1. 绪论

1.1 选题意义

交通拥堵是城市交通中非常常见的问题。城市化的推进和民众生活水平的提高同时也来了城市车辆密度的持续上升，交通拥堵也因此愈发常见，影响民众的出行质量。在此背景下，普通居民和城市管理者对于交通拥堵预测能力的需求与日俱增。

1.2 国内外交通拥堵识别和预测模型发展现状

对于交通拥堵的判断，不同国家及地区都有各自的官方标准。当前我国所采用的标准为。一般而言，城市主、次干道的拥堵等级是由道路车流速度决定的。

近年来，短期交通拥堵预测算法在国内外蓬勃发展，目前可分为三类：概率推理模型、浅机器学习模型、深度学习模型。常见的用于拥堵预测的概率推测模型包括模糊逻辑、隐马尔科夫模型、高斯混合模型、贝叶斯网络等。这些模型的共同缺点是面对大规模的数据，它们的复杂度会显著提升。浅机器学习则包括人工神经网络、回归模型、决策树以及支持向量机。其中，支持向量机如今在拥堵预测算法领域仍受到广泛的应用**[1]**。人工神经网络的复杂度较概率推理模型略有减少，但是其输入必须是含标签的。因此，对于无标签的原始数据，数据预处理必须进行聚类分析。这使得该类模型面对大规模数据的处理时间显著提升，且缺少适应性。相较于浅神经网络，深度学习算法可以在对数据无任何事先了解的情况下提取出其中的特征。而对于隐变量错综复杂的交通预测问题，深度学习对数据特征的敏感性吸引越来越多的研究使用它作为预测算法的核心。

1.3 目的和研究范围

本研究尝试设计有效的短期交通拥堵预测算法。其中，“短期”是指预测一到两天以内的交通数据。考虑到交通信息在宏观上具有一定的周期性，收到Chen等的研究的启发，我们尝试借助卷积神经网络（Convolutional Nerural Network，下文简称CNN）来实现交通拥堵的大范围预测。在本研究中，我们探索了通过合理的交通数据预处理和CNN的结构设计，测试CNN对大范围交通数据的预测性能，并进行可视化展示。

2. 研究问题

2.1 问题定义及重要参数

2.1.1 拥堵状态定义

本研究中，我们以一个空间网格内的平均速度作为拥堵状态的表征。在不考虑特殊路段的情况下，拥堵指数的计算直接取决于网格平均速度。网格速度到拥堵等级的映射参考[2]。

2.1.2 研究问题定义

在本研究中，我们对短期大范围交通拥堵预测问题的定义如下：给定一个地图网格和一个网格速度的序列{}, 。其中，和分别表示数据所在网格的行索引与列索引，t表示数据的时间索引。我们的目标是预测所有网格在时间点T的网格速度, ，从而达到预测单位网格拥堵程度的效果。

2.2 数据预处理

2.2.1 数据质量综述

本研究所用的数据来自于滴滴盖亚计划的开源数据，内容是2016年11月成都市中心的所有网约车订单数据。数据内容包括网约车的司机ID，订单ID，Unix时间戳，网约车在GCJ-02坐标系下的经度、维度。数据本身呈现很好的结构性，其中包含了详细经纬度信息和时间序列信息，能够从中提取时空特征信息。

2.2.2 预处理步骤

滴滴公司提供的数据较为完整，无需进行复杂的数据清洗和补全。但为了获得前文定义问题中所要求的网格速度序列，我们对原数据进行了一系列的变换

原数据中网约车位置采用的坐标系是国内的GCJ-02火星坐标系。出于网格化需求，我们需要对应的UTM平面坐标系坐标。同时，为了实现可视化，我们也需要绝大多数地图使用的WGS-84坐标。因此，坐标数据需要分别进行两组坐标变换

