可上网查阅： www.sciencedirect.com

ELSEVIER

科学导向

IFAC Papers Online
CONFERENCE PAPER ARCHIVE

IFAC PapersOnLine 53-5 (2020) 512-517

利用决策树、逻辑回归和神经网络进行交通拥堵预测

Tariku Sinshaw Tamir [1,2] , Gang Xiong [1,3] , Senior Member, IEEE, Zhishuai Li [1,2] , Hao Tao[4] , Zhen Shen[5] , Bin Hu [1,5] （通讯作者）, Heruye Mulugeta Menkir [6]

[1] 中国科学院自动化研究所复杂系统管理与控制国家重点实验室，中国北京 100190

[2] 中国科学院大学人工智能学院，中国北京 100049

[3] 中国科学院云计算中心广东省三维打印与智能制造工程技术研究中心，中国东莞 523808

[4] 中国船舶研发设计中心，中国武汉 430064

[5] 中国科学院自动化研究所智能系统与技术北京市工程研究中心，中国北京 100190[6] 德布雷马科斯大学，技术研究所，埃塞俄比亚德布雷马科斯 269 号

tamir@ia.ac.cn, gang.xiong@ia.ac.cn, lizhishuai2017@ia.ac.cn, aiar.th@stu.xjtu.edu.cn, zhen.shen@ia.ac.cn, binhu@ia.ac.cn

摘要：交通拥堵是全球面临的一个严重问题，在很大程度上对城市社区造成了各种影响，包括压力增大、交货延迟、燃料浪费和金钱损失。因此，准确的拥堵预测算法是减少这些不幸的根本。本文对决策树、逻辑回归和神经网络等交通拥堵预测系统进行了比较研究。本文利用五天的交通信息（1,231,200 个样本）来驱动预测模型。TensorFlow 和 Clementine 机器学习平台用于模型的数据预处理、训练和测试。混淆矩阵表明，决策树具有更好的预测性能，在 python 编程环境下，其准确率（97%）、宏观平均精度（95%）、宏观平均召回率（96%）和宏观平均 F1_score（96%）均领先于其他两种方法。此外，在 clementine 环境中对三种预测模型的性能进行了验证，决策树以 97.65% 的准确率优于其他所有模型。

*关键词*混淆矩阵 数据预处理 Python TensorFlow Clementine 环境 ITS

1. 导言

近年来，世界各地的交通流量逐渐增加，造成了污染、拥堵和事故。交通拥堵是最广为人知的主要问题，已成为大多数城市面临的挑战，并可能导致运输效率低下、温室气体大量排放和城市居民生活质量下降。为了提供路由服务和确保开放安全，有必要提供及时的预测信息，以实时响应和增加结构变化规划来减少持续的拥堵问题。由于城市化和稳态交通需求的持续增长，交通拥堵给公众造成了数十亿美元的持续损失（Wang 等，2017）。

在智能交通系统（ITS）应用中，预测准确的交通信息（如旅行时间、速度和拥堵状态）是一项重要工作。然而，交通状况的动态变化增加了预测的难度。事实上，城市环境中的公路和高速公路等道路状况会影响相应道路的行驶速度和拥堵状态（Liu 等人，2017 年）。

一般来说，城市地区的交通拥堵预测是一项异常麻烦的工作。迄今为止

最常用的交通预测方法是模拟和理论建模。如今，交通状况的实时海量数据集为设计不同的统计和数据驱动方法提供了优势。安装在每个道路网络中的广泛可用的无线传感器技术可用于获取车辆轨迹数据，以设计数据驱动的机器学习方法（Antoniou 等人，2007 年）。

本文接下来的内容安排如下。第 2 节讨论相关文献。第 3 节介绍所使用的方法。第 4 节展示了预测模型的建模和评估，第 5 节是结论意见。

2. 相关工作

近年来，人们在不同道路网络的交通拥堵预测领域开展了大量研究。

D'Andrea 和他的合作者 Marcelloni 开发了一种专家系统，通过当前和过去最近的交通速度来检测每个道路网络的交通拥堵情况（D'Andrea and Marcelloni, 2017）。同样，还提出了一种在网格框架中预测拥堵交通流量的可扩展方法（Gidofalvi，2015）。 Anwar 和合作者

