

Traffic Trajectory Pattern Recognition and Prediction in Chengdu Based on Data Mining Techniques

Peiru Li[1]

[1]（*School of Computer Science and Engineering*, *Beihang University*, *Beijing* 100000）

Abstract This research utilizes traffic trajectory data from Chengdu, applying clustering, classification, and regression techniques in data mining to uncover the potential travel patterns in Chengdu's traffic. The findings aim to provide scientific support for traffic management and planning. Specifically, the clustering analysis reveals the heterogeneity of traffic flow and traffic patterns, classification models help identify and predict different types of traffic behaviors, while regression analysis offers a quantifiable model for travel time prediction. The main contributions of this research include the following: **Traffic Network Preprocessing and Model Construction**: The Chengdu traffic network is preprocessed by integrating node and road information, followed by the use of Hidden Markov Models (HMM) for network matching to construct the Chengdu traffic network model.**Voting+ Model**: Based on an ensemble learning approach (Voting), combined with feature engineering, we develop the Voting+ model to identify features of the Chengdu traffic network and perform road segment classification. **ETA Prediction Using Transformer with Self-Attention Mechanism**: Leveraging existing GPS data and incorporating discrete factors such as weather and holidays, we design a Transformer-based model with self-attention for estimating the travel time (ETA), analyzing the variation patterns in travel time predictions.**Next-hop Prediction Modeling as Classification**: The next-hop prediction task is modeled as a classification problem, transforming the network's next-hop prediction into a vector-based endpoint prediction problem using a dual-attention mechanism, improving the accuracy of trajectory predictions.

Keywords: Traffic Trajectory Data, Data Mining, Clustering Analysis, Predictive Modeling, Traffic Behavior Classification

摘要 本研究基于成都市交通轨迹数据，应用数据挖掘中的聚类、分类和回归等技术，旨在揭示成都市交通出行的潜在规律，从而为交通管理与规划提供科学支持。通过这些技术手段，本研究能够全面分析交通数据，揭示出影响交通流动的关键因素，并为优化交通规划、提高管理效率提供理论依据和实践指导。本研究的主要贡献如下：交通路网预处理与模型构建：对成都市交通路网进行预处理，整合路网节点与道路信息，并利用隐马尔可夫模型（HMM）进行路网匹配，构建了成都市交通路网模型。Voting+模型：基于集成学习方法（Voting），结合特征工程，构建了 Voting+模型，实现了成都市交通路网特征的识别与路段分类任务。基于自注意力机制的 ETA 预测：利用现有 GPS 数据，并结合天气、节假日等离散性因素，设计了基于自注意力机制的 Transformer 模型，建立了交通到达时间（ETA）估计模型，分析并预测了到达时间的变化规律。将下一跳预测任务建模为分类问题，通过双头注意力机制，将路网的下一跳预测转化为向量终点预测问题，从而优化了交通轨迹的预测精度。

关键词 交通轨迹数据、数据挖掘、聚类分析、预测建模、交通行为分类

随着城市化进程的不断加快，城市交通问题日益成为影响城市居民日常生活质量和城市可持续发展的关键因素。交通流量的波动直接影响到城市的通行效率、交通安全以及居民的出行体验。因此，对交通轨迹数据进行深度分析和挖掘，揭示交通出行规律，不仅对优化交通规划、提高交通管理效率具有重要意义，而且为缓解城市交通拥堵、减少能源消耗、提高出行舒适性提供了理论依据和实践指导。

在数据科学领域，数据挖掘作为一种重要的技术手段，通过对大量复杂数据的分析，可以有效地发现潜在的规律和模式。本研究基于成都市的交通轨迹数据，应用数据挖掘中的聚类、分类和回归等技术，旨

在挖掘成都交通出行的潜在规律，为交通管理与规划提供科学支持。具体而言，通过聚类分析能够揭示出交通流量和交通模式的异质性，分类模型则有助于识别和预测不同类型的交通行为，而回归分析则为交通出行的预测提供了量化的模型。主要贡献包括以下几点：

（1）对成都交通路网进行预处理，整合路网节点和道路信息，并利用隐马尔可夫模型（HMM）进行路网匹配，构建成都交通路网模型。

（2）基于集成学习方法（Voting），结合特征工程，构建了 Voting+模型，实现了识别成都交通路网的特征，完成路段分类任务。

（3）利用现有 GPS 数据，结合天气、节假日等离散性因素，设计基于自注意力机制的 Transformer 模型，建立交通到达时间估计（ETA）模型，分析并预测到达时间的变化规律。

