

学校代码: 10286
分 类 号: TP302.7
密 级: 公开
U D C: 004.9
学 号: 190001



东南大学
SOUTHEAST UNIVERSITY
硕士 学位 论文

基于深度强化学习的出行模式与时间选择

研究生姓名: 王于凯

导师姓名: 刘志远 教授

申请学位类别 专业硕士 学位授予单位 东南大学

一级学科名称 交通运输工程 论文答辩日期 2022年5月27日

二级学科名称 交通运输规划与管理 学位授予日期

答辩委员会主席 评 阅 人

2023年4月4日

東南大學
碩士學位論文

基于深度强化学习的出行模式与时间选择

专业名称: 交通运输工程

研究生姓名: 王于凯

导师姓名: 刘志远 教授

SOUTHEAST UNIVERSITY LATEX THESIS
TEMPLATE USER MANUAL
HOW TO WRITE A MASTER THESIS IN
AN ELEGANT WAY

A Thesis Submitted to
Southeast University
For the Professional Degree of Master of Engineering

BY
WANG Yukai

Supervised by
Prof. LIU Zhiyuan

School of Transportation
Southeast University
April 2023

东南大学学位论文独创性声明

本人声明所呈交的学位论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得东南大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

研究生签名: _____ 日期: _____

东南大学学位论文使用授权声明

东南大学、中国科学技术信息研究所、国家图书馆、《中国学术期刊（光盘版）》电子杂志社有限公司、万方数据电子出版社、北京万方数据股份有限公司有权保留本人所送交学位论文的复印件和电子文档，可以采用影印、缩印或其他复制手段保存论文。本人电子文档的内容和纸质论文的内容相一致。除在保密期内的保密论文外，允许论文被查阅和借阅，可以公布（包括以电子信息形式刊登）论文的全部内容或中、英文摘要等部分内容。论文的公布（包括以电子信息形式刊登）授权东南大学研究生院办理。

研究生签名: _____ 导师签名: _____ 日期: _____

摘 要

本文提出了一个新的东南大学 L^AT_EX 硕士研究生毕业论文模板，并说明了如何更优雅地写出一篇漂亮而无用的文章。

关键词： T_EX, L^AT_EX, 学位论文

Abstract

This article proposes a new Southeast University master degree thesis L^AT_EX template and explains how to elegantly write an article which is beautiful but full of shit.

Keywords: T_EX, L^AT_EX, Thesis

目 录

摘要	I
Abstract	III
第一章 绪论	1
1.1 研究工作的背景及意义	1
1.2 国内外研究	2
1.2.1 基于随机效用的离散选择模型	2
1.2.2 机器学习模型	3
1.2.3 强化学习模型	4
1.2.4 既有文献总结	5
1.3 本文的贡献与创新	5
1.4 技术路线图	6
第二章 深度强化学习相关知识	7
2.1 强化学习	8
2.1.1 相关术语	8
2.1.2 基于价值的强化学习	8
2.1.3 基于策略的强化学习	9
2.1.4 价值与策略相结合的强化学习方法	10
2.2 深度学习	11
2.2.1 神经网络	11
2.2.2 激活函数	13
2.2.3 损失函数及其优化算法	15
2.3 深度强化学习	16
2.3.1 深度 Q 网络	17
2.3.2 近端策略优化	17
2.3.3 深度确定性策略梯度	18
2.3.4 深度强化学习算法的对比与选择	19
2.4 本章小结	20
第三章 仿真实验场景的设计与构建	23
3.1 城市交通仿真平台的可行性分析	23

3.1.1	Vissim 介绍	23
3.1.2	SUMO 介绍	24
3.1.3	MATSim 介绍	26
3.1.4	仿真平台的选择	26
3.2	基于 SUMO 的城市交通仿真平台	28
3.2.1	平台设计目标	28
3.2.2	功能模块简介	28
3.3	实验场景的选择与搭建	29
3.3.1	路网的编辑与生成	29
3.3.2	出行模式的设计	31
3.3.3	流量的生成	34
3.4	本章小结	35
第四章	基于深度强化学习的出行模式与时间选择方法	37
4.1	马尔可夫决策过程框架	38
4.1.1	动作空间	38
4.1.2	状态空间	39
4.1.3	奖励函数	41
4.2	基于深度 Q 网络算法的模式与出发时间选择算法	42
4.2.1	神经网络结构设计	42
4.2.2	超参数的选择	44
4.2.3	模型的优化	46
4.3	基于聚类的深度强化学习方法	48
4.3.1	DBSCAN 聚类方法	48
4.3.2	聚类参数的选择	50
4.3.3	深度强化学习模型的改进	50
4.4	本章小结	52
第五章	模式与出发时间选择方法的训练与评估	53
5.1	模型的训练与分析	53
5.1.1	实验场景的设置	53
5.1.2	智能体的聚类与选取	54
5.1.3	训练及结果分析	56
5.2	模型的评估	59
5.2.1	模型泛化能力检验	60
5.2.2	部分信息感知的模型检验	60
5.2.3	模型参数的灵敏性分析	63

5.3 本章小结	66
第六章 总结与展望	69
6.1 工作总结	69
6.2 论文创新点	70
6.3 展望	71
致 谢	73
参考文献	75
作者简介	79

第一章 绪论

1.1 研究工作的背景及意义

城市旅行需求在过去几十年里由于人口和城市化的快速增长而急剧上升。增长速度远超过了交通基础设施的扩展。需求与供应失衡的后果就是全球范围内普遍存在的交通拥堵，这也说明了各种出行需求管理策略的出现和必要性。然而，这些策略的成功实施在很大程度上取决于如何理解和建模出行者的出行选择。为了获得准确的出行需求预测并实施有效的需求管理策略，研究人员和政府机构了解出行者在出行时如何进行决策将会至关重要。一旦决策者知道出行者在何时何地以及将采取什么模式出行，就可以提供有效的解决方案来缓解拥堵。因此，出行决策建模成为交通研究的关键。

研究人员通常将出行决策描述为不同维度的备选方案选择，例如出发时间、目的地、方式和路线。这些选择问题通常被描述为离散或者连续选择模型。早期的出行决策模型只考虑了一个维度，即从该选择维度的一组相互排斥的备选方案中选择了一个备选方案。然而在实际生活中，需要结合不同行为维度的进行多维决策才足以支持日益增长的拥堵管理策略应用。与传统的单维度出行选择模型不同，多维度模型考虑了不同选择之间的相关性。与模式和出发时间相关的两个关键且相关的选择从微观角度来看，这两个选择反映了个体对其出行的偏好。从宏观角度来看，它们决定了交通网络的时空旅行需求。我们强调，交通方式的吸引力以及其可能的选择取决于其服务水平。这种服务水平可能受到诸如拥堵定价、公共交通优先、各种类型激励等众多政策措施的影响。因此，为了评估这些政策措施，有必要建立一个同时考虑出行方式和出发时间选择的建模框架。

大多数现有的关于建模联合出行方式和出发时间选择的研究都是使用不同类型的离散选择模型（DCM），主要寻求随机效用最大化。特别是，考虑到它们在描述不同选择替代方案之间的相关性方面的能力，通常使用诸如多项式 logit (MNL)、嵌套 logit (NL)、交叉嵌套 logit (CNL) 等模型。利用随机效用最大化的模型依靠着其强大的理论依据而被广泛地应用。基于随机效用的模型是可以解释出行选择的基本理论，而对于复杂的决策过程建模的适用性，尤其是在选择预测中，可能会受到随机效用函数中线性结构的限制。对于多维选择问题，不同维度之间的关联结构也需要预先确定。尽管这些模型在理论上是解决旅行选择问题的典型解决方案，但在复杂决策过程中，由于随机效用函数的表述限制，它们的适用性可能受到限制。由于缺乏适应性以及出行者对旅行信息的不完美感知，由离散选择模型所推导出的出行选择可能不一定会导致最佳结果。

效用的随机成分不仅可以解释出行者对与观察信息的局限性，而且可以考虑决策者的不完全信息和偏好的随机变化。然而，以下事实支持了对基于学习方法的出行选择模

型的需求。首先，乘客在模式选择的决策过程，是由不同出行方式的服务水平信息告知和指导的。这些知识通常是通过各种方式获得的（包括出行经验），并且会随着时间动态变化。第二，出行决策受到一些行为因素的影响，其中部分乘客更倾向于（更少）选择（改变）他们已经习惯的模式。第三，交通系统的随机性和时间依赖性最有可能引起出行者的自适应模式切换决策，在这种决策中，出行者可能会根据以往的经验更新他们对每种出行模式的预期效用。传统方法不能解决决策过程中涉及的时间维度。因此，与传统的选择建模方法相比，基于学习的出行决策模型更可取。

近年来，深度强化学习已成为应对复杂决策问题的关键机器学习方法之一，原因在于它在复杂环境中具有较强的学习能力。这种学习能力正是传统离散选择模型所缺乏的，可以充分用于出行选择建模或出行推荐。出行选择的决策过程是一个复杂的过程，会受到环境的影响而不断地变化，通过建立传统的出行选择模型来解释出行行为的方法过于理想化。而此类场景很好地契合了深度强化学习“无模型、自学习、数据驱动”，使用深度强化学习的方法可以将此类复杂的模型使用深度神经网络进行描述，提取不同外界环境的特征数据如等待时间、出行成本等构建状态输入，再对出行者的出行行为进行优化，利用大数据训练网络增加其真实性和可靠性。相较于传统的离散选择模型，深度强化学习的方法对复杂的场景适应能力有极大的提升，并且适用的场景更加广泛。

1.2 国内外研究

1.2.1 基于随机效用的离散选择模型

在出行选择的模型中，通常使用基于随机效用的离散选择模型对不同维度的选择行为进行建模。从 McFadden^[1] 在 1973 年提出 Multinomial Logit (MNL) 模型以来，Logit 系列模型被广泛应用于出行决策问题。MNL 方法解决出行方式和出发时间选择问题的原因主要是由于 MNL 模型具有简单、易于应用和计算效率高等特点，而且在解决出行选择问题上已被广泛应用并取得了一定的成果。刘炳恩等^[2] 使用非集计离散选择模型，基于北京居民出行调查数据，建立了交通方式选择 MNL 模型，并通过参数标定和命中率计算验证了模型的有效性。结果表明，该方法能较全面地考虑影响因素，尤其是出行者的个人特性，提高了模型的预测精度和实用性。此外，MNL 模型还具有广泛的理论基础和应用场景，能够在一定程度上满足研究者对出行选择行为进行建模的需求。CHANG 等^[3] 在探讨音乐会参与者在出行方式和到达时间选择方面的行为，运用多项式 Logit 模型探讨了影响其出行选择的最有效因素，帮助预测在计划特殊事件中每种出行方式的时间依赖性旅行需求。然而，MNL 存在一个被广泛承认的问题：它假设了不相关的替代方案之间的独立性，也称为 IIA（独立不相关）特性。这意味着未被观察到的特征在不同选项之间是不相关的，然而在一些出行选择问题中，这个假设不成立。例如，在离散的出发时间选择中，相邻的出发时间区间的未观察到的特征往往表现出显著的相关性。

为了解决这一问题，NL 模型^[4] 被提出来克服 IIA 的限制。NL 模型能够识别嵌套

组内不同替代方案之间的相关性，因此能更好地描述出行者在做出选择时的现实决策过程，从而提高模型的预测精度和实用性。诸葛承祥等^[5] 基于北京市居民出行调查数据，建立了两个方向的 Nested Logit 模型来分析通勤者的出行时间和出行方式选择特征，结果表明出行时间—出行方式选择模型比出行方式—出行时间选择模型更合理，可以为高峰时期的管理政策提供理论依据。Koppelman^[6] 提出了一个基于 MNL 和 NL 模型的整合模型，通过对非独立误差、异方差和协方差等方面进行拓展，使得模型具有更高的灵活性和行为丰富性，适用于长途城际出行的出行方式选择建模。该研究采用了逐步放宽假设的方法，证明了整合模型在统计拟合和行为解释方面的有效性和优越性。但 NL 模型的问题是无法充分考虑相隔较远的出发时间选项之间的相关性^[7]。因此，有序广义极值模型 (OGEV) 被提出来解决这个问题，它可以提供每一对备选方案的相关参数，更全面地考虑不同备选方案之间的相关性。经过 Bhat 在 1998 年的测试^[8]，得出的结论是，NL 和 OGEV 模型的性能都优于 MNL。

在此之后，不同的研究人员针对问题的多样性提出了更先进的 NL 模型，如 Papola^[9] 使用的交叉巢式 Logit (CNL) 模型。杨励雅等^[10] 构建了 OGEV 理论的交叉巢式 Logit 模型，研究居民居住地、出行方式和出发时间的联合选择行为，并进行了弹性分析。结果显示，交叉巢式 Logit 模型比传统巢式模型更优，出行者优先考虑出发时间的改变，通勤距离在 10-20 公里时出行时间变化对出行方式选择的影响最显著。Ding 等^[11] 使用交叉巢式 Logit 模型，对马里兰-华盛顿地区的通勤行程数据进行了分析，探究出行方式和出发时间的联合选择。结果表明，该模型比传统的 MNL 和 NL 更优。

另一种改进的离散选择模型是 De Jong 等^[12] 在 2003 年提出的混合 Logit(MMNL) 模型，它通过改变 MNL 模型的参数随给定分布变化来考虑个体之间的异质性。栾鑫等^[13] 以南京为例，基于随机效用最大化理论，分析了影响居民出行方式选择的多重变量，并建立了混合 logit 模型，分析了家庭特征、个人属性、出行信息、出行 OD 位置之间的相互作用机理，并表明随着出行时间的增加，慢行方式的竞争性优势逐渐减弱，而在市郊、长距离行程中，公共交通或小汽车的出行方式更受旅客欢迎。然而，MMNL 的一个限制是，它需要对整个人口的参数分布进行特定的假设^[14]。这种限制可以通过潜在类 (LC) 模型来解决，在该模型中，数据被假设为由不同的潜在类别或群体生成，每个潜在类别或群体对应一种特定的行为模式或特征，该模型可以通过将总体划分为离散数量的类来捕获未观察到的偏好异质性^[15]。

1.2.2 机器学习模型

另一种研究出行决策的主流方法是机器学习。与统计方法不同，在统计方法中，研究人员试图确定模型结构和需要估计的参数，机器学习方法关注的是数据本身，并试图找到不同参数之间的关联^[16]。相较于随机效用的模型，机器学习模型的结构更加灵活，方便其探索不同特征之间的关联。针对出行决策的建模，主要有以下几种主流的机器学习方法：决策树模型，神经网络模型，以及支持向量机^[17]。与随机效用离散选择模型相

比，这些机器学习方法可以处理大型数据库。

决策树是一种基于数据集构建决策规则的分类模型，可以用于解决分类、回归等问题。在出行方式和出发时间选择问题中，可以使用决策树算法对不同变量对出行决策的影响进行分析，进而得到不同条件下的出行方式和出发时间的最佳选择方案。该方法的优势在于易于理解和解释，适用于大量数据和高维数据，并且具有较高的预测准确率和解释性。研究^[18] 使用南京经济适用房居民出行调查数据和谷歌地图补充数据，利用分类决策树分析经济适用房居民通勤时间模式特征。石庄彬^[19] 等利用梯度提升决策树方法研究老年人出行方式选择的决策机理，发现出行特征和建成环境是影响老年人出行方式选择的最重要因素，且建成环境对老年人出行方式选择的影响存在非线性效应和群体差异。Arentze 等^[20] 探讨了决策树模型在预测出行行为选择中的应用，提出了替换确定性行动分配规则为概率性规则的方法，并建议采用似然度量替代传统的命中率等模型拟合度量。实证结果表明，新的方法和度量能够为离散和连续选择问题提供更多信息。