10.1016/j.ifacol.2021.04.138

讨论了另一个框架，该框架使用基于光谱聚类的方法来监测相连的拥堵道路集（Anwar 等人，2016 年）。在扩展方法中，开发了一种定制的基于密度的带噪声应用空间聚类（DBSCAN）算法，用于检测和分析重复拥堵的网格集群（An 等人，2016 年）。

考虑到交通流密度和道路类型，Liang 和合作者提出了一种新的预测模型，利用单个路段及其相邻上游路段的当前流入、流出和交通量等参数，估算出该路段下一时间步的交通量，从而确定其拥堵状况（Liang 和 Wakahara，2014 年）。更重要的是，Xiangjie 及其合作者使用一种名为支持向量机的机器学习算法对该模型进行了改进，以预测下一时间步的交通速度和交通量，进而估计路段的拥堵状态（Kong 等人，2016 年）。此外，Xiaolei 和合作者通过使用受限玻尔兹曼机和循环神经网络开发了基于深度学习的预测模型，以预测下一时间步所有路段的交通拥堵情况（Ma 等人，2015 年）。

开发了一种高效的单参数预测模型来估计交通状况。自回归模型与其他预测算法相结合，保证了交通流预测性能的最优化（Kong 等人，2013 年；Davoodi 等人，2016 年）。通过提取奇异点概率，开发了一种以均方根误差为指标的人工神经网络组合模型（Qian 等人，2017 年）。

文献中的大部分工作都集中在对特定道路网络预先确定和固定数量的交通流参数（如平均速度、密度、队列长度、流速等）进行交通拥堵预测。然而，这些因素对交通拥堵的影响或贡献因地而异。也就是说，在某些地点，所有车辆的平均车速都很低，但这并不能反映出交通拥堵的存在。因此，本文提出了一种自适应预测模型，它能很好地结合大量道路网络中不同类型的交通状况。首先，使用 Tensorflow 机器学习框架来建立拥堵预测系统模型。其次，Clementine 环境验证了预测模型的性能。此外，本文还以准确度、精确度、召回率和 F1_score 为常用指标，对决策树、神经网络和逻辑回归等三种机器学习算法进行了比较研究。

3. 方法

3.1. 数据资源

GCM（加里-芝加哥-密尔沃基）走廊由十六个城市化县和连接三个城市的两千五百（2,500）英里公路组成。如表 1 所示，它由十个输入属性和一个目标属性组成。

训练集使用四天的交通信息（984,960 个样本），测试集使用一天的交通信息（246,240 个样本）。

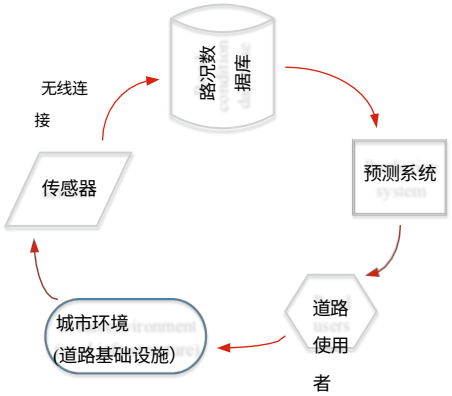


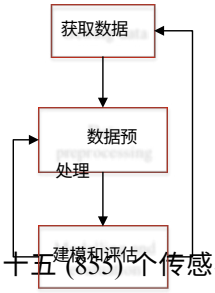
图 1：一般传感器集成预测系统一般传感器集成预测系统

表 1：预测系统属性

输入属性	目标属性
日期	
时间	
道路方向 道路类型	
道路传感器的链接 ID 道路长	拥堵程度
度	
行车时间 道路	
流量 车速	
道路占用率	

3.2. 使用的方法

预测系统的设计结合了获取数据、处理数据、预测算法建模和最终评估预测性能等操作。一般的系统设计流程如图 2 所示。



道路上安装了八百五十五（855）个传感器，每个传感器每天收集二百八十八 (288) 个数据流（每 5 分钟一个样本）

。每个传感器收集所在位置的实时交通信息，并通过无线连接定期发送到中央服务器。图 1 所示的通用传感器集成预测系统以闭环方式工作，可有效地向道路用户提供拥堵估计。每个流样本