（4）将下一跳预测任务建模为分类问题，通过双头注意力机制，将路网的下一跳预测转化为向量终点预测问题，优化交通轨迹的预测精度。

本研究通过多种数据挖掘技术对成都交通轨迹数据进行深入分析，不仅为城市交通管理提供了理论支持，也为后续的智能交通系统设计和优化提供了实践依据。

1 相关工作

交通轨迹数据挖掘是智能交通系统（ITS）中的一个重要研究方向，其目标是通过分析从车辆、行人、交通设备等收集到的动态位置信息，提取出有价值的交通规律和模式。随着大数据技术、物联网（IoT）和移动设备的广泛应用，交通轨迹数据的收集、存储与处理能力得到了显著提升，为交通轨迹数据挖掘提供了丰富的基础。

在交通数据预处理上，交通轨迹数据通常具有大量的噪声和不完整性，因此，数据预处理的目标是提高数据的质量，去除冗余信息，补充缺失值并进行轨迹的平滑和校正。Xu[1]于 2016 年提出了一种基于空间和时间上下文的轨迹数据清洗方法，能够有效去除异常值并填补轨迹数据中的空缺。许多研究还采用插值技术、聚类分析和卡尔曼滤波方法来平滑轨迹数据，减少误差和波动。

在路网匹配上，不同算法有各自的优缺点，选择合适的算法需要考虑应用场景、数据特点（如采样频率、噪声水平）以及计算资源。HMM 和增量算法适合处理低频率和噪声较大的数据，而全局算法则适用于精度要求较高的离线任务。

聚类分析是交通轨迹数据挖掘中常见的技术，用于识别不同类型的交通流模式。Liu[1]应用基于 DBSCAN（密度聚类）算法对轨迹数据进行聚类分析，发现了不同时间段和不同区域的交通模式。聚类分析能够揭示出交通流的结构性特点，如高峰时段、交通拥堵区域等，进而为城市交通管理和规划提供决策支持。Wu[3]则基于轨迹相似度度量，提出了一种基于 K-means 算法的交通模式识别方法，通过将轨迹聚合成多个簇，揭示出交通流的异质性。

交通行为预测是交通轨迹数据挖掘中的核心任务之一。分类分析方法被广泛应用于交通行为预测，如对车辆行驶状态（正常、加速、减速、停车等）进行分类，或者预测交通流量、道路通行能力等。Zhang[4]提出了一种基于决策树和随机森林的交通行为分类模型，用于对车辆在不同道路状况下的行驶行为进行预测。分类模型可以帮助交通管理部门实现对交通流的实时监控与预测，从而提高交通效率。

另外，近年来，深度学习方法在交通行为预测中表现出色。Dai[5]应用卷积神经网络（CNN）和循环神经网络（RNN）相结合的模型，预测交通轨迹中的下一步位置和行驶速度，为交通流量预测和路径规划提供精确模型。

到达时间预测（ETA）是智能交通系统中的一项关键任务，尤其在导航系统和共享出行服务中具有广泛的应用。回归分析技术在 ETA 预测中得到了广泛应用。研究者通过建立回归模型，利用历史交通数据、天气信息、节假日等多种因素，预测车辆或行人的到达时间。Chien[6]提出了一种基于支持向量回归（SVR）的方法，通过集成多维特征（如交通流量、交通事件、天气等）来实现准确的 ETA 预测。

近年来，基于深度学习的模型，如 LSTM（长短时记忆网络）和 Transformer，也被应用于 ETA 预测中。Chen[7]提出了一种基于 Transformer 的到达时间预测方法，能够捕捉复杂的时间序列依赖关系，并通过自注意力机制优化预测效果。

轨迹下一跳预测（Next Location Prediction）旨在基于已有轨迹数据，预测目标实体（如车辆、行人等）的下一位置。该任务在交通预测、路径规划和移动行为分析中具有广泛应用。许多研究采用深度学习方法来解决这一问题，如基于 RNN、LSTM、GRU 等神经网络模型，能够有效捕捉轨迹的时序特征并进行下一