支持向量机在处理高维度、非线性数据方面表现出色，因此被广泛应用于出行方式和出发时间选择问题的研究中。与传统的离散选择模型不同，支持向量机方法不需要假设各个出行选择行为之间的独立性和相互排斥性，因此可以更准确地反映实际出行决策中各种因素之间的复杂关系。另外，还可以通过设置不同的核函数，处理非线性问题，提高预测精度。

然而，机器学习方法很少能捕捉到对出行行为研究较为重要的因素，包括时间价值(VOT) 和弹性^[21]。此外，使用机器学习方法作为模型的主要框架还存在一个限制是机器学习模型对训练数据很敏感，在样本不足或有偏倚的情况下，会导致欠拟合或过拟合问题^[22]。

1.2.3 强化学习模型

强化学习作为一种被广泛应用的学习机制，是利用环境的反馈评价作为学习的输入，学习主体拥有较强的环境适应能力的机器学习方法，因此适用于重复日变的交通决策场景中。强化学习被用来解决各种领域的顺序决策问题，如机器人控制、电子游戏和系统优化等。强化学习的理论为人类行为提供了可解释的心理学和神经科学视角，即人类如何在给定的环境中计划自己的行为。此外，强化学习框架提供了智能决策的数学形式化形式，在智能体控制中具有强大而广泛的适用性，可直接应用于控制理论中顺序决策问题的求解。在交通领域，强化学习方法也受到了广泛的应用，例如交通流管理、自动驾驶，以及路线规划 [14]。近期，一些研究已经采用强化学习方法来建模出行者日常活动计划以及出行决策。

现有的出行决策与出行需求预测的研究工作多使用基于价值的强化学习方法，Janssens[15] 在 2007 年使用 Q-learning 的强化学习方法解决活动调度问题。Vanhulse[16] 等在 2009 年通过基于 Q-learning 的方法构建 MATSim 结构方程模型。Medhat[17] 等在 2008 年开发了一个更全面的动态公共交通路径和出行活动选择模型，称为 MILITRAS 系统，其

中的模型使用了预先设定的奖励(效用)函数。

近几年,深度强化学习在控制复杂智能体的决策行为上取得了巨大的成功,并将强化学习算法与许多神经相关因素的研究相结合,激发了大量使用人工神经网络作为通用函数逼近器的强化学习方法的研究。Hausknecht[18]等在2016年发表的著作研究了使用深度强化学习方法与多智能体合作行为。值得注意的是,它将多智能体研究中的矩阵博弈推广到更复杂的状态和行动空间。

1.2.4 既有文献总结

综上所述,国内外学者对出行选择问题的研究已取得了较为丰富的成果,并且形成了较为完备的理论体系,可为后续研究提供重要的理论技术支撑。其中,基于随机效用的模型是可以解释出行选择的基本理论,而对于复杂的决策过程建模的适用性,尤其是在选择预测中,可能会受到随机效用函数结构的限制。对于多维选择问题,不同维度之间的关联结构也需要预先确定。而机器学习方法关注的是数据本身,并试图找到不同参数之间的关联。相较于随机效用的模型,机器学习模型的结构更加灵活,方便其探索不同特征之间的关联。但是由于机器学习方法对数据的依赖性,所以使用机器学习方法时对数据的内容与质量要求也较高。目前使用强化学习方法解决出行选择的研究较少,并且多关注于单维的选择问题。在多维的出行选择问题,简单的强化学习方法在应对大规模的状态空间与动作选择集时会遇到维度灾难的情况。因此,本文将引入深度学习与强化学习结合的方法,提升复杂的场景适应能力,减少数据依赖性,解决多维出行选择中存在的维度灾难等问题。

1.3 本文的贡献与创新

在使用强化学习的仿真环境中,可以根据不同出行场景将出行者主要分成两种:有电子地图导航和无电子地图导航。在有电子地图导航的场景中,出行者信赖电子地图导航,会根据导航信息一般选择行程最短的模式和路径行驶。在此场景中,出行模式选择问题将转变为备选路径的行程时间预测和预估价问题。可以考虑使用预计到达时间(ETA)的计算方法解决[22]。在无电子地图导航的场景中,出行者只能依靠自身过往经验,根据经验记忆选择效用最大的出行模式和路径。这种场景下,出行者的每次决策都会得到环境带来的不同反馈,与强化学习的思想相契合。因此这种场景可以使用强化学习的模型解决。

相较于传统的离散选择模型,强化学习存在以下三点优势:

1. 强化学习模型直接与环境交互,减少了传统离散选择模型的假设限制。传统离散选择模型需要对环境的条件预先假设并检验,在复杂多变的环境下传统模型的弊端将会体现。
2. 强化学习的模型会减少采集数据的成本。一项新的交通政策在实施前需要大量的

仿真验证，传统模型需要采集大量的现实数据来验证模型的有效性。强化学习可以基于智能体已知的场景，通过更改仿真环境中的基础设施或策略，使得智能体学习处理未知场景下的决策行为。

3. 考虑智能体记忆能力，贴近实际决策过程。在强化学习的模型中，智能体做出动作后会根据以往经验以及自身的探索不断优化不同决策行为的价值及策略，这与实际中人在进行决策时的惯性一致。

1.4 技术路线图

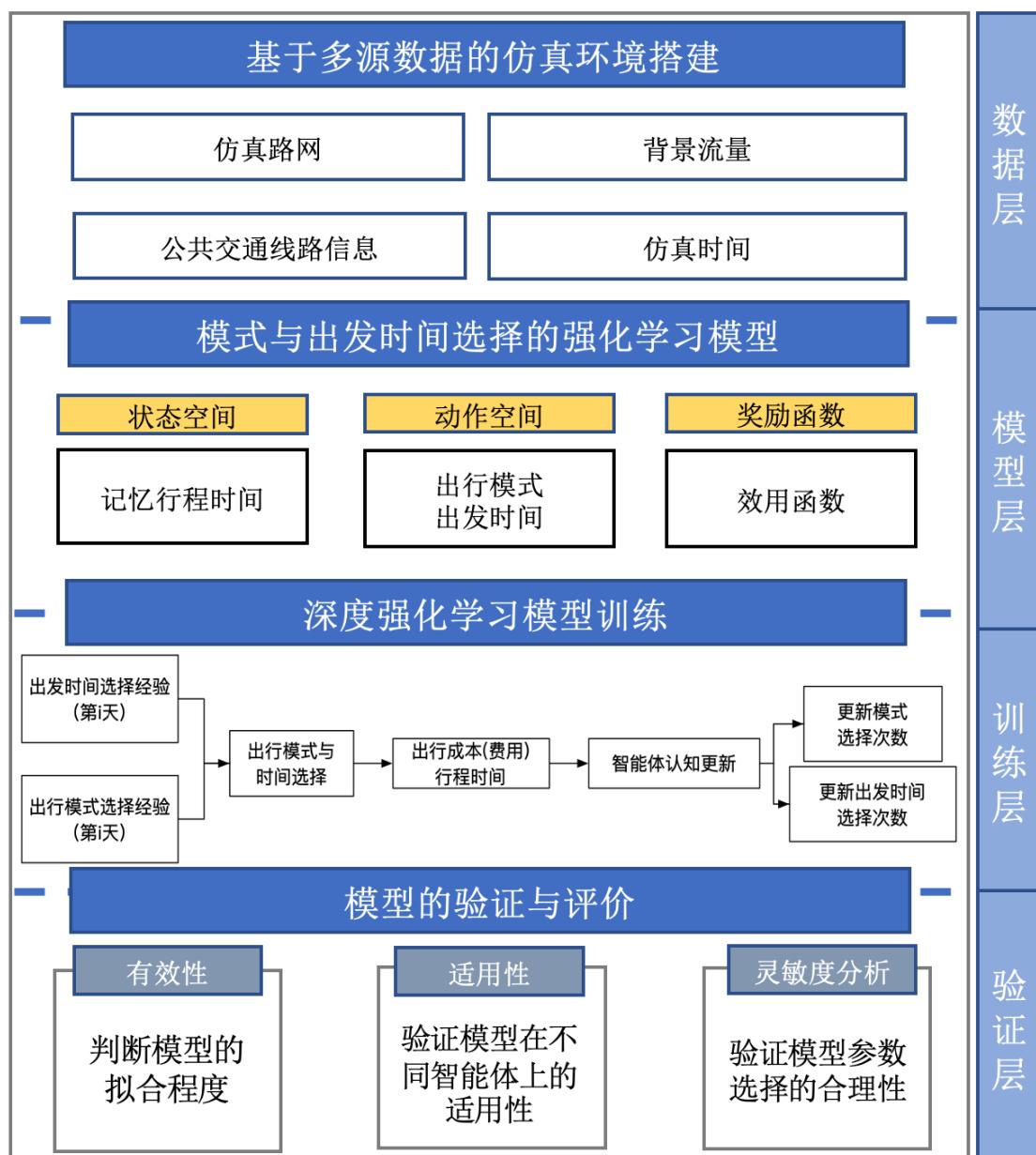


图 1-1 技术路线图

第二章 深度强化学习相关知识

在交通出行领域，如何合理地选择出行模式和时间，以达到高效、舒适、安全的出行，一直是研究者和决策者们关注的热点问题。传统的交通规划方法通常是基于流量预测和传统的数学模型来制定规划和决策，这种方法在一定程度上可以解决一些问题，但面临的挑战也日益增加。首先，传统方法很难处理复杂的交通场景和非线性的关系。其次，传统方法需要大量的数据和人工经验才能有效应对，但这些数据往往难以获取或者成本较高。此外，交通规划需要考虑的因素非常多，如出行模式、路线、时间等，而这些因素之间的复杂关系往往非常难以把握。

针对传统交通规划方法存在的问题，近年来，深度强化学习技术被引入到交通领域，成为了一种新的解决方案。深度强化学习通过学习交通系统的歷史数据，可以自动化地寻找规律和优化策略，以提高交通系统的效率和性能。通过引入深度强化学习技术，可以更好地解决交通出行领域中的一些问题。

对于模式选择问题，传统的方法主要是基于规则或者基于概率模型的方法。这些方法通常需要手动定义模型和规则，且模型和规则的适用性和可扩展性受到限制。而深度强化学习算法可以通过自主学习和适应环境的方式，学习到更加精准的出行模式选择策略，且不需要事先手动定义模型和规则。例如，可以通过深度强化学习来学习到乘客在不同时间、地点和情境下的出行偏好，以及在不同的出行模式之间做出选择的决策过程。

对于时间选择问题，传统的方法通常是基于历史数据或者基于概率模型的方法。这些方法存在着数据依赖性和模型精度的问题。而深度强化学习算法可以通过自主学习和适应环境的方式，学习到更加精准的出行时间选择策略，且不需要事先手动定义模型和规则。例如，可以通过深度强化学习来学习到乘客在不同时间、地点和情境下的出行偏好，以及在不同的出行时间之间做出选择的决策过程。

因此，引入深度强化学习算法可以有效地解决传统方法存在的问题，提高交通系统的智能化水平，优化交通出行效率，改善城市交通环境，为人们出行提供更加便捷、安全和可持续的选择。

本章主要将介绍了深度强化学习的相关知识。首先介绍强化学习的基本概念和相关术语，接着介绍深度强化学习基础知识中的深度学习，包括神经网络、激活函数、损失函数及其优化算法。然后详细介绍现在主流的深度强化学习算法，包括深度 Q 网络、近端策略优化、深度确定性策略梯度，并对这些算法进行了对比与选择。

2.1 强化学习

强化学习是一种基于马尔可夫决策过程的算法。在强化学习中，智能体根据环境状态和规定的策略进行交互，并根据环境给出的奖励信号产生新的状态。这个过程会不断循环，直到智能体完成设定的目标^[23]。强化学习算法利用产生的奖励数据来优化其行为策略，以获得最大的回报。本节将首先介绍强化学习算法的相关术语，然后根据智能体动作的选取方式，将强化学习方法分为基于价值、基于策略、以及基于价值和策略的三类方法，并对它们进行综述。

2.1.1 相关术语

智能体指的是具有独立思考能力且能够与环境进行交互的实体。在交通场景中，智能体可以是行人、车辆、信号灯等。

状态表示智能体对周围环境的感知，它是智能体感知历史的一个快照。所有状态的集合构成状态空间。

动作是智能体在某个状态下采取的行动。智能体可以采取的所有动作构成动作空间。

策略是智能体在当前状态下选择采取哪个动作的控制准则。它通常使用概率密度函数来表示，在每个状态下智能体采取各个动作的概率。

奖励是环境对智能体采取某个动作后的反馈效果。奖励可以为正向反馈或负向反馈。

回报是智能体从当前时刻开始采取行动到结束时所能获得的累积奖励之和。

状态转移是智能体采取某个动作后从当前状态转移到下一个状态的过程。状态转移通常具有随机性，这种随机性源自于环境^[24]。

2.1.2 基于价值的强化学习

基于价值的强化学习使智能体通过行动与奖励联系起来，通过试验和错误进行学习。智能体的主要目标是通过学习在不同情况下采取的最佳行动，随着时间的推移使其累积奖励最大化。在基于价值的强化学习中，智能体学习预测在特定状态下采取特定行动的价值。一个行动的价值通常被定义为智能体在特定状态下采取该行动并遵循特定政策所能获得的预期累积奖励。

在强化学习中，对于任意时刻 t ，在策略 π 下对状态 s_t 执行动作 a_t 会产生一个对应的奖励 R_t 。由于在强化学习研究背景下的问题具有马尔可夫性质，因此系统的总回报 U_t 与当前时刻的奖励 R_t 和未来时刻的奖励 R_{t+n} 有关。因此，可以表示为以下等式：

$$U_t = R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \dots + \gamma^n R_{t+n} \quad (2.1)$$

式中， γ 是折减因子。

在 t 时刻的回报 U_t 中，未来的奖励是与未来的状态和动作相关，而两者都具有随机性，所以需要通过对 U_t 求解期望值 $Q(s_t, a_t)$ 来消除随机性^[25]。

$$Q(s_t, a_t) = E[U_t | S_t = s_t, A_t = a_t] \quad (2.2)$$

因此， $Q_\pi(s_t, a_t)$ 可以用来表示状态动作对 (s_t, a_t) 的价值。其中， $Q(s, a)$ 是强化学习中的动作价值函数。通过寻找在 t 时刻所有策略 π 中动作价值函数 Q_π 的最大值，可以获得最优策略 π 的动作价值函数 $Q^*(s_t, a_t)$ 。

$$Q^*(s_t, a_t) = \max Q(s_t, a_t) \quad (2.3)$$

对最优策略 π 中的动作集 A 取最大值，即可获取每一次的最优动作 a^* 。

$$a^* = \operatorname{argmax} Q^*(s_t, a_t) \quad (2.4)$$

在基于价值的强化学习模型中，其主要目的就是逼近最优的策略 π 的动作价值函数 $Q^*(s_t, a_t)$ 。可以利用神经网络等方法近似动作价值函数进行求解。

由式2.4中动作价值函数 $Q^*(s_t, a_t)$ 可以得到价值最高的动作空间 A^* 。在强化学习中，一般使用神经网络的方法近似函数 $Q^*(s_t, a_t)$ ，网络的输入为状态 s ，网络的输出为不同动作的价值。则有：

$$Q(s, a; \mathbf{w}) \rightarrow Q(s, a) \quad (2.5)$$

式中， \mathbf{w} 是价值网络 (Value Network) 的参数。可以通过不同状态下的奖励 R 利用时序差分算法更新价值网络，使得网络的参数 \mathbf{w} 更加精确。

$$Q(s, a; \mathbf{w}) \approx R_t + \gamma \cdot Q(s, a; \mathbf{w}) \quad (2.6)$$