图 2.一般工作流程

3.2.1. 获取数据

TensorFlow 和 Clementine 机器学习平台用于数据处理和预测算法建模。原始传感器数据必须经过系统处理，才能设计出有效可靠的预测模型。在数据预处理之前要进行的整体数据操作包括读取数据、检查数据是否正确、分析数据是否有误以及对数据进行分析。

数据信息，并计算整个数据集结构中遗漏的数据值。

3.2.2. 数据预处理

在设计机器学习算法时，数据预处理阶段需要花费更多的时间和精力。数据预处理首先要将原始传感器数据归类为不同类型的数据结构，包括分类数据（方向、道路类型、道路传感器的 LinkID、拥堵程度）和数值数据（日期、时间、道路长度、行驶时间、道路容积、汽车速度、道路占用率）。这两种数据类型都应进行预处理，为构建预测模型做好准备。以下三项任务与数据预处理有关。

A. 不必要的数据：

表 2：删除错误的属性标签

属性名称	错误数据
拥堵等级	未知拥堵等级道路方向
	方向类型未知道路
	STH_{bound}

表 2 中显示的不必要的错误属性标签必须删除。此外，还应该找出最不重要的属性。名为 "道路传感器的链接 ID "的属性在预测过程中没有任何作用，因此不予考虑。

从预测模型中可以看出，属性 "汽车速度 "与拥堵程度具有很高的相关系数（即 96% 的相关系数）。因此，应将其从属性中剔除。否则，预测模型可能会受到单一属性的严重影响。

B. 填补缺失值

在数据预处理中，有不同的方法来处理缺失数据：（1）忽略元组；（2）手动填充缺失值；（3）自动使用平均值、模式和中位数填充；（4）选择最可能的值。在本文中，缺失值是通过自动填写模式来处理的。

C. 为分类数据指定数值

数据集中的分类属性需要转换为数值。对于表 3 所示的三类属性，将分别分配数值。

表 3：分配数值

属性类型	值数值 方向	北界、南界
	疆界、西疆界和东疆	界

数据预处理完成后，数据集将剩下八个输入属性和一个目标属性。因此，在设计预测模型之前，需要确定训练数据集和测试数据集。训练数据集用于构建预测模型。训练数据集的大小是确保模型更好预测的一个影响因素。为了评估预测性能，测试数据集被输入到设计好的机器学习算法中，然后通过通用指标来衡量性能。

使用 TensorFlow 和 Clementine 机器学习平台设计了三种拥堵预测系统，包括决策树、逻辑回归和神经网络。准确度、精确度、召回率和 F1_score 是分别评估这些模型的典型指标。

混淆矩阵用于总结分类算法的性能。表 4 显示了路网的四级分类混淆矩阵，包括非拥堵、轻度拥堵、中度拥堵和高度拥堵。

表 4：四级混淆矩阵

		预测值			
	拥堵程度	非	灯光	中型	高
实际价值	非	NN	NL	NM	NH
	灯光	LN	LL	LM	LH
	中型	明尼苏达州	ML	MM	MH
	高	HN	HL	HM	HH

在哪里？

NN：真实预测不拥堵

NL：预测不拥堵为轻度拥堵 NM：预测不拥堵为中度拥堵 NH：预测不拥堵为高度拥堵 LN：预测不拥堵为轻度拥堵 LL：预测轻度拥堵为真正拥堵

LM：轻度拥堵被预测为中度拥堵 LH：轻度拥堵被预测为高度拥堵 MN：ML： 中度拥堵被预测为轻度拥堵 MM： 中度拥堵被预测为真正拥堵

道路类型、	高速公路、匝道、干道	分别为 0、1、2 和 3
	高速公路快车道、地方道路和高速公路可逆车道	0、1、2、3、4、5 和 6
拥堵程度	非拥堵，轻度拥堵、中度拥堵和高度拥堵	九十分别为 0、1、2 和 3