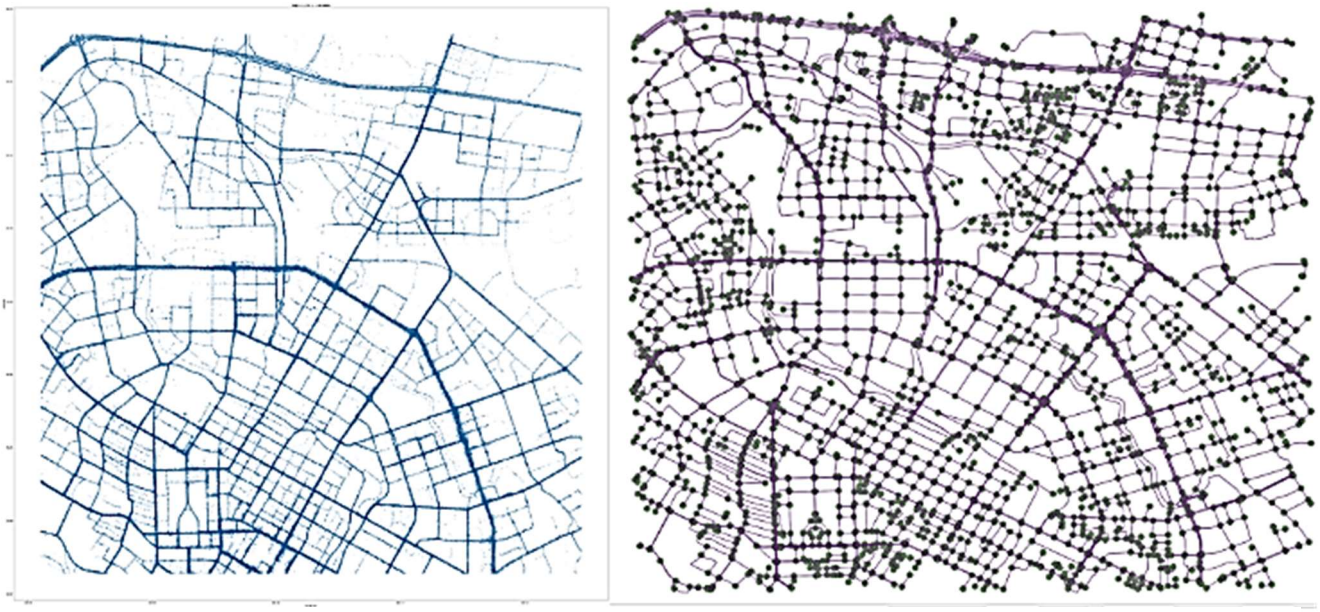

Fig. 1 Road Webs before and after Preprocess

图 1：预处理前后路网图

位置预测。Yuan[8]提出了一种基于 LSTM 的轨迹下一跳预测方法，通过深度学习捕捉轨迹的时空模式，成功预测了不同交通方式下的下一跳位置。

另外，双头注意力机制也被用于该任务，通过加强时间和空间特征的学习能力，提高了预测的准确度。Zhou[9]提出了基于双头注意力机制的下一跳预测模型，能够同时捕捉到历史轨迹中的空间和时间信息，从而精确预测目标实体的下一位置。

随着多种传感器和数据源的出现，交通轨迹数据常常与其他类型的数据（如天气、社会事件、交通设施等）融合使用。多源数据融合能够提供更加全面的信息，优化交通流预测。Li[10]提出了一种基于图神经网络（GNN）和多源数据融合的交通流预测模型，能够有效地结合轨迹数据、道路属性以及天气等因素，进一步提升交通流的预测精度。

2 数据预处理及路网匹配

2. 1 数据预处理

在轨迹数据中，GPS 信号本身会受到诸多因素的影响，如建筑物遮挡、天气状况、设备精度等，导致位置数据具有一定的噪声。这些噪声可能表现为位置点的异常波动、跳跃等，这对于后续的路网匹配会带来较大影响。因此，在进行轨迹映射到路网之前，必须先进行噪声处理，以确保数据的准确性和稳定性。在轨迹数据中。

考虑到 GPS 信号本身的偏移，首先对其 GPS 经纬度进行修正：将数据转换为 WGS84 坐标系后，使用移动平均法对经纬度坐标进行平滑处理，以及使用距离阈值对异常点进行检测和去除。对于每个坐标点 (lng_i,lat_i) 我们使用一个窗口大小为 w 的滑动窗口来计算窗口内所有点的平均经纬度。对于第 i 个点，平滑后的坐标(lng_i',lat_i')。计算公式如下：

$$lng_i' = \frac{1}{w}\sum_{j=i-\lfloor w/2 \rfloor}^{i+\lfloor w/2 \rfloor} lng_j$$

$$lat_i' = \frac{1}{w}\sum_{j=i-\lfloor w/2 \rfloor}^{i+\lfloor w/2 \rfloor} lat_j$$