随后，我们将这些时空数据点归类到棋盘状的栅格中。对于网约车轨迹数据而言，原本的轨迹也随之被栅格化为网格的序列。每个时刻的每个网格单元都对于三维矩阵的一个位置。这种预处理方式可以实现交通数据的张量化，从而可输入CNN进行处理。栅格化的颗粒度对试验有一定影响。只有颗粒度精细到一定程度才能充分保留网格内的关键信息，而颗粒度太大则会导致数据过量，减缓处理速度。经过试验后，我们选择了一块15\*15、单元边长为70米的网格作为研究对象，同时将时间颗粒度设置为100秒。

网格平均速度的计算分为两步：个体特征计算和个体特征集计。在计算个体特征时，我们需要计算每个网格中每个网约车的瞬时速度。瞬时速度的计算方法是将相邻时刻的欧式网格位移距离代替每个点的瞬时速度。而在集体特征集计部分，我们计算了每辆车在同一网格下的平均速度，然后将同一网格所有车的均速求均值，作为该网格的平均速度。

2.3 探索性数据分析

2.3.1 周期性分析

宏观上看，不同工作日之间的网格数据会表现出一定的周期性。以网格平均速度为例，以下选取2016年11月第一周(11.1-11.7)数据，由图[1]可见，每周每天相同时间的平均速度趋势相似，同时平均车速折线两端局部最低分别对应早晚高峰（7：00-9：00；16：00-18：00）。其中5号6号为周末，在早高峰时段相对车速较高。

图表, 直方图

描述已自动生成

**图1 2016年11月第一周网格平均速度趋势图**

以2016年11月的周二为例，从图[2]来看相同时间的平均速度趋势具有高度一致性，在早/晚高峰时段一致性尤为明显

图表

描述已自动生成

**图2 2016年11月周二网格平均速度趋势图**

2.3.2 邻域相关性分析

根据Chen的研究**[3]**，一个交通网格在某一时刻的网格速度与其在空间和时间上邻近的网格的差距在一定的距离范围内可视为是连贯的。这是因为大部分时间段的交通状况很少出现突变的情况。因此我们认为，同时考虑一定时空邻域内的所有数据对于预测某个时空网格的数据是有帮助的。

2.4 卷积神经网络的应用

2.4.1 CNN概述

卷积神经网络是一种被广泛用于交通数据分析和预测的深度学习算法。CNN在处理二维数据（比如图像）上具有优秀的性能。因此，CNN被广泛用于交通拥堵预测算法。CNN的主要结构包括：输入层、卷积层、池化层、全连接层和输出层。CNN在大体量的数据集下能表现出良好的性能，同时对分类和回归问题具有较高的处理性能。因此，任何可以转化为二维数据的数据集都可投入CNN进行处理，而交通路网数据可以视为由经度和维度编号的二维数据。然而，目前上没有通用的对CNN各个层的设计策略。对于每个特定问题，都需要对CNN的各类参数进行调试。

2.4.2 项目应用

在本项目中，整个交通网络在同一时刻的数据可以视为一个由横纵坐标编码的二维的速度矩阵。CNN在提取二维数据的特征上具有显著优势，而一个交通网络上的信息同样可由经度和维度两个维度编号，因此我们在预处理中选择将同一时间的交通信息转换成二维矩阵。同时，CNN的每个隐藏层都可以同时处理来自多个通道的二维矩阵，并提取它们所包含的共有特征。这一特性启发我们将间隔一定周期的数据矩阵叠加在一起，从多个通道一起输入神经网络，从而尝试让模型记忆并预测出交通拥堵的周期性特征，提高预测性能。

3. 研究方法和实验设计

3.1 卷积神经网络的设计

3.1.1 输入层

考虑到交通数据的邻域连贯性以及数据集的规格，我们采用数据裁剪来增强数据。本项目的研究区域是一个15×15的网格。我们采用长宽均为10的正方形滑动窗口对原空间网格进行裁剪，使数据量增加到原来的6×6=36倍。

目标问题的本质是一个回归问题。因此输入张量的结构可分为标签层和数据层。标签层为待遇测时刻m的速度网格。对于数据层，考虑到网格速度的周期性以及连贯性，数据层被设计为过去h小时中，每隔1小时的网格速度，加上过去d天中，同一时刻的网格速度。