最常见的基于价值的强化学习算法是 Q-learning。Q-learning 是一种估计最佳动作价值函数的无模型方法，它代表了智能体在特定状态下采取特定动作并遵循最佳策略所能获得的预期累积奖励。一个状态-行动对的 Q 值使用贝尔曼方程进行更新，该方程指出，一个状态-行动对的最佳 Q 值等于即时奖励加上折现的最大预期未来奖励。Q-learning 是一个迭代过程，包括在智能体采取每个行动后更新 Q 值，并接受奖励形式的反馈。随着时间的推移，智能体学会了所有状态-行动对的最佳 Q 值，使其能够在每个状态下选择最佳行动，使其累积奖励最大化。

基于价值的强化学习方法已经在各种应用中取得了巨大的成功，包括游戏、机器人和自动驾驶汽车，使得智能体能够学习如何在复杂和不确定的环境中做出最佳决策。

2.1.3 基于策略的强化学习

基于策略的强化学习主要是为智能体在环境中采取行动寻找最佳策略，以使奖励最大化。策略是一种从状态到行动的映射，它告诉智能体在特定状态下应采取何种行动。

基于策略的强化学习的目标是找到一个策略，使智能体的预期奖励在一段时间内最大化。在强化学习中，使用概率密度函数 $\pi(a | s)$ 来控制智能体在不同状态下的动作选取，即策略函数。策略函数的输入为当前 t 时刻的状态 s_t ，输出为所有动作的概率值。依据策略函数得到的概率值对所有动作随机抽样后，确定在状态 s_t 下进行的动作 a_t 。当使用神经网络的方法近似策略函数时，则有：

$$\begin{cases} \pi(a | s; \theta) \rightarrow \pi(a | s) \\ \sum_{a \in A} \pi(a | s; \theta) = 1 \end{cases} \quad (2.7)$$

式中， θ 是策略网络的参数。

通过式2.2，对 $Q_\pi(s_t, a_t)$ 求取期望，通过积分消除概率密度函数 $\pi(\bullet | s)$ 中的动作 A 可以得到状态价值函数 V_π ：

$$V_\pi(s_t) = E_A[Q_\pi(s_t, A)] \quad (2.8)$$

状态价值函数 $V_\pi(s_t)$ 只与当前策略 π 和状态 s_t 有关。因此，状态价值函数可以用来评价当前状态下不同策略的价值。如果是离散的动作空间，状态价值函数 $V_\pi(s_t)$ 可以写作：

$$V_\pi(s_t) = \sum_a \pi(a | s_t) \cdot Q_\pi(s_t, a) \quad (2.9)$$

如果是连续的动作空间，则使用积分形式代替连加求和。由于连续动作空间的研究较复杂，并且大多数可以离散化，因此之后均为离散动作空间下的状态价值函数。通过式2.7中策略网络近似得到的策略函数 $\pi(a | s_t; \theta)$ ，可以近似状态价值函数：

$$V(s; \theta) = \sum_a \pi(a | s; \theta) \cdot Q_\pi(s_t, a) \quad (2.10)$$

基于策略的方法通常使用随机梯度上升法来更新策略。策略 π 由一组参数表示，使用策略梯度定理等技术计算出相对于这些参数的预期奖励的梯度。然后使用梯度上升法更新参数，以改进策略。策略学习是通过学习式2.9中的参数 θ ，得到价值最高的策略。这个过程中需要通过不断地改进策略网络参数 θ 的使 $V(s; \theta)$ 的值达到最大值。因此，可以将式2.9中的 $V(s; \theta)$ 对状态空间 S 求期望，将目标函数转化为 $J(\theta)$ ：

$$J(\theta) = E_S[V(S; \theta)] \quad (2.11)$$

基于策略的强化学习的缺点之一是它的计算成本很高，因为策略通常由大量的参数表示。此外，策略有时会卡在局部最优处，这可能使它难以找到全局最优策略。总的来说，基于策略的强化学习是一种在复杂和动态环境中寻找最优策略的强大方法。

2.1.4 价值与策略相结合的强化学习方法

在强化学习中，将策略网络与价值网络同时训练更新的方法称为策略价值结合学习方法。其目的为使智能体通过策略网络做出的动作价值越来越高的同时，使得价值网络

对动作价值的评价越来越精准。在策略价值结合学习方法中，可以把策略网络当作行动者（actor），价值网络当作裁判（critic）。价值网络会对智能体通过策略网络做出的动作进行评价，帮助更新策略网络参数，

通过联立式2.5与式2.10，可以得到通过神经网络方法近似后的价值函数 $Q(s, a; \mathbf{w})$ 与策略函数 $\pi(a | s; \boldsymbol{\theta})$ 。因此，状态价值函数可以写作：

$$\begin{cases} V(s; \boldsymbol{\theta}, \mathbf{w}) = \sum_a \pi(a | s; \boldsymbol{\theta}) \cdot q(s, a; \mathbf{w}) \\ \sum_{a \in A} \pi(a | s; \boldsymbol{\theta}) = 1 \end{cases} \quad (2.12)$$

此时，可以把策略网络当作行动者（actor），价值网络当作批评者（critic）。价值网络会对智能体通过策略网络做出的动作进行评价，帮助更新策略网络参数，使其目标函数 $J(\boldsymbol{\theta})$ 的值更大。行动者和批评者根据奖励和估计的状态-行动值进行更新。批评者通过最小化估计值和真实值（奖励和下一个状态的估计值之和）之间的平均平方误差来更新其对状态行动值的估计。行动者以估计的状态行动值为指导，通过最大化预期收益（未来奖励的总和）来更新其政策。

与其他强化学习算法相比，通过学习策略和价值函数，价值与策略相结合的强化学习方法可以比基于策略的方法更快地收敛，比基于价值的方法更稳定。它还可以处理高维的状态和行动空间，并且可以在实时环境中在线学习。价值与策略相结合的强化学习方法结合了基于政策和基于价值的方法的优点，可以同时学习最优政策和最优价值函数。然而，该方法需要仔细调整学习率和其他超参数以确保稳定的学习，而且它可能存在收敛问题和价值函数估计的偏差。

2.2 深度学习

深度学习在强化学习中发挥了重要作用，其中神经网络作为深度学习的核心，被广泛应用于强化学习中的状态表示、策略和价值函数估计等任务中。激活函数和损失函数也在强化学习中扮演着重要角色，对于神经网络的训练和优化具有至关重要的作用。本节将介绍深度强化学习中深度学习的应用，重点讨论神经网络在强化学习中的应用、常见的激活函数和损失函数，以及针对深度强化学习的优化方法。

2.2.1 神经网络

神经网络是一种机器学习模型，其灵感来自于人脑的结构和功能。它是由多个相互连接的节点或神经元组织成层，并形成一个系统。每个神经元接收来自其他神经元的输入，处理输入数据后产生一个输出信号。然后一个层的输出被用作下一层的输入，直至最终层输出结果。神经网络在训练期间从数据中学习经验并调整神经元之间连接的权重。权重决定了神经元之间的连接强度，它们使用优化算法进行更新，以达到预测输出和实际输出之间的误差最小化。神经网络已被应用于广泛的场景中，包括图像和语音识

别、自然语言处理以及时间序列预测等。目前在深度学习领域内使用的主流神经网络结构有前馈神经网络、递归神经网络和卷积神经网络。

神经网络最常用的架构是前馈神经网络，其输入数据是沿同一个方向流经各层。前馈神经网络主体是由一个输入层、多个隐藏层和一个输出层组成。各个层的职责各不相同：输入层接收输入数据，输出层产生神经网络的最终输出，隐藏层负责学习输入数据的特征。输入层的每个神经元代表输入数据的一个特征，而输出层的每个神经元代表神经网络预测的一个类别或一个值。隐蔽层中每个神经元的输出是通过对输入和权重的线性组合来计算的，并加入一个偏置项。然后，输出通过一个激活函数，将非线性引入网络。

隐藏层中每个神经元的输出可以按以下方式计算：

$$\begin{cases} z = \mathbf{w} * \mathbf{x} + b \\ a = f(z) \end{cases} \quad (2.13)$$

其中 z 是输入和偏置的加权和， \mathbf{w} 是权重向量， \mathbf{x} 是输入向量， b 是偏置项， $f(x)$ 是激活函数， a 是神经元的输出。

隐藏层中每个神经元的输出被用作下一层的输入，在最后一层的输出是神经网络的最终输出。在训练过程中，神经网络通过调整神经元之间连接的权重和偏差，使预测输出和实际输出之间的差异最小。

反向传播是一种用于训练神经网络的算法，通过调整神经元之间连接的权重和偏置项来训练。它是一种基于梯度的优化算法，计算损失函数相对于权重和偏置的梯度，并按照负梯度的方向更新它们。其中，损失函数衡量的是预测输出和实际输出之间的误差。回归问题最常用的损失函数是平均平方误差，而分类问题则使用交叉熵损失。损失函数相对于权重和偏差的梯度可以用微积分的链式法则来计算。链式法则指出，一个复合函数的导数等于其组成部分的导数的乘积。在神经网络的背景下，链式法则被用来计算损失函数相对于每个神经元输出的导数，然后通过网络传播误差来调整权重和偏差。

损失函数对于神经元输出的梯度可以按以下方式计算。

$$\frac{\partial L}{\partial a} = \frac{\partial L}{\partial z} * \frac{\partial z}{\partial a} \quad (2.14)$$

其中 L 是损失函数， a 是神经元的输出， z 是输入和偏置项的加权和。

式2.14右边的第一项是损失函数相对于加权和的导数，可以用激活函数的导数来计算。第二项是加权和相对于神经元输出的导数，也就是权重向量。然后，损失函数相对于权重和偏置的梯度可以通过在神经网络中各神经层向后传播误差来计算。首先计算输出层的误差，然后使用链式法则通过隐藏层向后传播，最后使用计算出的梯度和学习率更新权重和偏置项。权重通常在训练前被随机初始化。学习率决定了权重和偏置更新的步长，通常使用试验和错误或网格搜索来选择。高的学习率可能会导致对最优权重的过度拟合，而低的学习率则会导致缓慢的收敛。

训练神经网络的挑战之一是过拟合，即模型学习到的数据过于适合训练数据而在新数据上表现不佳。当模型相对于可用于训练的数据量来说过于复杂时，就会出现过拟合。正则化技术可以用来防止过度拟合。最常用的正则化技术是 L1 和 L2 正则化。L1 正则化给损失函数增加了一个惩罚项，与权重的绝对值成正比，而 L2 正则化则是增加了一个惩罚项，与权重的平方成正比。惩罚项的作用是鼓励权重变小，这有助于防止过度拟合。

2.2.2 激活函数

激活函数是神经网络的一个关键组成部分，如果没有激活函数，神经网络本质上只是线性回归模型，这将严重限制其灵活性。激活函数是应用于神经网络中每个神经元的输出的函数，目的是在模型中引入非线性，这对于捕捉数据中的复杂模式来说是必要的。其计算过程主要是在神经元输入的加权和添加一个偏置项后，对结果进行非线性转换。之后，激活函数的输出值将被传递到网络的下一层作为输入数据。目前在深度学习中应用最广泛的激活函数有 Sigmoid 函数，ReLU 函数和 Softmax 函数。

Sigmoid 函数是一条平滑的 S 形曲线，接受任何输入并输出 0 到 1 之间的值。Sigmoid 函数的公式如下：

$$f(x) = \frac{1}{1 + e^{-x}} \quad (2.15)$$

其中 x 是该函数的输入。

Sigmoid 函数图像如图2-1所示。

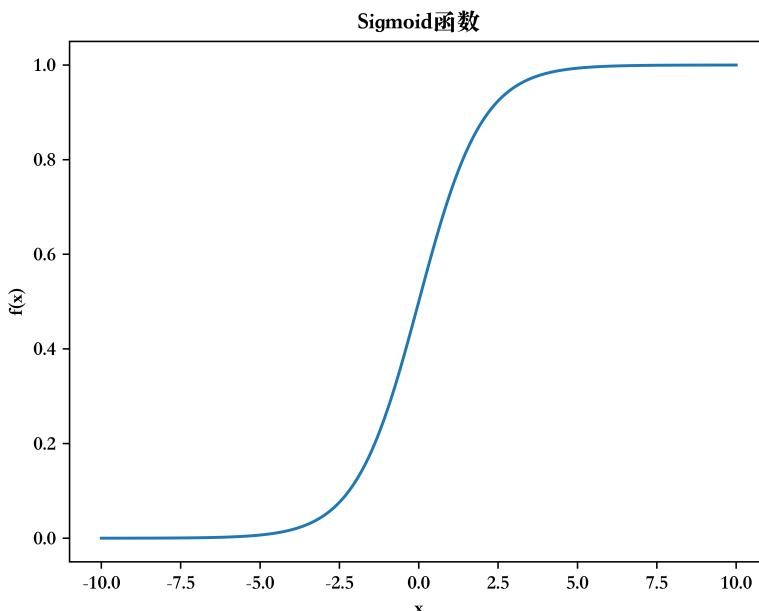


图 2-1 Sigmoid 函数图像

Sigmoid 函数是可微的，这意味着它的导数可以在任何一点处计算出来，这使得它很适合用于反向传播，也就是用于训练神经网络的算法。此外，Sigmoid 函数在 0 和 1

之间是有界的，这意味着它可以被解释为一个概率。然而，Sigmoid 函数也有一些缺点。Sigmoid 函数的主要问题之一是它存在梯度消失的问题。当 Sigmoid 函数的输入非常大或非常小时，函数的输出分别变得非常接近于 0 或 1，函数的导数也会变得非常小，这可能导致梯度在反向传播期间消失。因为梯度在网络中向后传播时变得非常小，这样会使得网络难以学习更加深度的表征。

ReLU 函数是神经网络中另一个常用的激活函数。它是一个片状线性函数，接受任何输入，如果输入是正的，就输出，否则就是 0。ReLU 函数的公式如下：

$$f(x) = \max(0, x) \quad (2.16)$$

ReLU 函数图像如图2-2所示。

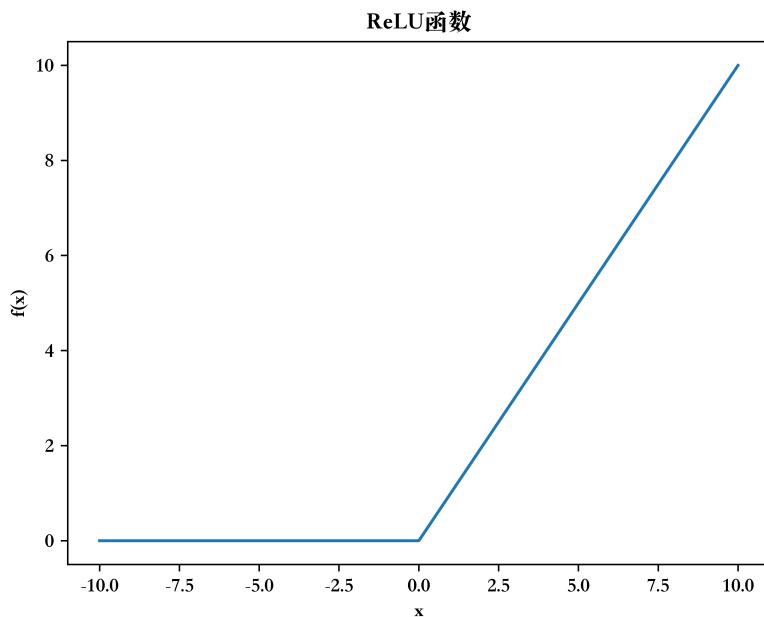


图 2-2 ReLU 函数图像

ReLU 函数的主要优点之一是它不存在梯度消失的问题。当 ReLU 函数的输入为正数时，该函数的导数为 1，这意味着在反向传播过程中，其计算出的梯度仍然很大。因为梯度在通过网络向后传播时不会消失，这使得网络更容易学习深度表征。ReLU 函数的另一个优点是它的计算效率高。由于该函数只是一个阈值操作，它可以用简单的逻辑运算来实现。然而，ReLU 函数的一个主要问题是，当 ReLU 函数的输入为负数时，该函数的输出为 0，这意味着神经元将会变得不活跃。这可能会导致整个神经元在训练过程中起不到任何作用，对网络的性能产生负面影响。