由于 GPS 信号的噪声，可能会出现两个相邻点之间的距离明显大于正常的行驶距离，所以需要检测异常值。这里我们通过 Haversine 公式计算两点间的地理距离，设置阈值检测：

$$a = \sin^2\left(\frac{\Delta\ \mathrm{lat}}{2}\right) + \cos\left(\ \mathrm{lat}_{\ 1}\right)\cdot\cos\left(\mathrm{lat}_2\right)\cdot\sin^2\left(\frac{\Delta \mathrm{lng}}{2}\right)$$

$$c = 2\cdot \mathrm{atan}\ 2(\sqrt{a},\sqrt{1-a})$$

$$d = R\cdot c$$

对于检测到的异常点，可以选择使用插值法修复，我们这里采用前后点平均值法修正。我们最后预处理结果数据构成的路网的 QGIS 可视化结果如图 1。

2. 2 路网匹配

在路网匹配中，HMM 用于根据带噪声的 GPS 数据推断出最可能的行驶路线。通过隐马尔科夫链模型，能够将车辆的实际行驶状态与观测数据（如 GPS 坐标）进行对比，从而推测出最符合路网的轨迹路径。Viterbi 算法是用于隐马尔科夫链中最优状态序列推断的动态规划算法。其目标是找到给定观测序列的最大化路径概率，通过马尔科夫链以及贝叶斯公式得到

的优化目标如下：

$$P(O_1, O_2, \dots, O_T, S_1, S_2, \dots, S_T)$$

$$= P(O_1 \mid S_1) \cdot \pi_1 \cdot \prod_{t=2}^{T} A_{S_{t-1}, S_t} \cdot B_{S_t}(O_t)$$

公式中 A 为状态转移概率，B 为状态 S 下 O 的发射概率。通过递归回溯计算最大概率：

$$\delta_t(i) = \max_{S_1, S_2, \dots, S_{t-1}} (\delta_{t-1}(i') \cdot A_{i',i} \cdot B_i(O_t))$$

使用开源架构 Trackit[1]可以进行基于 HMM 的路网匹配，其中一个路网匹配结果如下：

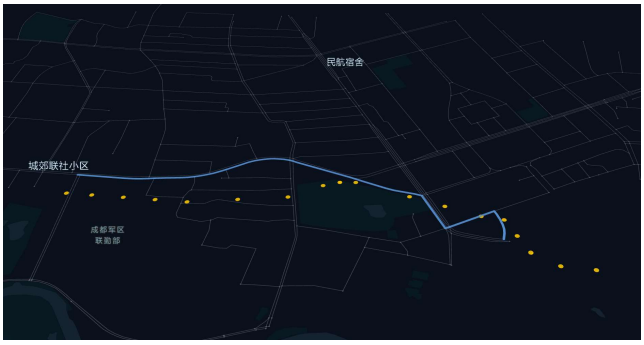

Figure 2: Road Network Matching Results

图 2：路网匹配结果

最终，我们得到的路网匹配主要数据存储方式如表 1 所示：

Table 1: Road Network Matching Table

表 1：路网匹配表格

字段名称	字段含义	类型
agent_id	GPS 点所属 agent_id	string
seq	GPS 点的序列 ID	int
sub_seq	GPS 点的子序列 ID，如果子序列>0，说明该点是在匹配后补出来的点，称之为后补点	int
link_id	对应路网字段	int
from_node	起始节点	int
to_node	终到节点	int
(lng,lat)	EPSG4326 坐标	/
(prj_lng, prj_lat)	GPS 点在匹配路段上对应匹配点的坐标	/

3 路段分类

3.1 特征工程

使用 2.1 中的哈弗辛距离，我们可以得到两个坐标点间地理距离。

然后，基于每三个连续点的圆形拟合来评估路径的弯曲程度。这个方法假设路段的弯曲是由三个连续的点定义的，从地理坐标系映射到直角坐标系后，可以得到以三点为顶点的三角形。求其外接圆半径，即可得到该段路径的近似曲率半径，取倒数即可得到曲率，计算过程如下：

$$d = 2 \cdot R \cdot \arcsin\left(\sqrt{\sin^2\left(\frac{\Delta\phi}{2}\right) + \cos(\phi_1) \cdot \cos(\phi_2) \cdot \sin^2\left(\frac{\Delta\lambda}{2}\right)}\right)$$

其中：

$$\Delta\phi = \text{lat}_2 - \text{lat}_1$$
$$\Delta\lambda = \text{lon}_2 - \text{lon}_1$$