3.1.2 隐藏层结构

隐藏层沿用由LeCun[3]设计的“卷积-池化”结构。特别地，考虑到卷积和池化操作可能会是特征图的尺寸缩小，因此我们选择在最后一个卷积层前增加一个上采样层来还原特征图的尺寸。

3.1.3 优化算法、数据集和目标函数

模型选择采用Adam优化器进行优化。我们选择前29天的数据作为训练集，其中20%作为验证集，用于早停策略的执行，避免模型严重过拟合。目标函数采用预测值与观测值之间的均方误差。

3.1.4 性能评价标准

本研究中，我们使用四个性能指标来评价模型性能：绝对平均误差（MAE），平均相对误差（MRE），均方根误差（RMSE）和决定系数（R2）。

3.2 预实验

为了尽量提高CNN性能，需要为模型找到合适的超参数。因此，在正式试验之前，我们选择用小数据集对超参数进行预实验，即用11月1到4日的数据预测11月5日的数据。预实验的测试变量为：输入张量的t和d，Adam优化器的学习率以及模型训练的批尺寸。

3.2.1 输入张量的内容

预实验分别取 与 作为输入张量的内容。模型性能如图[]所示

3.2.2 学习率

预实验测试了优化器学习率为0.0001到0.0005的情况，间隔为0.0001。模型性能如图[]所示。

3.2.3 批尺寸

预实验测试了训练批尺寸为32到128的情况，间隔为32。模型性能如图[]所示。

3.3 正式试验

3.2.1 实验目的

用CNN处理已预处理后的数据集，用前29天的数据预测第30天的全部网格速度数据。

3.2.2 实验设计

输入张量的t=2,d=2。Adam优化器的学习率为0.0002。训练批尺寸为64。隐藏层设计如图[]所示。前29天的速度数据在转换成特定张量以后输入模型进行训练。当模型训练结束后，我们用该模型预测第30天的数据。

3.2.3 实验结果

- 性能数据：表

- 可视化展示：图

4. 研究结论和讨论

4.1 研究结论

在本研究中，我们提出了一种用于预测短期交通拥堵状况的深度学习算法。整个算法基于卷积神经网络。考虑到交通数据往往具有连贯性和周期性，我们在输入层的选择上做出了一定创新，将隔日与隔小时的数据相结合同时作为输入。实验结果表明，模型对于一整日交通速度变化信息的预测表现出了较高的回顾性能。这说明该算法设计可以用于交通拥堵状况的预测，从而辅助交通管理决策。

4.2 讨论

4.2.1 算法的优点

该算法可以在短时间内迅速地捕捉并预测大范围空间路网内的速度信息，同时保持较高的预测精度。

如表[]所示，与传统的交通预测算法[4]做对比，本研究所训练的模型在预测性能方面取得了一定的提升。

|  |  |  |
| --- | --- | --- |
| 模型 | MAE | RMSE |
| k-NN | 1.388 | 3.065 |
| LSTM | 1.231 | 2.720 |
| SVM | 1.345 | 2.286 |
| **本研究模型** | **0.526** | **0.883** |

**表[] 与传统交通拥堵算法的性能对比**

4.2.2 算法的不足

- 基于网格化、忽略了不同路段的信息

- 完全基于历史信息，没有考虑天气、节假日等实时信息

- 模型的性能还不足以达到可应用的范围

- 训练数据完全基于网约车，无法提取出更多的路段车流信息，影响算法预测

- 对数据突变的预测能力较差

- 深度学习的性能优势离不开海量的数据。本研究所用数据相对较少。实现广泛应用需要更大体量的数据

4.2.3 项目的应用前景

5. 谢辞

参考文献

[4] Bogaerts, Toon, et al. "A graph CNN-LSTM neural network for short and long-term traffic forecasting based on trajectory data." *Transportation Research Part C: Emerging Technologies* 112 (2020): 62-77.