Softmax 函数是一种特殊的激活函数，常用于进行分类任务的神经网络的输出层。Softmax 函数在接受到一个输入矢量后，会输出一个和为 1 的数值矢量，可解释为概率。Softmax 函数图像如图2-3所示。

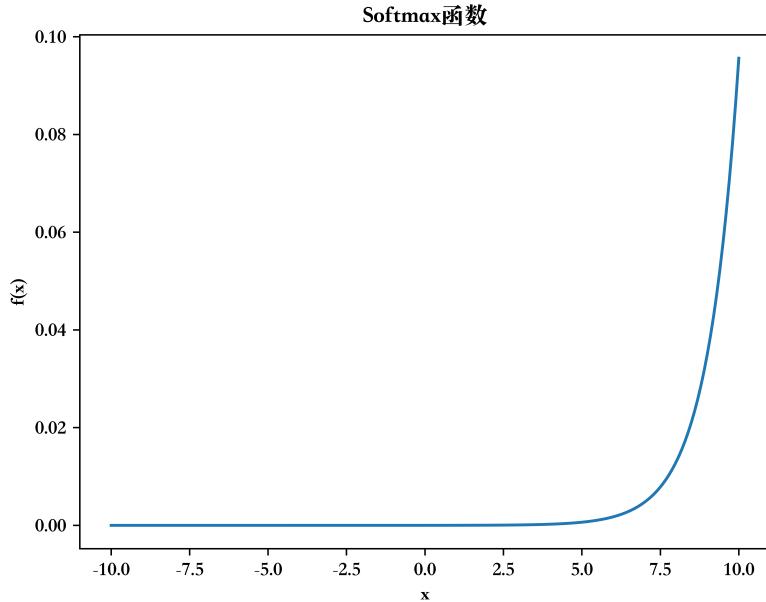


图 2-3 Softmax 函数图像

Softmax 函数由以下公式给出。

$$f(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}} \quad (2.17)$$

其中 x_i 是输入矢量的第 i 个元素，和是在矢量的所有元素上取的。

Softmax 函数的主要优势是可以确保网络的输出被解释为概率。这使得它非常适合用于分类任务，其目标是将输入分配到几个可能的类别中的一个。此外，Softmax 函数是可微分的，这意味着它可以用于反向传播来训练网络。

除了上面讨论的激活函数外，还有许多其他类型的激活函数被用于神经网络，包括双曲正切函数、指数线性单元 (ELU) 函数和缩放指数线性单元 (SELU) 函数等。这些激活函数都有自己的特性和使用场景，激活函数的选择取决于被解决的问题的具体需求，不同的激活函数可能更适合于不同类型的数据或任务。通过了解不同激活函数的特性，可以更好地设计神经网络，使其能够捕捉到实际数据中存在的复杂模式。

2.2.3 损失函数及其优化算法

损失函数是用来衡量神经网络的预测输出和实际输出之间差异的数学函数，其目标是提供一个衡量神经网络表现如何的标准。而损失函数优化是为了找到一组权重和偏置，使神经网络的预测输出与实际输出之间的差异最小。损失函数的选择取决于正在解决的具体问题。例如，对于回归问题，通常使用平均平方误差，而对于分类问题，通常使用交叉熵损失函数。

平均平方误差损失函数的公式如下：

$$L = \frac{1}{n} \cdot \sum_i y_i - \hat{y}_i^2 \quad (2.18)$$

其中 n 是数据集中的样本数, y_i 是第 i 个样本的实际输出, \hat{y}_i 是第 i 个样本的预测输出, 和是在数据集中的所有样本中取的。

交叉熵损失函数的公式如下:

$$L = -\frac{1}{n} \cdot \sum_i y_i \cdot \log \hat{y}_i + (1 - y_i) \cdot \log (1 - \hat{y}_i) \quad (2.19)$$

其中 y_i 是第 i 个样本的实际输出 (二元分类为 0 或 1, 多类分类为单次编码向量), \hat{y}_i 是第 i 个样本的预测输出 (0 和 1 之间的概率), 和是在数据集中的所有样本中取值。

神经网络最常用的优化算法是梯度下降法。梯度下降是一种迭代优化算法, 它沿着损失函数的负梯度方向更新神经网络的权重和偏置。负梯度指向最陡峭的下降方向, 这意味着沿着这个方向更新权重和偏置将导致损失函数值的减少。

梯度下降的更新规则如下:

$$w_i = w_i - \alpha \cdot \frac{dL}{dw_i} \quad (2.20)$$

其中 w_i 是第 i 个权重, α 是决定更新步长的超参数, dL/dw_i 是损失函数相对于第 i 个权重的偏导。

随机梯度下降 (SGD) 是梯度下降的一个变种, 它基于随机选择的小型训练数据子集来更新权重和偏差。这样做的好处是在计算上比梯度下降法更有效率, 因为它只需要计算一小部分数据的梯度。随机梯度下降的更新规则与梯度下降相似, 但使用在数据子集上计算的梯度而不是整个数据集, 更新规则:

$$w_i = w_i - \alpha \cdot \frac{dL_i}{dw_i} \quad (2.21)$$

其中 dL_i/dw_i 是损失函数相对于当前数据子集的第 i 个权重的偏导。

总之, 损失函数优化是深度学习的一个关键方面, 因为它直接影响到神经网络的性能。通过使用各种优化算法, 如梯度下降、SGD 和 Adam, 我们可以训练我们的神经网络来最小化损失函数, 并在各种任务中获得高精确度。

2.3 深度强化学习

深度强化学习是强化学习中应用深度学习的一种重要方法, 可以解决一系列复杂的任务。本节将介绍深度强化学习中常用的算法包括深度 Q 网络、近端策略优化和深度确定性策略梯度等。深度 Q 网络是基于 Q-learning 算法的一种深度学习算法, 可以直接学习环境状态和行为之间的映射关系, 以实现最优策略的学习。近端策略优化是一种基于梯度的方法, 用于直接优化策略函数, 以提高强化学习模型的性能。深度确定性策略梯度则是将近端策略优化与确定性策略相结合, 以实现高效的连续动作控制。本节将深入探讨这些深度强化学习算法的原理和应用。

2.3.1 深度 Q 网络

深度 Q 网络 (DQN) 是基于 Q-learning 算法的深度强化学习算法，其目的是使用深度神经网络来近似给定状态-动作对的最佳动作价值函数。在 Q-learning 中，智能体通过选择最大化动作价值 Q 的行动来学习最大化其预期的未来回报，Q 值是在给定状态下采取行动并在之后遵循给定策略的预期未来回报。

在式2.2和式2.3中分别给出了动作价值和最佳动作价值的定义，深度 Q 网络算法使用一个深度神经网络来近似动作价值函数。该神经网络将当前状态作为输入，为每个可能的行动输出一个动作价值，智能体所选动作的动作价值被用来更新神经网络的权重。在深度 Q 网络算法中，下一个状态的动作价值是用目标网络来估计的，目标网络是一个具有固定权重的主网络的复制模型。目标 Q 值被用来更新主网络的权重，主网络被用来估计当前状态的动作价值。DQN 中使用的平均平方误差损失函数，它用于衡量网络输出的 Q 值和目标 Q 值之间的差距。可根据式2.18得到用于更新神经网络权重的损失函数 $L(\theta)$ ：

$$L(\theta) = E \left[\left(r + \gamma \cdot \max_{a'} Q(s', a', \theta') - Q(s, a, \theta) \right)^2 \right] \quad (2.22)$$

其中， θ 是主网络的权重， θ' 是目标网络的权重， r 是在状态 s 下采取行动 a 后获得的即时奖励。

深度 Q 网络算法使用经验回放来提高学习效率和稳定性。经验回放存储了一个固定大小的经验缓冲区，神经网络通过从缓冲区中随机抽取经验进行训练。经验重放减少了连续经验之间的相关性，使学习过程更加有效。经验重放也有助于防止网络对最近的经验过度拟合。

深度 Q 网络算法使用 ε -greedy 探索来平衡对新行动的探索和对当前策略的利用。 ε -greedy 探索以 $1-\varepsilon$ 的概率选择具有最高 Q 值的行动，以 ε 的概率选择一个随机行动。

$$a_t = \begin{cases} \operatorname{argmax}_{a \in A} Q(s_t, a) & \text{当概率为 } 1 - \varepsilon, \\ \operatorname{rand}(a) & \text{当概率为 } \varepsilon. \end{cases} \quad (2.23)$$

总之，深度 Q 网络算法是一种被广泛使用的强化学习方法，用于训练强化学习问题中的动作价值函数。它通过使用神经网络来近似动作价值函数，解决了传统 Q-learning 算法的一些局限性，这使得它可以在类似的状态和行动中进行泛化。它还使用了一个经验重放缓冲器和一个目标网络来提高稳定性并防止过度拟合。DQN 算法已被成功应用于各种具有挑战性的决策问题中，包括游戏、机器人的控制等。

2.3.2 近端策略优化

近端策略优化 (PPO) 是一种近年来备受关注的深度强化学习策略优化算法，属于策略梯度方法，这意味着它将从当前策略收集的经验中学习。PPO 算法的基本原理是通过优化策略函数来寻找最优策略。通过式2.7可以得知，在强化学习中，策略函数 $\pi(a | s)$

通常是一个映射函数，它将当前状态作为输入，输出对应的行动。近端策略优化算法通过反复迭代，不断更新策略函数，使其逐渐趋于最优。

该算法的核心思想是限制每次策略更新的大小，避免策略函数发生大幅度变化，导致训练不稳定。具体而言，PPO 算法采用一种被称为“近端策略优化”的方法，通过在优化目标函数中增加一个约束项来限制每次策略更新的大小，这个约束项通常被称为“剪切项”，它会限制新旧策略之间的差异，并确保每次策略更新的大小不超过一个预设的阈值，以确保优化的稳定性。在近端策略优化算法中，优化目标函数是最大化经过剪切后的期望优势函数，而不是最大化期望回报函数。这里的优劣势函数表示当前策略相对于旧策略的性能提升程度。在近端策略优化算法中，使用剪切方法来限制当前策略和旧策略之间的差异，其优化目标函数可以写作：

$$L_{\text{CLIP}}(\theta) = \mathbb{E}_t \left[\min \left(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)} A_t, \text{clip} \left(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}, 1 - \epsilon, 1 + \epsilon \right) A_t \right) \right] \quad (2.24)$$

其中， θ 表示当前策略的参数， θ_{old} 表示旧策略的参数， $\pi_\theta(a_t|s_t)$ 表示在状态 s_t 下，当前策略选择动作 a_t 的概率， $\pi_{\theta_{\text{old}}}(a_t|s_t)$ 表示在状态 s_t 下，旧策略选择动作 a_t 的概率， A_t 表示在时刻 t 的优势函数，用于表示在给定状态 s_t 和行动 a_t 下，相比于平均水平的预期奖励，当前策略的表现优劣程度。具体地， A_t 的定义如下：

$$A_t = Q(s_t, a_t) - V(s_t) \quad (2.25)$$

式2.24中，目标函数 $L_{\text{CLIP}}(\theta)$ 的第一项表示策略更新的目标是最大化期望回报函数，第二项表示对策略更新进行剪切，确保新策略不会偏离原来的分布太远。通常来说， ϵ 的取值较小，可以取 0.1 或 0.2 等较小的数值。当 ϵ 取较小值时，第二项的影响较小，策略更新更倾向于最大化期望回报函数。

在近端策略优化算法中，更新价值函数的方法通常是通过均方误差（MSE）损失函数来实现，类似于式2.22。近端策略优化算法已经在多个实际应用场景中得到了广泛的应用，然而，近端策略优化算法也存在一些不足之处。例如，PPO 算法对于大规模离散动作空间的问题处理较为困难，同时其算法复杂度较高，需要消耗大量的计算资源。此外，近端策略优化算法在处理一些特殊场景下，如存在不确定性的环境、存在噪声的环境等，可能会出现训练不稳定的问题。

2.3.3 深度确定性策略梯度

深度确定性策略梯度（DDPG）算法是一种结合了深度神经网络和确定性策略梯度的强化学习算法，主要用于解决连续动作空间的问题，这类问题中，智能体需要在一个连续的动作空间中选择动作，因此传统的强化学习算法无法直接应用于该类问题。深度确定性策略梯度算法通过结合深度神经网络和确定性策略梯度的方法，解决了这一问

题。与传统的 Q-Learning 算法相比，深度确定性策略梯度算法在处理连续动作空间问题时，可以直接输出动作值，而无需在离散动作空间中搜索最优动作。

深度确定性策略梯度算法的主要思路是通过 Actor-Critic 模型来学习动作值函数，同时通过确定性策略梯度的方法来更新策略函数。其主要由四个部分组成：策略网络、价值网络、经验回放缓存和目标网络。其中策略网络和价值网络都采用深度神经网络来进行参数化，在训练时首先需要通过经验回放缓存来收集一定数量的状态转移样本，然后从中随机采样一批样本，用于网络的训练。策略网络的作用是输出在当前状态下最优的动作值。价值网络的作用是评估策略网络输出该动作值优劣的评估值，然后根据这个评估值计算出相应的策略梯度，最后通过反向传播算法更新策略网络的参数。经验回放缓存的作用是记录智能体在环境中的经验，并从中随机采样用于网络的训练。目标网络的作用是解决训练不稳定的问题，其参数是由价值网络参数每隔一段时间拷贝而来的。

深度确定性策略梯度算法的主要优点可以总结为以下几个方面。首先，深度确定性策略梯度算法采用 Actor-Critic 模型，可以直接输出动作值，因此可以直接处理连续动作空间问题。其次，深度确定性策略梯度算法引入目标网络和经验回放缓存技术，可以提高网络的训练稳定性。此外，深度确定性策略梯度算法使用深度神经网络处理高维状态空间问题，可以有效提取状态特征信息，提高智能体的决策效果。最后，深度确定性策略梯度算法结合了确定性策略梯度和 Q-Learning 的思想，可以同时处理具有连续动作空间和延迟奖励的问题。

然而，深度确定性策略梯度算法也存在一些缺点：1. 使用经验回放缓存和目标网络等技术来提高训练的稳定性的同时会导致了训练时间的增加。2. 深度确定性策略梯度算法有许多超参数需要调节，例如网络结构、学习率、优化器等，这些超参数的不同取值会影响算法的表现，参数调节较为复杂。3. 深度确定性策略梯度算法在处理连续动作空间问题时，通常需要引入噪声，但对噪声的选择和调节会对算法的表现产生较大影响，对噪声较为敏感。

2.3.4 深度强化学习算法的对比与选择

在解决出行模式和时间选择问题时，本章已经介绍了三种深度强化学习算法：深度 Q 网络、近端策略优化和深度确定性策略梯度。现在需要对这三种算法进行对比，并最终在出行模式与时间选择的场景下适用的方法。

深度 Q 网络算法是一种基于 Q-Learning 算法和深度神经网络的强化学习算法。其优点是训练速度快、稳定性高，而且适用于离散动作空间和连续动作空间。缺点是难以处理连续状态空间问题。