在得到哈弗辛距离后，曲率 k 计算如下：

$$1/k = \frac{d_1 \cdot d_2 \cdot d_3}{4 \cdot \sqrt{s \cdot (s - d_1) \cdot (s - d_2) \cdot (s - d_3)}}$$

同时，计算这一路径的曲率方向，即使用外积：

$$\text{cross\_product} = (lat_2 - lat_1) \cdot (lon_3 - lon_2) - (lon_2 - lon_1) \cdot (lat_3 - lat_2)$$

计算出三点的曲率后，我们可以得到平均曲率和曲率和（可以直接判断环岛、主干线等）。

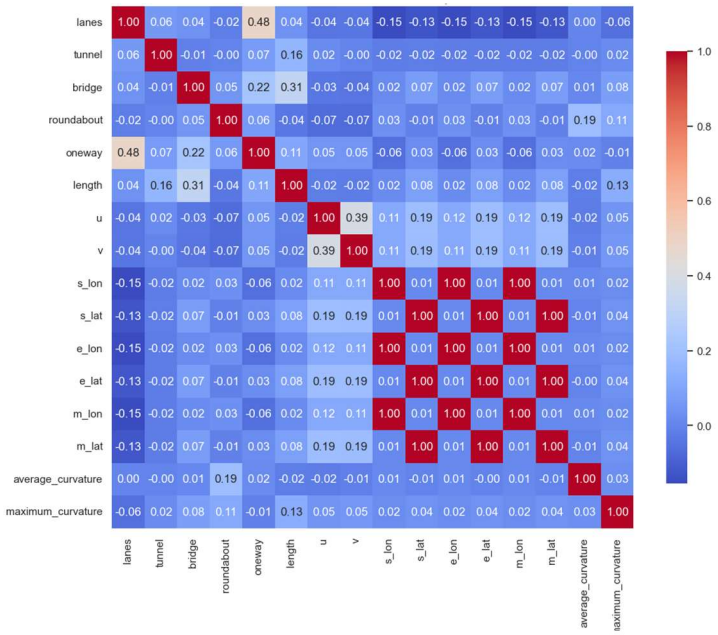

Figure 3: Correlation Heatmap

图 3：相关性分析热力图(红色表示正相关，蓝色表示负相关，值越接近 1，相关性越强)

在提取隐藏特征后，考虑到变量相关性存在，进行相关性分析，相关性热力图如图 3 所示。

注意到以下特点：

(1) lanes 和 oneway 有一定的正相关性（约 0.48）。这可能表明单向道路的车道数可能存在某种规律，例如单向道路车道数可能更

稳定（比如日常生活中常见为单车道）。

(2) 起点和终点的经纬度相关性较低，但同一类经纬度之间存在强相关性（例如 s_lon 与 e_lon、s_lat 与 e_lat）可能说明轨迹跨度较大，路径间起点和终点相对独立。

(3) length 和 bridge（约 0.31）length 与 bridge 的正相关性表明，桥梁可能更倾向于出现在较长的路段上。

(4) 环岛与平均曲率有一定的正相关性，说明在环岛区域中，道路的曲率较大（即更弯曲）。

(5) 车道数与最大曲率之间有弱负相关性，这可能说明车道数较多时，路段通常更加直线化，最大曲率较小

最后，我们考虑到量级的影响，对连续型数据进行归一化，同时对离散型变量进行独热编码。选择若干特征作为模型输入。

除此之外，考虑到道路的连接情况，如高速公路几乎不会与城市街道相连，我们构建邻接情况图（图 5）。

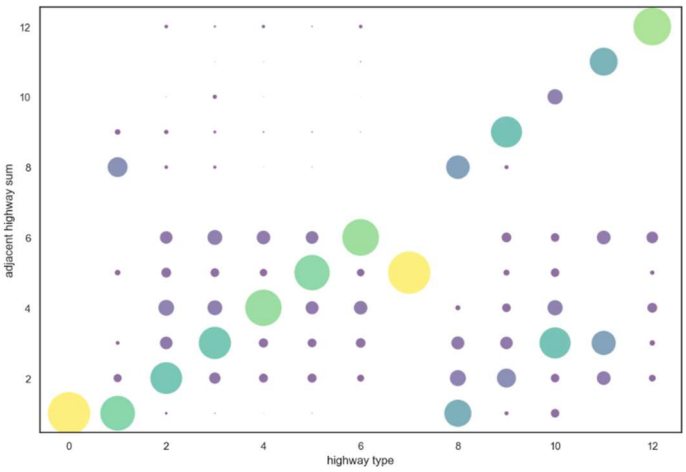

Figure 4: Adjacent Situation

图 4：道路连接情况

可以看出不同道路连接情况差异很大。因此，我们将某一道路的连接道路集合作为模型输入之一，作为判断道路类型的特征。

3.2 Voting+ 模型构建

Voting 是一种集成学习（Ensemble Learning）方法，通过结合多个模型的预测结果来提高分类或回归的整体性能。它是集成学习中的一种简单而常用的技术，特别适合在模型预测结果不一致时，通过集体决策获得更稳定、更准确的结果。