MH：中度拥堵被预测为高度拥堵 HN：预测高度拥堵为不拥堵 HL：预测高度拥堵为轻度拥堵 HM：预测高度拥堵为中度拥堵 HH：预测高度拥堵为真正拥堵

4. 建模与评估

4.1. 在 python 编程环境中建模

4.1.1. 逻辑回归

逻辑回归预测算法是根据路网数据集设计的。因此，模型生成的混淆矩阵如表 5 所示，可用于计算性能评估指标。

表 5：基于 Python 的逻辑回归混淆矩阵

		预测值			
	拥堵程度	非	灯光	中型	高
实际价值	非	168496	4461	25	0
	灯光	6968	30798	529	0
	中型	42	3517	6637	5
	高	0	7	35	956

将测试数据集输入逻辑回归预测模型，就能得出四个拥堵等级的预测概率。此外，图 3、图 4、图 5 和图 6 分别显示了 "非拥堵"、"轻度拥堵"、"中度拥堵 "和 "重度拥堵 "等每种概率的 10 个二进制直方图。

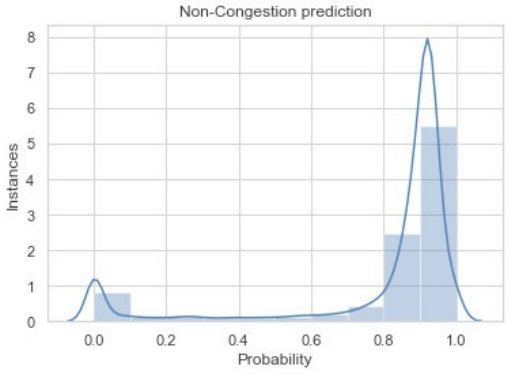


图 3.逻辑回归案例中不拥堵的概率

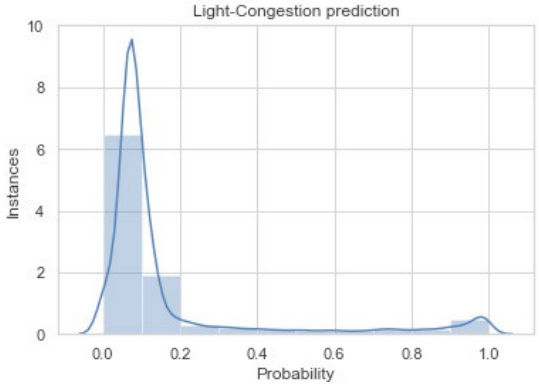


图 4.逻辑回归案例中轻度拥堵的概率

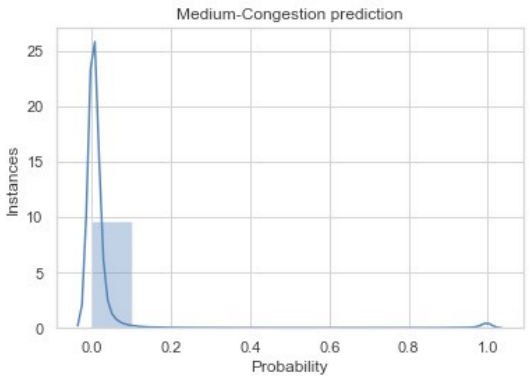


图 5.逻辑回归中的中等拥堵概率个案

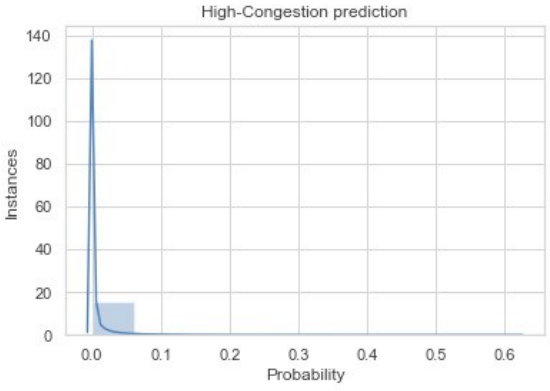


图 6.逻辑回归中的严重拥堵概率个案