近端策略优化算法是一种基于策略梯度的强化学习算法。其优点是能够处理连续状态空间和连续动作空间问题，具有很好的表现。缺点是训练速度较慢，而且需要大量的超参数调节。在解决出行模式和时间选择问题中，我们的状态空间较大，因此 DQN 算法的表现可能会受到限制。

深度确定性策略梯度算法是一种基于 Actor-Critic 模型和深度神经网络的强化学习算法。其优点是能够直接处理连续动作空间问题和高维状态空间问题，而且能够处理具有延迟奖励的问题。缺点是训练时间较长，且对超参数的调节比较敏感。

表2.1总结并列出了三种算法的优缺点。

表 2.1 三种深度强化学习算法的对比

模型	优点	缺点
深度 Q 网络	训练速度快，稳定性高，可以处理高维状态空间和延迟奖励	只适用于离散动作空间，对参数的选择比较敏感
近端策略优化	收敛速度快，能够保持高样本效率	训练时间较长，需要手动调整超参数
深度确定性策略梯度	处理高维状态空间和延迟奖励，学习到高质量的策略	对参数的选择比较敏感，难以处理高维状态空间

因为问题的场景是一个离散的动作空间，而且深度 Q 网络算法的训练速度快、稳定性高，适合处理这种场景。同时，考虑到算法的可解释性和易于实现性，深度 Q 网络算法在这方面也具有优势。虽然近端策略优化算法在处理连续状态空间和连续动作空间问题上表现较好，但在本文的问题中，状态空间较大，训练速度也较慢，不太适合。深度确定性策略梯度算法可以处理连续动作空间和高维状态空间问题，但运算时间成本较高，且对超参数的调节比较敏感，因此也不太适合处理本文的问题。因此，综上所述，针对出行模式和时间选择问题，通过对三种主流的深度强化学习算法的优缺点分析，最终选择深度 Q 网络算法作为本文解决问题的主要深度强化学习方法。

2.4 本章小结

本章主要介绍了深度强化学习相关的知识和技术。在2.1小节中首先详细介绍了强化学习中的相关术语，包括智能体、状态、动作、策略、奖励、回报以及状态转移等，这些概念是深度强化学习算法的基础。随后，依据智能体选择策略的原理梳理了基于价值、基于策略以及基于价值和策略相结合的强化学习方法。这些方法各有特点和优缺点，基于价值的方法主要学习状态值函数或状态动作值函数，而基于策略的方法则直接学习策略函数，基于价值和策略相结合的方法则既学习状态值函数又学习策略函数。在实际应用中，需要根据问题的具体特点选择合适的方法来解决问题，以获得最佳的效果。介绍深度强化学习之前，在2.2小节简要介绍了深度学习的基本知识，包括神经网络、激活函数以及损失函数和优化算法等，这些知识对于理解深度强化学习算法是非常重要的。之后，在2.3小节详细介绍了三种深度强化学习算法中常用的方法，包括深度 Q

网络、近端策略优化、深度确定性策略梯度。这些方法使用深度神经网络来近似值函数或策略函数，使得智能体可以在高维状态空间中进行学习和决策。最后，通过对比了不同深度强化学习算法的优缺点，选择了深度 Q 网络的方法作为研究出行模式和时间选择的主要方法。

第三章 仿真实验场景的设计与构建

仿真平台选择与仿真环境的搭建在交通研究中扮演着非常重要的角色。随着城市化进程的不断加速和交通需求的不断增长，城市交通系统的规划和管理变得越来越复杂和困难。传统的基于观察和统计数据的研究方法已经无法满足交通系统优化和决策的需求，因此交通仿真技术应运而生。

交通仿真通过计算机技术对交通系统进行建模、仿真和分析，以获得交通系统行为和性能方面的信息。通过仿真平台的选择和搭建，可以对不同的交通系统进行模拟和测试，以评估不同交通系统的性能和效果。仿真平台可以帮助交通研究者理解交通系统的复杂性和动态性，预测未来交通系统的发展趋势，并提出优化方案和决策建议。

在交通研究中，不同的仿真平台有着各自的优缺点和适用范围。例如，商业软件VISSIM 可以模拟更加复杂和精细的交通系统，但是需要购买授权和付费，对于一些小型和非营利性项目来说，成本较高。相比之下，SUMO 是一款开源的、免费的仿真平台，可以轻松地进行定制和扩展，因此被广泛应用于学术界和非营利组织中。

在仿真平台的搭建方面，需要根据具体的研究问题和场景进行选择和设计。例如，在仿真场景的选择方面，需要考虑到不同的交通模式、路段的特点和流量情况等因素。在路网的编辑和生成方面，需要考虑到交通规划和设计的要求，并对道路的连接、交叉口的设计等进行合理的规划和调整。在流量的生成方面，需要根据实际情况进行仿真和预测，以准确地反映交通系统的行为和性能。

第三章将主要介绍仿真实验场景的设计与构建，首先分析不同仿真平台的优缺点，以及在深度强化学习的出行模式和时间选择研究中为何选择 SUMO 作为仿真平台。接着，详细介绍基于 SUMO 仿真平台的城市交通仿真平台，包括平台设计目标、功能模块简介。最后，完成本文实验所需场景的选择与搭建、路网的编辑与生成、出行模式的设计和流量的生成等方面。

3.1 城市交通仿真平台的可行性分析

3.1.1 Vissim 介绍

VISSIM 是由德国 PTV (Planung Transport Verkehr AG) 公司开发的一款功能强大的交通模拟软件，广泛应用于交通规划、设计、管理等领域。其模拟能力可以涵盖各种复杂的交通场景，如城市道路、高速公路、交叉口、公共交通、交通流、交通信号控制、公共交通路线规划和车辆路径规划等。图3-1是德国 PTV 公司在其官网提供的 VISSIM 多模式路网仿真样例。

VISSIM 的建模功能非常强大，用户可以使用图形用户界面轻松创建交通场景，包

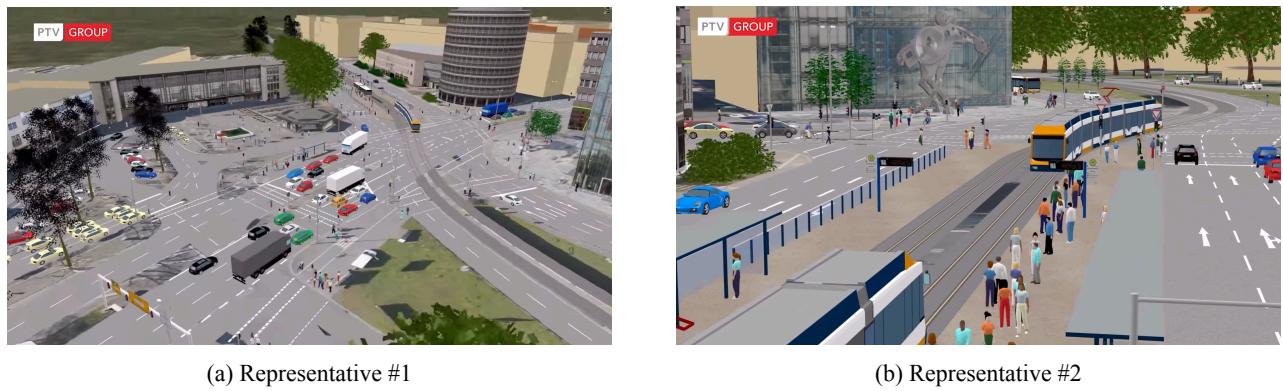


图 3-1 VISSIM 多模式路网仿真样例

括道路、交叉口和公共交通路线等。用户还可以调整交通流率、车辆类型、行人流量和其他参数，以建立一个真实的交通网络。智能交通生成器使用真实的交通数据，为一天中的不同时段、工作日和周末创建真实的交通流模式。同时，用户还可以根据自己的需要进行定制，调整交通量、速度和其他参数，以快速创建虚拟的交通场景。

VISSIM 的仿真模拟功能是其最核心的功能之一，用户可以通过 VISSIM 的仿真引擎高度准确地模拟各种交通场景，从简单的交叉口到复杂的城市网络。仿真引擎可以生成实时交通数据，如车辆速度、行驶时间和延迟时间等。用户可以对交通管理策略进行模拟，如交通信号控制、公共交通路线规划和车辆路径规划，并评估不同交通管理策略的性能，比较不同方案的效果，做出明智的决策。同时，用户可以通过虚拟实验，得出真实世界中的交通流动规律，从而更好地解决实际问题。

VISSIM 还提供了强大的分析功能，用户可以分析和可视化模拟结果，生成大量的图表和表格来分析交通流量、拥堵情况、车辆行驶时间、车辆速度和其他指标。分析功能还允许用户比较不同交通管理策略的性能，评估不同参数对交通流的影响。通过 VISSIM 的分析功能，用户可以更全面、更深入地了解交通流动规律，并进一步优化交通系统的设计和管理。

整体来说，VISSIM 是一款功能强大的交通模拟软件，可以广泛应用于交通规划、设计、管理等领域。VISSIM 的建模、仿真和分析功能都非常强大，可以帮助用户高度准确地模拟各种交通场景，评估交通管理策略的性能，并优化交通系统的设计和管理。作为交通专业人员、研究人员和政策制定者的重要工具，VISSIM 为实现更安全、更高效和可持续的未来交通系统做出了巨大的贡献。

3.1.2 SUMO 介绍

SUMO (Simulation of Urban Mobility) 是一款由德国科研机构 DLR 开发的开源交通仿真软件，广泛应用于城市交通规划、交通管理、交通研究等领域。该软件旨在提供高效、可扩展和高度自定义的交通仿真，以及良好的可视化和数据输出。SUMO 具有开

放性、高可靠性、高可扩展性和良好的可视化效果等特点，成为交通仿真领域的重要工具之一。图3-2展示了SUMO仿真软件的基本操作界面。

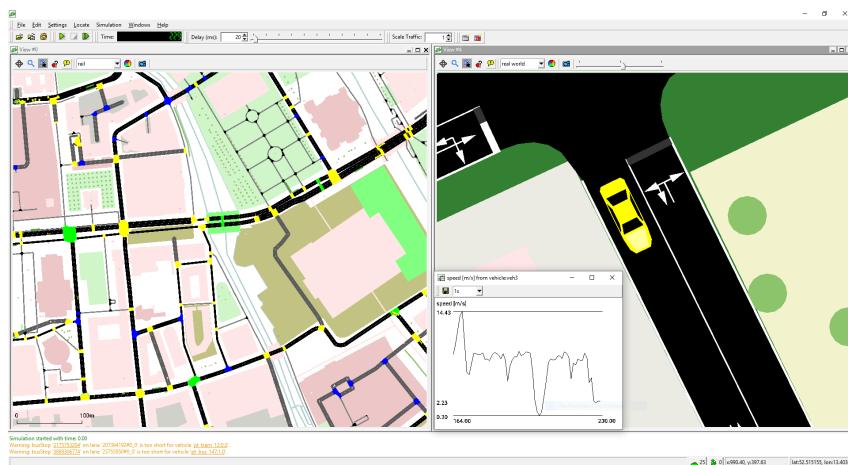


图 3-2 SUMO 仿真软件的操作界面

SUMO 的基本功能包括建模、仿真技术和可视化。用户可以使用该软件创建虚拟交通网络，并定义网络中的道路布局、交通流和车辆类型。SUMO 使用微观交通模拟引擎，可以模拟各种交通场景，如车辆路线、公共交通和行人运动，并生成实时交通数据。用户可以通过 SUMO 的图形用户界面对模拟结果进行可视化，也可以生成各种图表和表格总结模拟结果。

除了基本功能外，SUMO 还提供定制功能、多模式仿真、优化功能、并行化和集成等高级功能。用户可以通过 SUMO 定义影响模拟的不同参数、自定义脚本、创建自定义的车辆模型，并将其导入到模拟中。该软件支持多模式模拟，如汽车、公交车、火车和自行车，用户可以评估不同交通方式的性能，并对不同的交通计划进行比较。SUMO 还提供优化功能，可以优化交通信号灯时间、车辆路线和公共交通时间表，找到最佳的交通管理策略和交通计划。SUMO 提供并行化功能，允许用户在多个处理器上运行模拟，加快模拟时间，并模拟更大的交通网络。此外，SUMO 可以与其他软件工具集成，如交通流模型、地理信息系统（GIS）和数据分析工具，为用户提供一套全面的工具来模拟和分析各种交通情况。

总之，SUMO 是一个功能强大、用途广泛的模拟软件，为用户提供了一套全面的工具来模拟和分析各种交通情况。该软件支持多模式交通的模拟，包括汽车、公交车、自行车和行人，它还提供了一些高级功能，如交通需求生成、交通信号控制和车辆路由。SUMO 的灵活架构和开源代码库使其成为研究人员、交通专业人士和需要高度可定制模拟工具的决策者的热门选择。SUMO 有能力生成真实的交通数据，并提供对交通系统性能的洞察力，在设计、优化和评估交通系统方面发挥了关键作用，以实现更安全、更高效和可持续的未来。

3.1.3 MATSim 介绍

MATSim 是一款由瑞士苏黎世联邦理工学院 (ETH Zurich) 开发的开源交通仿真软件。该软件基于代理人建模，能够模拟不同交通方式的选择和相互作用，以及不同交通场景下的个人行为和整个交通系统的运行情况。此外，MATSim 提供了全面的可视化工具，使用户可以实时查看模拟的结果，包括交通流量、旅行时间等重要指标。图3-3展示了使用 MATSim 平台搭建的基于智能体的多模式研究场景。



图 3-3 使用 MATSim 平台搭建的基于智能体的多模式研究场景

MATSim 的模块化架构使用户能够根据自己的需求定制软件，以模拟新的运输系统或包括新的功能。这使该软件能够适应不同交通场景的具体要求，并成为研究人员和交通专业人士的热门选择。MATSim 还具有开源代码库的特点，允许用户修改软件并将其分发给其他人，使研究人员和交通专业人士能够合作和分享他们的工作。

MATSim 是一个功能强大的交通仿真软件，能够模拟多种交通方式和不同的交通场景，使用户能够评估不同交通系统的性能，并优化城市或区域内不同交通方式的使用。该软件已被广泛测试和验证，用于对全球不同城市和地区的交通系统进行建模和模拟，证明了其对交通系统进行精确建模和模拟的能力。MATSim 在设计、优化和评估交通系统以实现更安全、更高效和可持续的未来方面发挥了关键作用。

总之，MATSim 是一款功能强大、用途广泛的交通仿真软件，为用户提供了一套全面的工具来模拟和仿真不同的交通场景。该软件的模块化架构、多模式模拟能力、先进的可视化功能和开源代码库使其成为研究人员和交通专业人士的热门选择。MATSim 能够准确地模拟交通系统并评估不同交通方案的性能，在设计、优化和评估交通系统以实现更安全、更高效和可持续的未来方面发挥了关键作用。

3.1.4 仿真平台的选择

通过对三种主流交通仿真平台的介绍，表3.1总结梳理了各个仿真平台的优缺点。

SUMO、VISSIM 和 MATSim 是三款常用的交通仿真软件，它们各自具有一定的优劣势。其中，SUMO 是一款开源软件，用户可以免费使用和修改，而 VISSIM 则是商业软件，需要购买授权才能使用。MATSim 虽然是开源软件，但其复杂度和使用难度相对