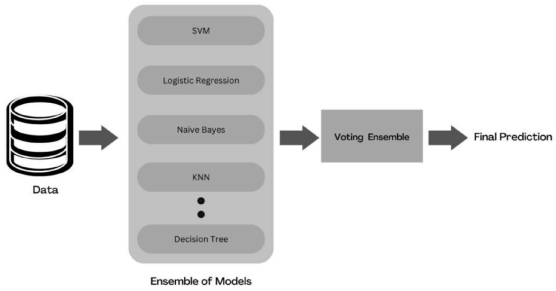

Figure 5: Voting Model

图 5：投票集成模型

在轨迹分类任务中，由于数据特征较多、分类目标较多，我们选择以树为基础的模型，其好处在于对数据特征的鲁棒性、特征重要性评估、高效的多目标分类、非线性决策能力强。最终我们选择包括 RandomForest[12]、XGBoost[13]、LGBM[14]、ExtraTrees[15] 在内的四种基学习器，以软投票的方式进行集成。

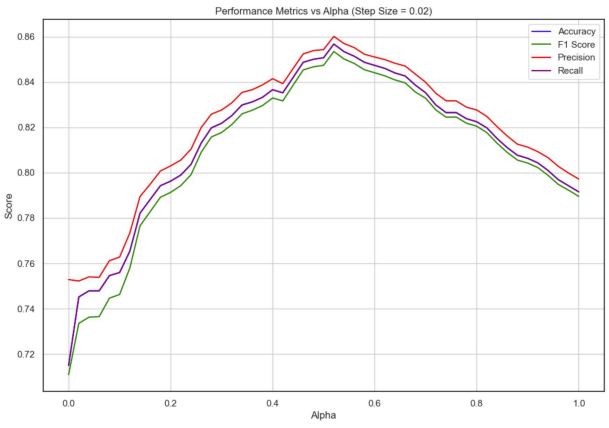

Figure 6: ACC, F1, Precision and Recall Curves

图 6：ACC、F1、准确率、召回率曲线

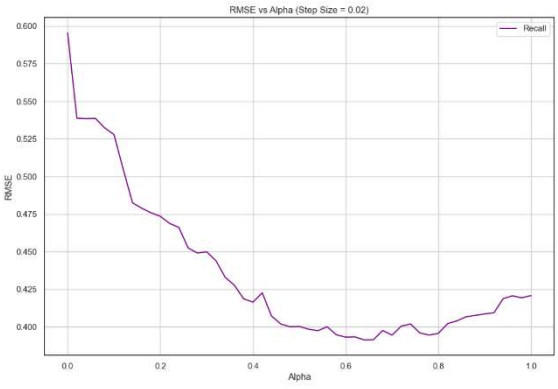

Figure 7: RMSE Curves

图 7: RMSE 曲线

训练好 Voting 模型后，把临界情况预测结果与 Voting 模型的预测结果进行加权调整，得到 Voting+ 模型。其加权比例以及 F1 得分、准确率、RMSE 等

指标的变化情况如图 6、图 7 所示。

最终确定按照 3:2 加权比例，得到的结果在测试集上效果为：

Table 2: Voting+ Results

表 2：Voting+模型结果

指标名称	数值
Accuracy	0.85675
F1 Score	0.85346
Precision	0.86012
Recall	0.85675
RMSE	0.39131

4 ETA 回归分析

在 ETA 回归分析中，需要考虑变量较多，包括静态数据，如道路网络信息、历史交通数据，以及动态数据，如实时交通信息、天气数据、 时间特征：节假日、工作日。由于部分数据难以在互联网上查找，所以我们仅考虑了天气、节假日信息（十一期间）、早晚高峰期以及道路特征，作为研究的出发点。