概率分布图按递减顺序展示了四级拥堵预测：不拥堵、轻度拥堵、中度拥堵和高度拥堵。在特定的测试数据集中，非拥堵的概率相对较高，而严重拥堵的概率较低。

4.1.2. 决策树

根据决策树预测算法可以生成表 6 所示的混淆矩阵。矩阵的相应值可计算出评价指标。

表 6：基于 Python 的决策树混淆矩阵

		预测值			
	拥堵程度	非	灯光	中型	高
实际价值	非	168235	4744	3	0
	灯光	1701	36393	198	3
	中型	0	260	9892	49
	高	0	0	51	947

通过测试数据集生成四级拥堵概率分布，作为决策树预测模型的输入。图 7、图 8、图 9 和图 10 分别显示了 "非拥堵"、"轻度拥堵"、"中度拥堵 "和 "重度拥堵 "等拥堵等级的 10 个二进制直方图。

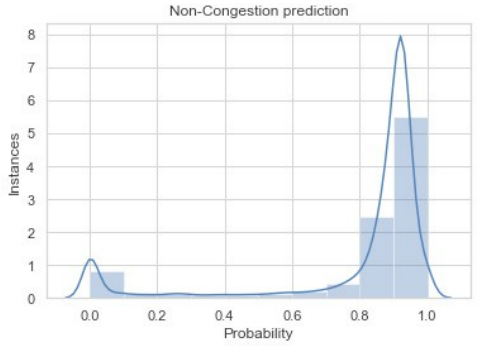


图 7.决策树情况下不拥堵的概率

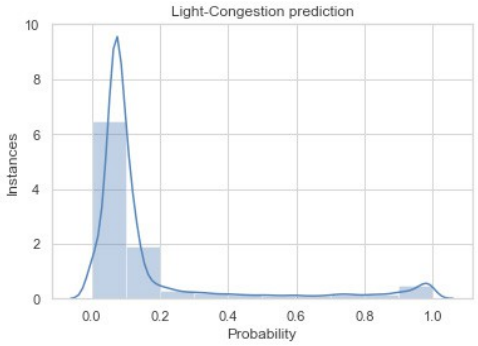


图 8.决策树情况下轻度拥堵的概率

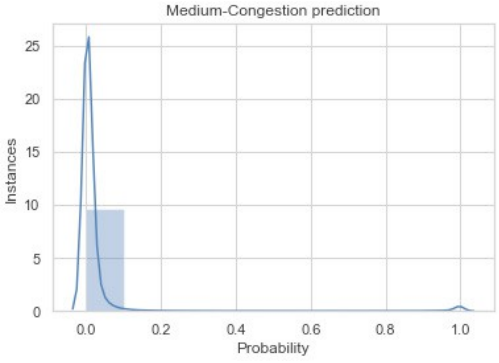


图 9.决策树情况下属于中度拥堵的概率

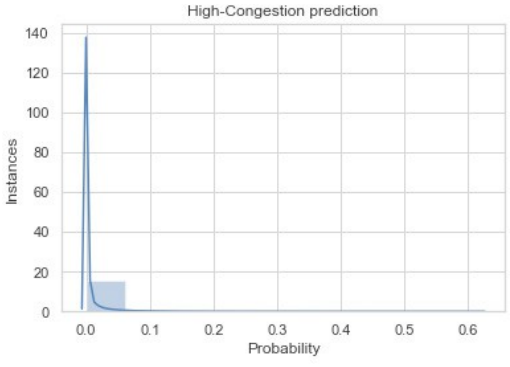


图 10.决策树情况下出现严重拥堵的概率

与逻辑回归预测方法类似，概率预测图按递减顺序显示四个拥堵等级：非拥堵、轻度拥堵、中度拥堵和高度拥堵。相对而言，在特定的测试数据集中，非拥堵的概率较高，而严重拥堵的概率较低。