表 3.1 不同仿真软件的对比

仿真软件	优点	缺点
VISSIM	城市和区域交通系统、交通行为建模、支持多种交通模式、交通流量和排放量测量评估	商业软件、非机动车和行人交通模式处理能力有限、复杂交通场景模拟效果可能不够准确、缺乏开源模型和工具
SUMO	速度快、处理大规模交通网络、开放性、开源模型和工具、多种交通模式和路段可调整性	交通行为建模较为简单、非机动车和行人交通模式处理能力有限、缺少全面的可视化工具
MATSim	模拟多种交通方式和不同的交通场景、模块化架构、提供全面的可视化工具、开源代码库	对于大规模交通网络的处理能力有限、个人行为的细节模拟较为复杂、某些交通模式和场景的支持仍不够全面、模拟速度相对较慢

较高。因此，使用 SUMO 平台作为仿真工具，能够在节约成本的前提下实现高质量的仿真。

在仿真速度方面，SUMO 的速度相对较快，适合处理大规模交通网络。而 VISSIM 的仿真速度较慢，在处理大规模网络时效果不如 SUMO。MATSim 在处理大规模网络时也存在一定的限制。因此，在大规模交通仿真方面，使用 SUMO 平台可以更快速地生成模拟结果，为交通规划和管理提供更快捷的决策支持。

SUMO 还提供了可定制的 API 和开源模型，便于进行二次开发和扩展，而 VISSIM 和 MATSim 的 API 和模型相对较为封闭，用户的自定义程度较低。这一特点使得 SUMO 平台能够更好地适应不同仿真需求，提高仿真的灵活性和可扩展性。同时，这也使得 SUMO 平台成为开发新的交通仿真工具和应用的理想平台，为深度强化学习等新兴技术的应用提供了广阔的发展空间。

除此之外，SUMO 对非机动车和行人等非机动交通模式的处理能力相对较强，而 VISSIM 和 MATSim 在这方面存在一定的局限性。这意味着 SUMO 平台可以更好地模拟现实交通场景，更准确地预测交通行为和流量，从而为交通规划和管理提供更高效的支持。另外，SUMO 平台上已经有相关的研究和开源工具，支持深度强化学习算法的应用。而在 VISSIM 和 MATSim 上的应用研究相对较少，缺乏相应的成熟工具和开源库。因此，使用 SUMO 平台能够方便地借鉴和复用这些成果，提高仿真的效率和准确性。

综上所述，SUMO 平台作为一款开源、高效、灵活和可扩展的仿真工具，非常适合

用于本文的基于深度强化学习的出行模式和时间选择仿真。

3.2 基于 SUMO 的城市交通仿真平台

在前一节中，介绍了仿真平台的选择，重点比较了 SUMO、VISSIM 和 MATSim 三个交通仿真软件的优缺点，并说明了为何在基于深度强化学习的出行模式和时间选择仿真中，使用 SUMO 作为仿真平台是一个较为理想的选择。本节将进一步介绍基于 SUMO 的城市交通仿真平台的设计与构建，其中包含了平台设计目标和功能模块简介两个方面。

3.2.1 平台设计目标

本研究的目的是利用深度强化学习方法研究出行者的出行模式和时间选择行为，并利用出行者的学习行为，实现提高出行效率。为此，我们采用 SUMO 仿真平台作为我们的仿真工具，以模拟不同的交通环境和交通管理策略。

SUMO 仿真平台是一个开源的、高度可定制的交通流量仿真器，支持多种车辆类型和行驶策略，如私家车、公共交通、自行车和行人等。它还提供了一个完整的仿真工具链，包括道路网络编辑器 Netedit、仿真器 SUMO、交互式仿真器 SUMO-GUI 和命令行接口 TraCI 等，支持多种路线选择算法和交通灯控制算法，用户可以根据他们的应用场景选择最适合的算法。然而，如果开发者直接在实验流中调用 TraCI 而不改进其上的结构以满足实际需求，将会导致代码结构混乱，代码可读性、复用性和扩展性较差等问题，这种方案只适用于简单试验性质的仿真，不能用于系统性的研究工作。因此，SUMO 仿真平台的设计目标是为强化学习算法训练和调试不同结构的区域路网提供支持，能够灵活地构造和修改仿真实验的网络道路拓扑结构和环境参数，复现经典深度强化学习算法并提供实验算法流程示例，同时对实时实验和模型数据具有一定的记录和可视化功能。

3.2.2 功能模块简介

SUMO 包括 Netedit、TraCI、SUMO-Tools 和 SUMO-Plugins 四个重要的功能模块。Netedit 允许用户创建、编辑和管理道路网络，TraCI 允许用户控制车辆和路口的运动和行为。SUMO-Tools 用于模拟和评估交通管理策略，SUMO-Plugins 用于扩展 SUMO 的功能和行为。这些功能模块和工具可以相互集成，为用户提供了一个强大和灵活的仿真环境，使得用户可以模拟不同的交通场景和交通管理策略，评估其效果和优化方案，进而提高交通流量的效率和可持续性。

Netedit 是 SUMO 中一个重要的功能模块，用于创建、编辑和管理道路网络。用户可以通过 Netedit 提供的图形界面添加和编辑道路、路口、车道和交通灯等元素，以模拟不同的交通环境和交通管理策略。同时，Netedit 支持多种地图格式，如 OpenStreetMap、Google Maps 和 Bing Maps 等。用户可以通过拖放操作和线条绘制工具轻松绘制道路网

络。Netedit 还提供一些有用的工具，如道路长度和宽度测量工具，帮助用户更准确地绘制道路网络。创建和编辑完成后，用户可以导出 Netedit 文件以供 SUMO-GUI 或其他仿真器使用。Netedit 还支持 Python 脚本，用户可以通过编写脚本自动化地创建和编辑道路网络。

TraCI 是 SUMO 中的另一个重要功能模块，它允许用户控制车辆和路口的运动和行为，以模拟不同的交通环境和交通管理策略。TraCI 提供了一系列 API 接口，用户可以通过 Python 脚本访问和修改仿真器中的车辆和路口状态，如车辆的位置、速度、目标路段和加速度等信息。用户也可以控制车辆的加速、刹车和转向，以及路口的红绿灯和车辆进出等操作。TraCI 还支持多种车辆类型和行驶策略，并提供有用的工具和功能，如路网查询、车辆生成和路由规划等，帮助用户更好地控制和管理交通流量。TraCI 可以与其他 SUMO 模块和工具集成，如 SUMO-GUI、SUMO-Tools 和 SUMO-Plugins 等，为用户提供更多的定制和扩展能力。

SUMO-Tools 是 SUMO 中用于模拟和评估交通管理策略的功能模块，如交通灯控制和路由优化。SUMO-Tools 包括一个流量分析器、一个行驶速度分析器和一个决策支持工具等，用户可以使用它们分析和比较不同的交通管理策略，以评估其效果和优化方案。SUMO-Plugins 是 SUMO 的可扩展模块，允许用户添加自定义的功能和行为，如新的车辆类型、路线选择器、行驶策略和交通管理策略等。SUMO-Plugins 可以是 Python 脚本、C++ 插件或 Java 插件等，支持多种路线选择算法和交通灯控制算法。SUMO-Tools 和 SUMO-Plugins 可以与其他 SUMO 模块和工具集成，如 SUMO-GUI 和 TraCI 等，为用户提供更多的定制和扩展能力。使用 SUMO-Tools 和 SUMO-Plugins，用户可以模拟和评估不同的交通管理策略，创建复杂的交通场景，并添加自定义的功能和行为，以满足特定应用程序的需要。SUMO-Tools 和 SUMO-Plugins 为用户提供了一个灵活的仿真环境，使得用户可以针对特定的交通问题和应用场景进行定制和扩展。

3.3 实验场景的选择与搭建

在基于 SUMO 的城市交通仿真中，实验场景的选择和搭建是十分重要的，它直接决定了仿真结果的可靠性和实用性。实验场景的选择和搭建包括了路网的编辑与生成、出行模式的设计和流量的生成等多个方面。在本节中，将详细介绍如何通过 SUMO 的功能模块和工具，搭建出逼近真实世界的仿真场景，从而进行有效的交通流量仿真和出行模式探究。

3.3.1 路网的编辑与生成

在本研究中，我们使用 SUMO 作为交通仿真引擎。SUMO 是一个开源的交通仿真软件，支持建模、仿真和分析各种交通场景，包括道路交通、公共交通、自行车交通和行人交通等。

为了研究基于深度强化学习的出行模式和时间选择，我们需要建立一个仿真环境来模拟交通流。这个仿真环境需要选择一个具有典型城市交通结构和各种出行方式的区域进行研究。因此，我们选择了中国苏州市的一个城市区域作为我们的研究对象，该区域面积大约为 20 平方公里，具有复杂的道路结构和丰富的出行方式，是一个理想的研究对象。图3-4是仿真场景选取的现实路网区域。



图 3-4 仿真场景选取区域

接着通过 OpenStreetMap (OSM) 来获取苏州市的路网几何和配置信息。OSM 是一个开源的地图服务，可以提供全球范围内的地图数据，包括道路、建筑和自然环境等。我们使用 OSM 提供的数据，利用 SUMO 软件进行仿真。在建立仿真环境时，我们考虑到苏州市的交通网络的实际情况，并进行了一些修正和调整。

SUMO 中的 netedit 模块是一个可视化的工具，用于编辑和修改道路网络。使用其添加一些缺失的路段和交叉口，还可以设置道路的形状和方向，以确保它与现有道路网络的拓扑结构相匹配，图3-5展示了在 netedit 模块中补齐 OSM 数据路网中缺失道路的场景。在实际道路网络中，路口通常是复杂的，并且需要进行精细调整才能更好地反映真实交通环境。可以在 Netedit 模块中设置路口的属性信息，图3-6展示了对交叉口的精细化操作界面，包括路口类型、交通信号灯、路口转向规则和优先级等。此外，还添加了一些交通规则和限制条件，如车速限制、交通信号灯和路口优先级等，以更真实地模拟交通环境。

图3-7是对选取区域使用 Netedit 模块搭建后的 SUMO 路网，包含 2,423 个节点和 4,970 条路段。

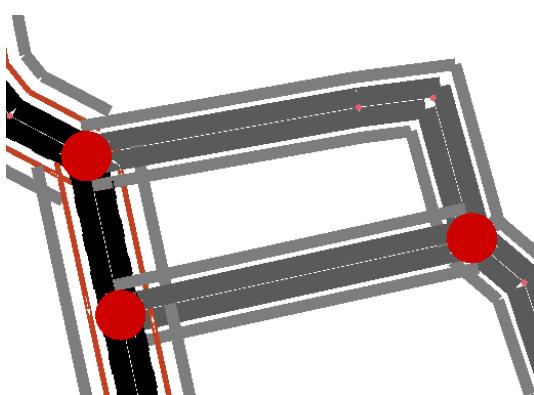


图 3-5 缺失道路的补齐



图 3-6 交叉口的精细化调整



图 3-7 SUMO 中的仿真路网

3.3.2 出行模式的设计

在本研究中，需要探究深度强化学习在出行模式和出发时间选择方面的应用。因此，在 SUMO 交通仿真软件中还需要建立一个具有多模式出行的仿真环境。共考虑私家车、公共交通（包括公交车和地铁）和自行车三种出行方式。私家车是最常见的出行方式之一，而公共交通和自行车则是城市交通中的重要组成部分。为了实现这种多模式交通网络的仿真，需要对这些出行方式进行适当的建模和参数化。首先，确定私家车、公共交通和自行车在 SUMO 仿真中的特征和参数。表3.2是在 SUMO 中设置的各个交通工具的基本参数，包括最大速度、最大加速度和最大减速度。

表 3.2 SUMO 中多模式交通的参数设置

交通工具	最大速度 [m/s]	最大加速度 [m/s^2]	最大减速度 [m/s^2]
私家车	120	2	4.5
公交车	70	1.5	3.0
地铁	80	0.9	1.5
自行车	25	1.2	2

其次，需要将这些参数输入到 SUMO 仿真的配置文件中。对于私家车，利用 SUMO 的 Car-Following 模型来建模其行为；对于公共交通，可以利用 SUMO 中的公交车和地铁模型来建模其行为；对于自行车，可以利用 SUMO 的 Bicycle 模型来建模其行为。通过将这些模型组合在一起，实现多种出行方式的仿真。

针对仿真区域中的公共交通（包括公交和地铁），还需要配置相关的停靠站点。为了实现这一点，需要从地图服务应用程序中提取公共交通运营信息，然后进行地图匹配。提取的信息包括线路 ID、停靠站或车站 ID 及其地理位置。这些信息可以通过地图匹配技术与实际地图中的位置进行匹配，从而实现公共交通在仿真环境中的配置。表3.3是获取到仿真区域内主要公共交通的基本信息，包含线路信息和停靠站信息。

表 3.3 公共交通的线路及停靠站信息

线路	停靠站数量	停靠站信息
地铁 1 号线	4	养育巷、乐桥、临顿路、相门
地铁 4 号线	5	北寺塔、察院场、乐桥、三元坊、南门
公交 2 路	4	学士街、养育巷、乐桥、市一中
公交 9011 路	3	南门、工人文化宫东、市红十字会东
公交 9003 路	14	苏州饭店、网师园北、苏州日报社、平桥直街、乌鹊桥北、乌鹊桥路、乌鹊桥南、工人文化宫南、工人文化宫、三元坊、苏州图书馆、饮马桥、乐桥、市一中、双塔
公交 9004 路	13	苏州饭店、网师园北、苏州日报社、平桥直街、虎丘山庄、水云路、黄桥、方洲花园、星海名城、幸福广场、苏州新区火车站、市一中、乐桥

在多模式交通网络中，不同的出行模式需要不同的道路和路径支持。SUMO 支持在

路网中建立自行车道，以支持自行车出行模式的模拟。SUMO 中的自行车道可以通过添加专用的边缘或中央车道来实现，以便自行车可以安全地与其他车辆分离行驶。在自行车道上，SUMO 也支持不同于汽车的最大速度、加速度和减速度等参数的配置，以适应自行车的行驶特性。图3-8展示了在 SUMO 路网搭建中的自行车车道。

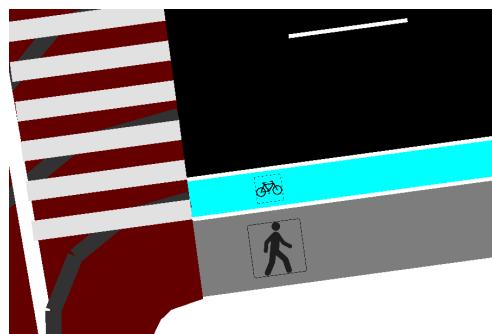


图 3-8 路网中的自行车车道

最终综合私家车、公共交通以及自行车的多模式出行设计，建立了一个具有复杂交通结构的仿真环境，可以用于研究基于深度强化学习的出行模式和时间选择。通过 SUMO 的仿真功能，可以对不同出行模式和时间选择的影响进行量化分析，帮助更好地了解并探究这些问题。图3-8展示了基于 SUMO 搭建的多模式仿真路网。