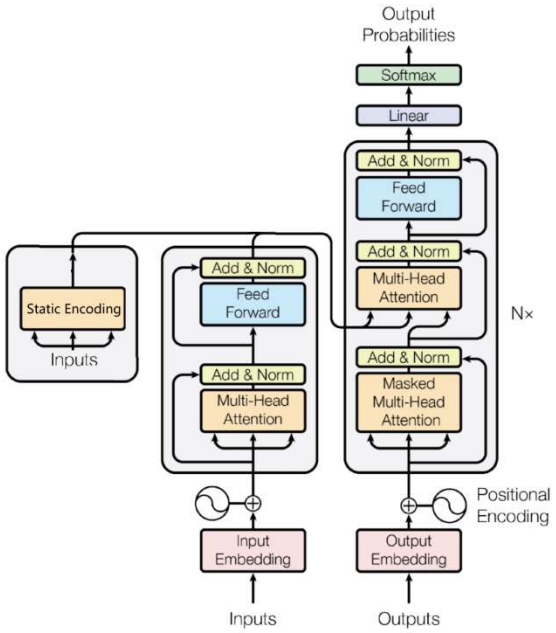

Figure 8: Transformer + Static Encoder Attention

图 8: 静态变量机制

4.2 日静态变量

考虑到天气、节假日等作为一段轨迹内几乎不变的因素，这些均可以视为轨迹内的静态变量。首先通过爬虫程序爬取天气、节假日信息后，对数据集重新整合。借鉴 NLP 领域嵌入层的概念，将这些变量映射为高维度向量（即为每个离散变量单独生成一个 **Query** 向量），然后让 Query 和 Key 进行点积，计算离散变量在每个输入上的重要性，通过 Softmax 权重对 Value 进行加权，得到最终的输出。公式如下：

$$\text{Attention}\ (Q_{\text{cat}}, K, V) = \text{Softmax}\ (\frac{Q_{\text{cat}} K^T}{\sqrt{d_k}})V$$

如图 8 所示，我们在 Tranformer 架构上进行调整，增加对静态变量的注意力机制。

4.2 轨迹序列控制

作为 ETA 合法轨迹，其最重要的是时间单调递增，同时，相邻时间间隔在非拥堵时段不应过长，因此，将解码器的输出转换为**增量 logits**，即不再预测时间偏移量（以最早日期为基准），转向回归预测时间的增加量以及积累值，即如下公式：

$$\text{x}_t = \log\left(1 + e^{\text{Logits}_t}\right) + \epsilon$$

$$\text{cumsum}_t = \sum_{i=1}^{t} \text{x}_i$$

除此之外，我们需要控制增量变化情况，即不能让时间间隔增量过快。所以对 MSE loss 函数进行修正，添加正则项，对不合理的时间间隔进行惩罚，其具体公式为：

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{base}} + \lambda \sum_{t=1}^{T} max(0, x_t - \epsilon)$$

4.3 模型运行结果

实验硬件设施为 1 张 RTX3060Ti 显卡，对于数据，在按照轨迹序号进行分割后，使用零填充对齐。之后按照是否为节假日进行分割，通过分层抽样将训练集和测试集按照 8：2 进行分割，设置 batch 大小为 64，运行 200 个 epoch 内，loss 变化曲线与测试集上 RMSE、MAE 变化曲线如图 9 所示。

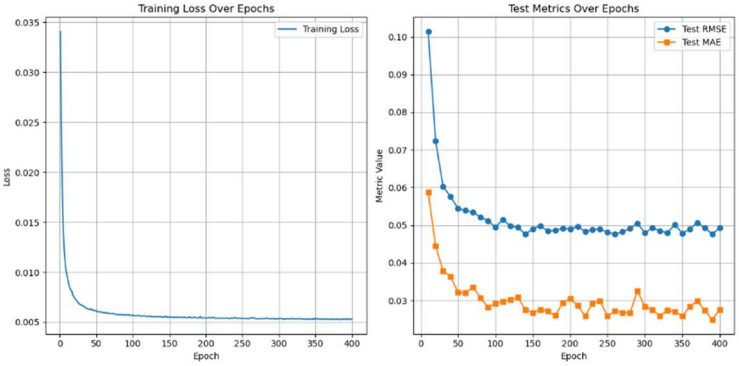

Figure 9: Loss Curve and Test Metric Curve

图 9: Loss 曲线与测试集性能

5 下一跳预测

5.1 问题建模

为了将下一跳预测任务建模为分类问题，我们需要首先对所有可能的轨迹点进行离散化。假设我们有一个空间网格，将整个空间划分为若干个小区域，每个区域可以被认为是一个类别。然后，对于每一个历史轨迹序列，我们可以用分类模型来预测下一个轨迹点落在哪个网格区域。