表 7 用准确度、精确度、召回率和 F1_score 等常用指标评估了两种拥堵预测模型的性能。

表 7：python 环境下的性能指标 指标 逻辑回归

	决策树 准确率	0.93
	0.97	
精确度	0.92	0.95
召回	0.85	0.96
F1_score	0.88	0.96

4.2. 在克莱门特环境中建模

4.2.1. 逻辑回归

如图 11 所示，在克莱门特环境中设计了逻辑回归预测模型。从表 8 所示的混淆矩阵来看，准确率达到了 95.69%。

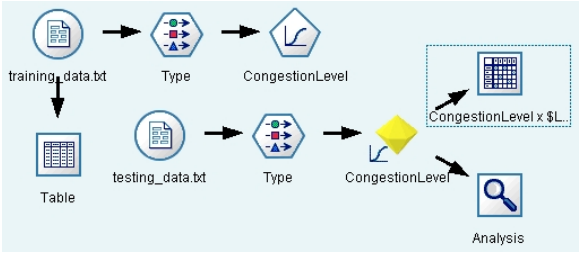


图 11.克莱门特环境下的逻辑回归模型

表 8：基于克莱门汀逻辑回归的混淆矩阵

		预测值			
	拥堵程度	非	灯光	中型	高
实际价值	非	168449	4533	0	0
	灯光	4519	33550	226	0
	中型	8	268	9911	14
	高	0	0	25	973

4.2.2. 决策树

决策树预测模型由以下参数设置组成：(1) 树深度为 30，(2) 修剪严重程度假定为 75，(3) 每个子分支的最小记录数假定为 4。此外，共生成了六百一十二条（612）决策规则。克莱门特模型如图 12 所示。根据表 9 所示的混淆矩阵计算出的准确率为 97.65%。

图 12.克莱门汀环境中的决策树模型

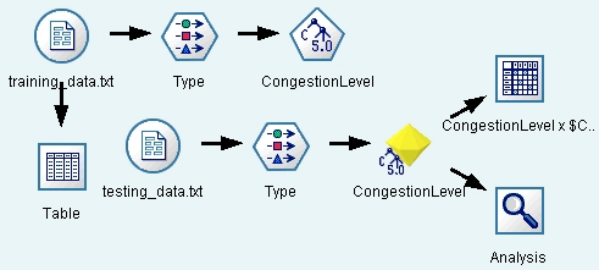


表 9：基于克莱门汀决策树的混淆矩阵

		预测值			
	拥堵程度	非	灯光	中型	高
实际价值	非	169894	3087	1	0
	灯光	1634	36485	176	0
	中型	0	230	9936	35
	高	0	0	65	933

4.2.3. 神经网络

神经网络模型由八个输入属性和一个目标属性构成，包括以下参数。(1) 网络有 8 个神经元的输入层（即 8 个输入属性）；(2) 网络有 1 个 10 个神经元的隐藏层（即隐藏层数和神经元数由实验设定）；(3) 网络有 4 个神经元的输出层（即 4 个拥塞层）；(3) 使用 *trainlm* 作为训练函数；(4) 使用 sigmoid 类型的激活函数；(5) 应用准确率作为性能函数。clementine 环境下的网络模型如图 13 所示，根据表 10 所示的混淆矩阵计算出的准确率为 93.75%。

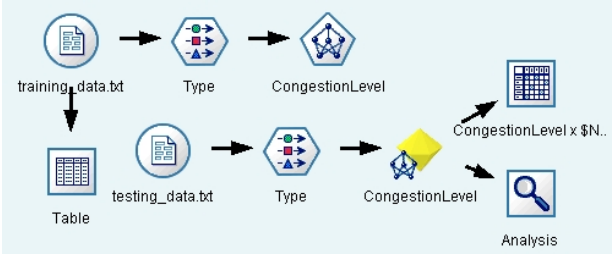


图 13.克莱门汀环境下的神经网络模型 表 10：基于克莱门汀的神经网络混淆矩阵

		预测值			
	拥堵程度	非	灯光	中型	高
实际价值	非	168100	4841	28	0
	灯光	6943	31019	312	0
	中型	11	743	9447	13
	高	0	0	21	998