图 3-9 基于 SUMO 搭建的多模式仿真路网

3.3.3 流量的生成

本研究使用的旅行需求模型基于随机游走模型，该模型将人口分布、基础设施和交通需求等因素考虑在内，以生成一定数量的旅行需求。每个旅行需求包含起点、终点、出行方式和出行时间等信息。对于不同的出行模式，设置了不同的行驶速度和路线选择策略。在仿真过程中，通过修改路网配置文件，调整车道数量和长度等参数，以适应不同的交通流量和路况。图3-10展示了配置文件中的部分出行需求所包含的信息。其中，ID 为标识符，用于区分不同的车辆。每个车辆都有一个不同的 ID，以便在数据中跟踪和管理。DEPART 为出发时间，表示车辆从起点开始行驶的时间。它通常以秒为单位表示，并用于确定车辆在仿真中的出发顺序。ROUTE 为路线，表示车辆在行驶过程中经过的道路网络中的一系列边缘（或段）。边缘通过它们的 ID 连接在一起，形成车辆将行驶的整个路径。

```

<vehicle id="99.00" depart="99.00">
|   <route edges="503535567#3 -503535567#3 -503535567#2 -503535567#1 -503535567#0 -503535567#1 -503535567#2 -503535567#3 -503535567#4 -503535567#5 -503535567#6 -503535567#7 -503535567#8 -503535567#9 -503535567#10 -503535567#11 -503535567#12 -503535567#13 -503535567#14 -503535567#15 -503535567#16 -503535567#17 -503535567#18 -503535567#19 -503535567#20 -503535567#21 -503535567#22 -503535567#23 -503535567#24 -503535567#25 -503535567#26 -503535567#27 -503535567#28 -503535567#29 -503535567#30 -503535567#31 -503535567#32 -503535567#33 -503535567#34 -503535567#35 -503535567#36 -503535567#37 -503535567#38 -503535567#39 -503535567#40 -503535567#41 -503535567#42 -503535567#43 -503535567#44 -503535567#45 -503535567#46 -503535567#47 -503535567#48 -503535567#49 -503535567#50 -503535567#51 -503535567#52 -503535567#53 -503535567#54 -503535567#55 -503535567#56 -503535567#57 -503535567#58 -503535567#59 -503535567#60 -503535567#61 -503535567#62 -503535567#63 -503535567#64 -503535567#65 -503535567#66 -503535567#67 -503535567#68 -503535567#69 -503535567#70 -503535567#71 -503535567#72 -503535567#73 -503535567#74 -503535567#75 -503535567#76 -503535567#77 -503535567#78 -503535567#79 -503535567#80 -503535567#81 -503535567#82 -503535567#83 -503535567#84 -503535567#85 -503535567#86 -503535567#87 -503535567#88 -503535567#89 -503535567#90 -503535567#91 -503535567#92 -503535567#93 -503535567#94 -503535567#95 -503535567#96 -503535567#97 -503535567#98 -503535567#99 -503535567#100 -503535567#101 -503535567#102 -503535567#103 -503535567#104 -503535567#105 -503535567#106 -503535567#107 -503535567#108 -503535567#109 -503535567#110 -503535567#111 -503535567#112 -503535567#113 -503535567#114 -503535567#115 -503535567#116 -503535567#117 -503535567#118 -503535567#119 -503535567#120 -503535567#121 -503535567#122 -503535567#123 -503535567#124 -503535567#125 -503535567#126 -503535567#127 -503535567#128 -503535567#129 -503535567#130 -503535567#131 -503535567#132 -503535567#133 -503535567#134 -503535567#135 -503535567#136 -503535567#137 -503535567#138 -503535567#139 -503535567#140 -503535567#141 -503535567#142 -503535567#143 -503535567#144 -503535567#145 -503535567#146 -503535567#147 -503535567#148 -503535567#149 -503535567#150 -503535567#151 -503535567#152 -503535567#153 -503535567#154 -503535567#155 -503535567#156 -503535567#157 -503535567#158 -503535567#159 -503535567#160 -503535567#161 -503535567#162 -503535567#163 -503535567#164 -503535567#165 -503535567#166 -503535567#167 -503535567#168 -503535567#169 -503535567#170 -503535567#171 -503535567#172 -503535567#173 -503535567#174 -503535567#175 -503535567#176 -503535567#177 -503535567#178 -503535567#179 -503535567#180 -503535567#181 -503535567#182 -503535567#183 -503535567#184 -503535567#185 -503535567#186 -503535567#187 -503535567#188 -503535567#189 -503535567#190 -503535567#191 -503535567#192 -503535567#193 -503535567#194 -503535567#195 -503535567#196 -503535567#197 -503535567#198 -503535567#199 -503535567#200 -503535567#201 -503535567#202 -503535567#203 -503535567#204 -503535567#205 -503535567#206 -503535567#207 -503535567#208 -503535567#209 -503535567#210 -503535567#211 -503535567#212 -503535567#213 -503535567#214 -503535567#215 -503535567#216 -503535567#217 -503535567#218 -503535567#219 -503535567#220 -503535567#221 -503535567#222 -503535567#223 -503535567#224 -503535567#225 -503535567#226 -503535567#227 -503535567#228 -503535567#229 -503535567#230 -503535567#231 -503535567#232 -503535567#233 -503535567#234 -503535567#235 -503535567#236 -503535567#237 -503535567#238 -503535567#239 -503535567#240 -503535567#241 -503535567#242 -503535567#243 -503535567#244 -503535567#245 -503535567#246 -503535567#247 -503535567#248 -503535567#249 -503535567#250 -503535567#251 -503535567#252 -503535567#253 -503535567#254 -503535567#255 -503535567#256 -503535567#257 -503535567#258 -503535567#259 -503535567#260 -503535567#261 -503535567#262 -503535567#263 -503535567#264 -503535567#265 -503535567#266 -503535567#267 -503535567#268 -503535567#269 -503535567#270 -503535567#271 -503535567#272 -503535567#273 -503535567#274 -503535567#275 -503535567#276 -503535567#277 -503535567#278 -503535567#279 -503535567#280 -503535567#281 -503535567#282 -503535567#283 -503535567#284 -503535567#285 -503535567#286 -503535567#287 -503535567#288 -503535567#289 -503535567#290 -503535567#291 -503535567#292 -503535567#293 -503535567#294 -503535567#295 -503535567#296 -503535567#297 -503535567#298 -503535567#299 -503535567#300 -503535567#301 -503535567#302 -503535567#303 -503535567#304 -503535567#305 -503535567#306 -503535567#307 -503535567#308 -503535567#309 -503535567#310 -503535567#311 -503535567#312 -503535567#313 -503535567#314 -503535567#315 -503535567#316 -503535567#317 -503535567#318 -503535567#319 -503535567#320 -503535567#321 -503535567#322 -503535567#323 -503535567#324 -503535567#325 -503535567#326 -503535567#327 -503535567#328 -503535567#329 -503535567#330 -503535567#331 -503535567#332 -503535567#333 -503535567#334 -503535567#335 -503535567#336 -503535567#337 -503535567#338 -503535567#339 -503535567#340 -503535567#341 -503535567#342 -503535567#343 -503535567#344 -503535567#345 -503535567#346 -503535567#347 -503535567#348 -503535567#349 -503535567#350 -503535567#351 -503535567#352 -503535567#353 -503535567#354 -503535567#355 -503535567#356 -503535567#357 -503535567#358 -503535567#359 -503535567#360 -503535567#361 -503535567#362 -503535567#363 -503535567#364 -503535567#365 -503535567#366 -503535567#367 -503535567#368 -503535567#369 -503535567#370 -503535567#371 -503535567#372 -503535567#373 -503535567#374 -503535567#375 -503535567#376 -503535567#377 -503535567#378 -503535567#379 -503535567#380 -503535567#381 -503535567#382 -503535567#383 -503535567#384 -503535567#385 -503535567#386 -503535567#387 -503535567#388 -503535567#389 -503535567#390 -503535567#391 -503535567#392 -503535567#393 -503535567#394 -503535567#395 -503535567#396 -503535567#397 -503535567#398 -503535567#399 -503535567#400 -503535567#401 -503535567#402 -503535567#403 -503535567#404 -503535567#405 -503535567#406 -503535567#407 -503535567#408 -503535567#409 -503535567#410 -503535567#411 -503535567#412 -503535567#413 -503535567#414 -503535567#415 -503535567#416 -503535567#417 -503535567#418 -503535567#419 -503535567#420 -503535567#421 -503535567#422 -503535567#423 -503535567#424 -503535567#425 -503535567#426 -503535567#427 -503535567#428 -503535567#429 -503535567#430 -503535567#431 -503535567#432 -503535567#433 -503535567#434 -503535567#435 -503535567#436 -503535567#437 -503535567#438 -503535567#439 -503535567#440 -503535567#441 -503535567#442 -503535567#443 -503535567#444 -503535567#445 -503535567#446 -503535567#447 -503535567#448 -503535567#449 -503535567#450 -503535567#451 -503535567#452 -503535567#453 -503535567#454 -503535567#455 -503535567#456 -503535567#457 -503535567#458 -503535567#459 -503535567#460 -503535567#461 -503535567#462 -503535567#463 -503535567#464 -503535567#465 -503535567#466 -503535567#467 -503535567#468 -503535567#469 -503535567#470 -503535567#471 -503535567#472 -503535567#473 -503535567#474 -503535567#475 -503535567#476 -503535567#477 -503535567#478 -503535567#479 -503535567#480 -503535567#481 -503535567#482 -503535567#483 -503535567#484 -503535567#485 -503535567#486 -503535567#487 -503535567#488 -503535567#489 -503535567#490 -503535567#491 -503535567#492 -503535567#493 -503535567#494 -503535567#495 -503535567#496 -503535567#497 -503535567#498 -503535567#499 -503535567#500 -503535567#501 -503535567#502 -503535567#503 -503535567#504 -503535567#505 -503535567#506 -503535567#507 -503535567#508 -503535567#509 -503535567#510 -503535567#511 -503535567#512 -503535567#513 -503535567#514 -503535567#515 -503535567#516 -503535567#517 -503535567#518 -503535567#519 -503535567#520 -503535567#521 -503535567#522 -503535567#523 -503535567#524 -503535567#525 -503535567#526 -503535567#527 -503535567#528 -503535567#529 -503535567#530 -503535567#531 -503535567#532 -503535567#533 -503535567#534 -503535567#535 -503535567#536 -503535567#537 -503535567#538 -503535567#539 -503535567#540 -503535567#541 -503535567#542 -503535567#543 -503535567#544 -503535567#545 -503535567#546 -503535567#547 -503535567#548 -503535567#549 -503535567#550 -503535567#551 -503535567#552 -503535567#553 -503535567#554 -503535567#555 -503535567#556 -503535567#557 -503535567#558 -503535567#559 -503535567#560 -503535567#561 -503535567#562 -503535567#563 -503535567#564 -503535567#565 -503535567#566 -503535567#567 -503535567#568 -503535567#569 -503535567#570 -503535567#571 -503535567#572 -503535567#573 -503535567#574 -503535567#575 -503535567#576 -503535567#577 -503535567#578 -503535567#579 -503535567#580 -503535567#581 -503535567#582 -503535567#583 -503535567#584 -503535567#585 -503535567#586 -503535567#587 -503535567#588 -503535567#589 -503535567#590 -503535567#591 -503535567#592 -503535567#593 -503535567#594 -503535567#595 -503535567#596 -503535567#597 -503535567#598 -503535567#599 -503535567#600 -503535567#601 -503535567#602 -503535567#603 -503535567#604 -503535567#605 -503535567#606 -503535567#607 -503535567#608 -503535567#609 -503535567#610 -503535567#611 -503535567#612 -503535567#613 -503535567#614 -503535567#615 -503535567#616 -503535567#617 -503535567#618 -503535567#619 -503535567#620 -503535567#621 -503535567#622 -503535567#623 -503535567#624 -503535567#625 -503535567#626 -503535567#627 -503535567#628 -503535567#629 -503535567#630 -503535567#631 -503535567#632 -503535567#633 -503535567#634 -503535567#635 -503535567#636 -503535567#637 -503535567#638 -503535567#639 -503535567#640 -503535567#641 -503535567#642 -503535567#643 -503535567#644 -503535567#645 -503535567#646 -503535567#647 -503535567#648 -503535567#649 -503535567#650 -503535567#651 -503535567#652 -503535567#653 -503535567#654 -503535567#655 -503535567#656 -503535567#657 -503535567#658 -503535567#659 -503535567#660 -503535567#661 -503535567#662 -503535567#663 -503535567#664 -503535567#665 -503535567#666 -503535567#667 -503535567#668 -503535567#669 -503535567#670 -503535567#671 -503535567#672 -503535567#673 -503535567#674 -503535567#675 -503535567#676 -503535567#677 -503535567#678 -503535567#679 -503535567#680 -503535567#681 -503535567#682 -503535567#683 -503535567#684 -503535567#685 -503535567#686 -503535567#687 -503535567#688 -503535567#689 -503535567#690 -503535567#691 -503535567#692 -503535567#693 -503535567#694 -503535567#695 -503535567#696 -503535567#697 -503535567#698 -503535567#699 -503535567#700 -503535567#701 -503535567#702 -503535567#703 -503535567#704 -503535567#705 -503535567#706 -503535567#707 -503535567#708 -503535567#709 -503535567#710 -503535567#711 -503535567#712 -503535567#713 -503535567#714 -503535567#715 -503535567#716 -503535567#717 -503535567#718 -503535567#719 -503535567#720 -503535567#721 -503535567#722 -503535567#723 -503535567#724 -503535567#725 -503535567#726 -503535567#727 -503535567#728 -503535567#729 -503535567#730 -503535567#731 -503535567#732 -503535567#733 -503535567#734 -503535567#735 -503535567#736 -503535567#737 -503535567#738 -503535567#739 -503535567#740 -503535567#741 -503535567#742 -503535567#743 -503535567#744 -503535567#745 -503535567#746 -503535567#747 -503535567#748 -503535567#749 -503535567#750 -503535567#751 -503535567#752 -503535567#753 -503535567#754 -503535567#755 -503535567#756 -503535567#757 -503535567#758 -503535567#759 -503535567#760 -503535567#761 -503535567#762 -503535567#763 -503535567#764 -503535567#765 -503535567#766 -503535567#767 -503535567#768 -503535567#769 -503535567#770 -503535567#771 -503535567#772 -503535567#773 -503535567#774 -503535567#775 -503
```

表 3.4 连续五个工作日早高峰期的出行需求配置

总需求 [veh/h]	7:00-7:30	7:30-8:00	8:00-8:30	8:30-9:00
周一	$D_1 \sim \mathcal{N}(2700, 108^2)$	$D_1 \sim \mathcal{N}(5000, 200^2)$	$D_1 \sim \mathcal{N}(3800, 152^2)$	$D_1 \sim \mathcal{N}(2500, 100^2)$
周二	$D_2 \sim \mathcal{N}(3300, 132^2)$	$D_2 \sim \mathcal{N}(4400, 176^2)$	$D_2 \sim \mathcal{N}(4000, 160^2)$	$D_2 \sim \mathcal{N}(3300, 132^2)$
周三	$D_3 \sim \mathcal{N}(3000, 120^2)$	$D_3 \sim \mathcal{N}(4800, 192^2)$	$D_3 \sim \mathcal{N}(3200, 128^2)$	$D_3 \sim \mathcal{N}(3000, 120^2)$
周四	$D_4 \sim \mathcal{N}(2800, 112^2)$	$D_4 \sim \mathcal{N}(4200, 168^2)$	$D_4 \sim \mathcal{N}(3500, 140^2)$	$D_4 \sim \mathcal{N}(3500, 140^2)$
周五	$D_5 \sim \mathcal{N}(1500, 60^2)$	$D_5 \sim \mathcal{N}(4000, 160^2)$	$D_5 \sim \mathcal{N}(4900, 196^2)$	$D_5 \sim \mathcal{N}(3600, 144^2)$

件的情况下，所提出的方法可能不再适用，因为这些事件对出行选择的影响没有被智能体所体验和学习。

3.4 本章小结

在本章中，对仿真实验场景的设计与构建进行了详细讨论。在3.1节中分析了城市交通仿真平台的可行性，通过对比多种仿真平台，最终选择了基于SUMO的城市交通仿真平台。在3.2中介绍了该平台的设计目标是为研究城市交通问题提供一个高度可定制、易于操作的实验环境。同时介绍了平台的功能模块，包括路网编辑与生成、出行模式设计以及流量生成。在3.3节中，通过实验场景的选择与搭建，确保了实验场景能够满足不同研究需求。在路网编辑与生成部分，讨论了如何创建和优化路网结构，以便更好地模拟实际城市交通。出行模式的设计部分重点关注了如何根据不同的交通需求和策略进行出行模式安排。最后，在流量生成部分，介绍了如何根据实际数据生成合理的交通流量，以便在仿真环境中进行准确的交通分析。