使用滑动窗口法，将历史轨迹数据转化为时间序列数据。每个时间步的特征作为输入序列，根据这些历史信息预测下一个轨迹点所在的网格区域，从而输出下一个轨迹点所在的区域类别。通过将空间划分为多个网格，模型预测的输出就是一个类别标签，表示物体的下一个轨迹点属于哪个网格区域。

考虑到这种时序化关系，我们构建了 BiLSTM 模型，并加入 4.1 中所使用的注意力机制，即关注到不同日期、情境下人们驾车目的地的不同。从而实现处理时间序列数据，同时捕捉过去和未来的时序信息，以增强预测的准确性。

其存在的难点在于：

(1)网格大小选择：空间离散化的第一步是将整个空间划分为多个网格区域。网格的大小会直接影响预测的精度。较大的网格会导致信息丢失，使得空间位置预测较为模糊，而较小的网格则会带来类别数量激增，增加计算复杂度并可能导致数据稀疏性。

(2)空间不均匀分布问题：不同区域的轨迹数据可能分布不均匀，有些区域可能非常密集，而有些区域的数据较少。为了避免模型对稀疏区域的预测不准确，可能需要考虑在训练过程中使用加权损失函数，赋予数据较少的区域更高的权重，或者通过数据增强 技术生成稀疏区域的额外样本。

5.2 实验结果

考虑到离散型特征可能对训练无意义或者产生负面作用，对不同类型的特征进行组合（天气、节假日等），我们进行消融实验，并与 Bert 模型进行对比。最终得到只有降水量、节假日对下一跳预测结果有较好影响，对比 RMSE 结果如表 3。

Table 3: Ablation Results

表 3：消融实验结果

Model	Rain	Holiday	Temperature	RMSE
LSTM1	√		√”	0.07091
LSTM2		√”		0.06314
LSTM3	√	√		0.05908
LSTM4				0.09712
Bert1	√”		√”	0.10761
Bert2		√”		0.09128
Bert3	√	√		0.08437
Bert4				0.12095

参 考 文 献

[1] Xu, Y., Li, Q., & Wei, H. (2016). A data cleaning method for vehicle trajectory data in urban transportation systems. Transportation Research Part C: Emerging Technologies, 68, 33-47.

[2] Liu, J., Zhang, H., & Li, Y. (2017). A clustering method for discovering *traffic flow patterns using GPS trajectory data. International Journal of* Geographical Information Science, 31(7), 1414-1432.

[3] Wu, X., Zhou, Z., & Zhang, X. (2018). Pattern discovery in vehicle trajectory data using K-means clustering. IEEE Transactions on Intelligent Transportation Systems, 19(8), 2574-2583.

[4] Zhang, W., Chen, T., & Li, J. (2019). Predicting vehicle behavior based on GPS trajectory data using decision trees and random forests. Journal of Transportation Engineering, Part A: Systems, 145(4), 04019022.

[5] Dai, L., Song, H., & Zhang, H. (2020). Traffic behavior prediction using deep learning models: A survey. Transportation Research Part C: Emerging Technologies, 113, 17-32.

[6] Chien, S., Ding, Y., & Wei, C. (2014). A support vector regression model for short-term traffic flow prediction with weather information. Transportation Research Part C: Emerging Technologies, 45, 71-83.

[7] Chen, L., He, Z., & Li, X. (2021). ETA prediction based on Transformer for urban traffic networks. Transportation Research Part B: Methodological, 141, 163-177.

[8] Yuan, Z., Zheng, Y., & Xie, X. (2018). T-Drive: Driving directions based on taxi trajectories. ACM Transactions on Intelligent Systems and Technology, 9(3), 1-26.

[9] Zhou, G., Wei, F., & Yu, J. (2020). A double-attention mechanism for next location prediction from trajectory data. ACM Transactions on Knowledge Discovery from Data, 14(5), 1-22.

[10] Li, J., Li, Z., & Zhang, Y. (2022). Multi-source data fusion for traffic flow prediction based on graph neural networks. IEEE Transactions on Intelligent Transportation Systems, 23(7), 5300-5312.

[11] https://github.com/zdsjjtTLG/TrackIt

[12] Breiman, L. (2001). "Random forests." Machine Learning, 45(1), 5–32.

[13] Chen, T., & Guestrin, C. (2016). "XGBoost: A scalable tree boosting system." Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.

[14] Ke, G., Meng, Q., Finley, T., et al. (2017). "LightGBM: A highly efficient gradient boosting decision tree." Advances in Neural Information Processing Systems, 30.

[15] Geurts, P., Ernst, D., & Wehenkel, L. (2006). "Extremely randomized trees." Machine Learning, 63(1), 3–42.