表 11 以准确度作为通用指标，对三种预测模型的性能进行了评估和总结。

表 11：克莱门汀环境中的性能指标

指标	逻辑回归	决策树	神经网络

在克莱门特环境中进行了验证。这些模型通过提供实时拥堵信息，为人们和城市规划者做出了卓越贡献。整个预测算法的性能是通过准确度、精确度、召回率和 F1 分数等常用指标来评估的。这些指标都是通过相应的混淆矩阵计算得出的，结果表明决策树的预测能力优于其他两种预测模型。建议通过将路况数据库与卫星系统相连接，进一步改进预测模型。此外，还可以结合基于深度学习的预测算法，利用海量实时数据集提高预测能力。

鸣谢

本研究得到广东省科技计划项目（2019B1515120030）、国家自然科学基金资助项目（U1909204、61701471、61773381、U1811463、61773382 和 61872365）、东莞市科技计划项目（2019B1515120030）和东莞市科技计划项目（2019B1515120030）的部分资助。
创新人才项目（熊刚）。

参考资料

AN, S., YANG, H., WANG, J., CUI, N. & CUI, J. 2016.挖掘城市从配备全球定位系统的车辆移动数据中发现经常性拥堵演变模式。*信息科学》*，373，515-526。

Antoniou, C., Koutsopoulos, H. & Yannis, G. 2007.交通使用马尔可夫链模型进行状态预测。*2007 年欧洲控制会议，希腊科斯，2007 年 7 月 2-5 日。*

ANWAR, T., VU, H. L., LIU, C. & HOOGENDOORN, S. P. 2016.动态城市道路网络中拥堵分区的时空跟踪。*运输研究记录》*，2595，88-97。

D'ANDREA, E. & MARCELLONI, F. 2017.通过 GPS 轨迹分析检测交通拥堵和事故。*专家系统与应用》*，73，43-56。

DAVODI, N., SOHEILI, A. R. & HASHEMI, S. M. 2016.A macro-考虑驾驶员反应的交通流模型时间和距离。*非线性动力学》*，83，1621-1628。

Gidofalvi, G. 2015.可扩展的选择性交通拥堵通知。*第 4 届 ACM SIGSPATIAL 移动地理信息系统国际研讨会，美国华盛顿州贝尔维尤，2015 年 11 月 3 日*：ACM Press.

KONG, Q., ZHAO, Q., WEI, C. & LIU, Y. 2013.大规模城市道路网络的高效交通状态估计。*IEEE Transactions on Intelligent Transportation Systems,* 14, 398- 407.

KONG, X., XU, Z., SHEN, G., WANG, J., YANG, Q. & ZHANG, B. 2016.基于浮动车轨迹数据的城市交通拥堵估计与预测》。*未来新一代计算机系统》*，61，97-107。

LIANG, Z. & WAKAHARA, Y. 2014.动态路线引导系统的实时城市交通量预测模型。*EURASIP Journal on Wireless Communications and Networking,* 2014, 85.

LIU, B., CHENG, J., CAI, K., SHI, P. & TANG, X. 奇异点概率改进 LSTM 网络在长期交通流预测中的性能。2017

精确度	0.95	0.97	0.93

5. 结论

本文介绍了三种机器学习算法，包括决策树、逻辑回归和人工神经网络，以解决在给定多变量交通流参数的情况下预测交通拥堵传播模式的问题。本文使用 Python 编程环境来设计这三种拥堵预测模型，并在此基础上对其进行了优化。

新加坡，328-340 页。

Ma, X., YU, H., WANG, Y. & WANG, Y. 2015.大规模利用深度学习理论预测交通网络拥堵演变。*PLOS ONE*，10，e0119044。

QIAN, Z., LI, J., LI, X., ZHANG, M. & WANG, H. 2017.建模异构交通流：实用方法。*运输研究 B 部分：方法论》*，99，183- 204。

WANG, J., HU, F. & LI, L. Deep Bi-directional Long Short-Term Memory Model for Short-Term Traffic Flow Prediction.2017 Cham.Springer International Publishing, 306-316.