第四章 基于深度强化学习的出行模式与时间选择方法

强化学习是一种通过迭代地改进策略来最大化累计奖励或回报的机器学习方法。在应用强化学习到具有马尔可夫属性的序列决策过程中，需要先构建一个马尔可夫决策过程，该过程定义了环境的演变，考虑到强化学习代理所采取的行动。强化学习代理通过行动探索和开发不断地与环境互动，根据当前状态 s_t 进行行动。每次行动会使环境演变成一个新状态 s_{t+1} ，并获得相应的奖励 r_t ，反馈给代理以改善其决策逻辑。这个过程一直迭代，直到代理成功学习到一个能够最大化累计奖励的策略 π ，也就是一个决策者。因此，强化学习的关键在于根据奖励不断迭代改进策略。

在本研究中，将每个出行者视为具有学习能力的智能实体，通过马尔可夫决策过程来建模每个出行者跨越多个连续日的交通出行行为。每个出行者能够选择的行动包括不同组合的出行方式和出发时间。最终由个人采取的行动 a_t 决定了环境演变到的下一个状态 s_{t+1} 。该状态应反映个人关于行程本身以及相关环境的最新知识。选择此行动所获得的奖励 r_t （与旅行成本相关）有助于改善个人的决策逻辑，这样个人就能逐渐学习到最优的行动策略，并最大化累计奖励。图4-1展示了在出行模式与时间选择问题背景下的强化学习中智能体与环境的交互示意图。

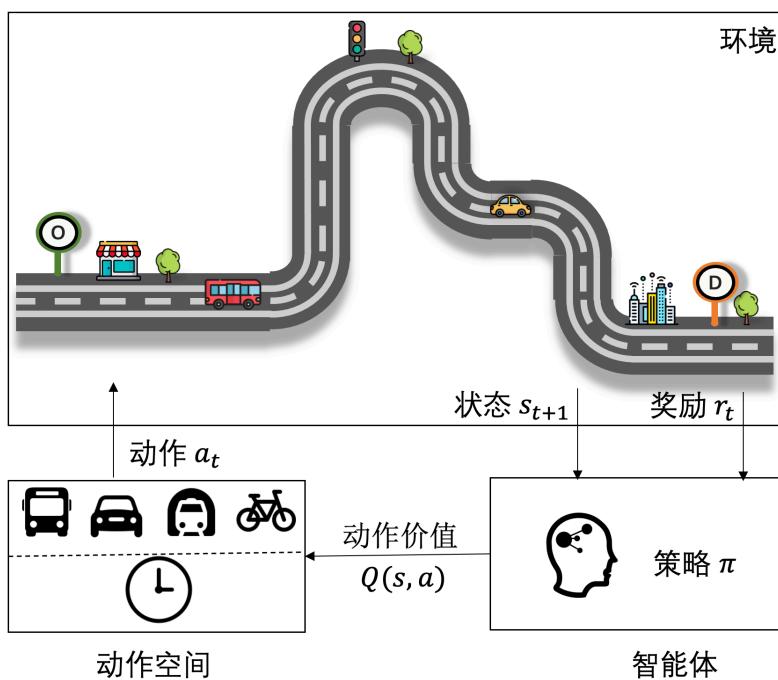


图 4-1 基于强化学习的出行模式与时间选择示意图

本章主要建立基于深度强化学习的出行模式与时间选择模型。首先，将使用马尔可夫决策过程框架来建模出行决策过程，包括定义状态空间、动作空间和奖励函数。然后，

将介绍基于深度神经网络算法的出行模式和出发时间选择算法，并讨论如何设计合适的神经网络结构、选择超参数以及对模型进行优化和训练。最后，通过训练结果评估模型的性能。

4.1 马尔可夫决策过程框架

马尔可夫决策决策过程是建模和优化出行模式和时间选择的先决条件。从数学角度来说，它是一个五元组 (S, A, P, R, γ) ，其中 S 表示状态空间， A 表示动作空间， P 表示状态转移概率， R 表示奖励函数， γ 为折扣因子。接下来，将进一步阐述如何构建和解决本文研究的特定问题下的马尔可夫决策决策过程。

4.1.1 动作空间

动作空间是强化学习中的一个关键概念。在建模动作空间时需考虑出行者的个性化特征。例如，不同出行者对于出行方式和出发时间的偏好不同，因此他们的动作空间也会不同。对此，可以引入个性化因素对动作空间进行建模。例如，可以考虑出行者的年龄、性别、职业、家庭状况等因素，进一步细化动作空间的描述，提高模型的预测能力和适应性。此外，动作空间的大小和粒度也会影响到模型的性能和可解释性。如果动作空间过大，模型的训练和预测会变得非常困难，同时也会增加模型的计算复杂度和存储空间需求。而如果动作空间过小，模型的表达能力就会受到限制，无法对真实情况进行有效建模。因此，需要在合理范围内对动作空间进行定义和限制，以平衡模型的性能和可解释性。在实际应用中，建模动作空间的过程也需要考虑到数据的可用性和质量。

在出行模式与时间选择中，动作空间包括出行方式和出发时间两个方面。在选择出发时间时，出行者需要考虑到交通拥堵、出行时间和其他因素对行程的影响。例如，在高峰期出发可能需要更长的旅行时间，而在非高峰期出发可能可以更快地到达目的地。因此，在建模动作空间时，需要综合考虑各种因素，以便在代理决策时提供准确的信息。

其中，出行模式的选择包含私家车、公共交通或自行车。在公共交通中，可以选择乘坐公交车或地铁，以及在公交地铁间的换乘，但是不考虑三种交通方式之间的换乘。这是因为在与多模式换乘中涉及到很多变量，比如停车地点、换乘时间等，考虑到乘客在选择时的不会多次更换交通模式的实际决策情况以及研究的复杂程度，本研究将不考虑不同模式之间的多次换乘行为。

对于出发时间，每个出行者都有一个初始或期望出发时间 t_0 。实际上，在早高峰出行时，出行者往往会考虑调整出发时间，以避免交通拥堵和延误。这种现象被称为“出行时间弹性”(travel time elasticity)，指的是出行者在面对交通拥堵或不确定性时，可以调整出行时间以获得更好的出行体验或更高的效率。根据交通经济学的研究，出行时间弹性的大小受到许多因素的影响，包括个人时间成本、出行目的、所处的交通环境以及出行者的偏好等。因此，在出行模式和时间选择算法中，需要考虑出行时间弹性，以更

准确地推荐出行模式和出发时间。在本文的出行模式与时间选择场景中，出发时间可以在一个有限的时间窗口 $[t_{\min}, t_{\max}]$ 内进行调整。这个时间窗口是由最早和最晚的出发时间 t_{\min} 和 t_{\max} 确定的。出发时间的调整是以离散间隔为单位进行的，而不是连续方式进行。在本文的出行模式与时间选择场景中，离散间隔的设置是为了将出发时间的选择问题转化为一个有限的状态空间的问题，从而可以应用深度强化学习算法进行求解。如果使用连续方式进行出发时间的调整，状态空间将是无限的，这将导致算法难以收敛并且计算复杂度较高。因此，将出发时间的调整以离散间隔为单位进行，可以有效地简化问题，并且使算法更易于实现和计算。

对于上述所提到的交通方式和出发时间，需要进行适当的编码以便于智能体在模型中进行操作。在本文中，采用离散化编码的方式，将交通方式和出发时间分别离散化为一组离散的选项。例如，对于交通方式，可以将私家车、公交车、地铁和自行车分别编码为 m_1 、 m_2 、 m_3 和 m_4 。对于出发时间，可以将 t_0 和时间窗口 $[t_{\min}, t_{\max}]$ 离散化为一组时间步长，例如每 5 分钟一步。这样，代理可以从一组离散的选项中进行选择，并决定最佳的出行方式和出发时间。动作空间的描述采用了向量的形式， \mathbf{a} 包括交通方式 \tilde{m} 和出发时间 \tilde{t} 。交通方式可以是可用交通方式 m_1, m_2, \dots, m_N 中的任意一种，而出发时间必须在时间窗口 $[t_{\min}, t_{\max}]$ 内。因此，动作空间可以表示为：

$$\mathbf{a} = \begin{bmatrix} \tilde{m} \\ \tilde{t} \end{bmatrix} = \begin{bmatrix} \text{出行模式} \\ \text{出发时间} \end{bmatrix} \in \begin{bmatrix} \{m_1, m_2, \dots, m_N\} \\ [t_{\min}, t_{\max}] \end{bmatrix} \quad (4.1)$$

4.1.2 状态空间

状态空间是强化学习中非常重要的一个概念，它定义了智能体在决策时需要考虑的连续环境。在设计状态空间时，需要仔细考虑应用场景和问题的特征，以确保状态空间中包含的信息能够有效地指导代理的决策。同时，也需要注意状态空间的维度和大小，以便使智能体能够有效地处理状态，并且可以在有限的时间内完成状态的学习和更新。对于出行模式与时间选择的问题，状态空间被设计为不仅需要包含有关行程的最新信息，还包括智能体早期经验。这种状态空间设计在很大程度上类似于理性人类的决策机制，即从经验中学习。由于本研究的重点不在于经验选择建模而在于选择指导或推荐，因此将假设智能体能够充分感知环境，因此掌握行程的全部信息。

行程信息首先包括每种交通方式的旅行距离 L 和记忆旅行时间 \bar{T} 。前者是模式 m 的旅行距离，而后者是该模式 m 的平均经验旅行时间，这样可以利用经验和在交通随机性存在的情况下保持稳健性。作为行程信息的另外两个变量是初始出发时间 t_0 和出

发时间差或偏移量 Δt 。将上述所有内容组合起来，得行程信息的状态向量 s_{trip}^m ：

$$s_{\text{trip}}^m = \begin{bmatrix} L_m \\ \bar{T}_m \\ t_0 \\ \Delta t \end{bmatrix} = \begin{bmatrix} \text{出行距离} \\ \text{记忆行程时间} \\ \text{初始出发时间} \\ \text{出发时间差} \end{bmatrix} \quad (4.2)$$

环境信息基本上包括有助于不同交通方式旅行成本的因素。对于公共交通，考虑到两个因素，即可达性和票价。在这里，我们将可达性 p 定义为完成行程的第一和最后一段所需的总步行距离：

$$p = d_{\text{origin}} + d_{\text{destination}} \quad (4.3)$$

其中 d_{origin} 和 $d_{\text{destination}}$ 分别是从最近的公交车站或地铁站到起点和终点的步行距离。公共交通票价是必须支付的使用该服务的货币成本。对于私家车，我们考虑燃油价格作为影响因素，并将其放在状态中。因此，特定于环境信息的状态向量如下所示：

$$s_{\text{environment}} = \begin{bmatrix} p \\ f \\ o \end{bmatrix} = \begin{bmatrix} \text{可达性} \\ \text{票价} \\ \text{燃油价格} \end{bmatrix} \quad (4.4)$$

为了将用户的时间价值纳入状态空间，将再添加一个状态向量来代表用户的偏好，并使用这个变量来调整不同状态下关联的时间价值。例如，对于高收入用户而言建议出行时间较长的状态造成更高的时间价值，因为这些用户可能有能力支付更昂贵但更快的交通方式。另一方面，对于低收入用户，时间价值会更低。基于此，我们提出的深度强化学习算法将学习考虑用户的收入水平来推荐旅行方式，并在用户特定的需求和偏好上权衡旅行时间和成本，从而做出适当的建议。

$$s_{\text{VOT}} = [\alpha] = [\text{时间价值}] \quad (4.5)$$

将式4.2、式4.4与式4.5相结合，可以得到在完全信息假设下的完整状态空间。

实际人类在做选择时，通常无法获取所有的状态信息。因此，在模拟人类的选择行为，需要建立部分信息的状态空间向量。通过将某些状态变量排除在外，可以简化问题并减少计算复杂度。然而，这也会影响算法的性能和准确性，因为所选取的状态变量可能无法完全描述真实的状态。

实际上，在进行决策和选择时，人类通常无法获取所有的状态信息。这种现象在现实生活中尤为明显，因为面临许多复杂且不确定的情况。因此，在模拟人类的决策行为时，需要建立一个基于部分信息的状态空间向量，在众多可能的状态变量中挑选一部分来构建模型。通过将某些状态变量排除在外，可以简化问题并降低计算复杂度，从而加

快求解速度和提高算法的效率。为了平衡问题的简化程度与保持准确性之间的关系，研究人员需要在选择状态变量时做出权衡。这通常涉及到对问题的深入理解和多次尝试。在实际应用中，可以通过逐步增加或减少状态变量的数量来调整算法的性能和准确性，以找到最适合特定问题的状态空间表示。然而，这种简化往往需要在算法的性能和准确性方面付出一定代价。由于所选取的状态变量可能无法充分描述真实的状态，这可能导致算法在某些情况下无法做出最佳决策。

$$\mathbf{s}_{\text{reduced}} = \begin{bmatrix} \bar{T} \\ t_0 \\ \Delta t \end{bmatrix} = \begin{bmatrix} \text{记忆行程时间} \\ \text{初始出发时间} \\ \text{出发时间差} \end{bmatrix} \quad (4.6)$$

4.1.3 奖励函数

在强化学习中，智能体的目标是通过最大化长期奖励来学习最优的决策规则。通过奖励函数，智能体可以计算每个动作对于实现这个目标的预期收益。在本文的研究中，奖励函数可以参考旅行效用函数，旨在最小化总旅行费用。智能体将根据预期的长期奖励来选择动作，以便在未来的交互中最大化收益。在第 i 步获得的奖励计算公式如下：

$$r_i = \frac{E_1 - C_m^i}{E_2}, \quad (4.7)$$

其中， C_m^i 是交通方式 m 的总旅行费用， E_1 和 E_2 是映射和缩放成本到奖励的两个常数。总旅行费用又可以分解为三个部分，即总旅行时间 T_m^i 、行程延误 $\delta(t^i)$ 和其他与旅行相关的成本 F_m^i ：

$$C_m^i = \alpha T_m^i + \delta(t^i) + F_m^i, \quad (4.8)$$

其中， α 是时间的价值。这个公式描述了在一个旅行过程中，成本是如何被划分的。智能体可以通过调整其动作来优化它所接收到的奖励，并在行程中实现更好的效用。

只考虑总旅行时间的问题往往会忽略实际到达时间。也就是说，尽管总的旅行时间很短，但到达时间可能与理想的时间相差甚远。因此，在成本计算中引入了计划延迟惩罚^[26]。假设每个人都只有一个期望的到达时间，早到和晚到都会产生一个所谓的日程延误成本。当实际到达时间偏离期望时间时，计划延迟成本会以线性方式增长。从数学上讲，它表示为：

$$\delta(t^i) = \begin{cases} \beta(t + T_m - t_d) & \text{if } t + T_m - t_d < 0, \\ 0 & \text{if } t + T_m - t_d = 0, \\ \gamma(t_d - t - T_m) & \text{if } t + T_m - t_d > 0, \end{cases} \quad (4.9)$$

其中, β 和 γ 分别是早到和晚到的行程延误时间价值, t_d 是期望到达时间。通过考虑行程延误成本, 可以更准确地评估各种出行方式的效用。

除此之外, 其他的与出行相关的费用主要是指汽车的燃料费用和公共交通的票价。对于私家车来说, 燃料费用是与行驶距离成正比的, 而对于公共交通来说, 则是根据具体的交通工具的票价。因此, 其他出行相关费用可以表示为公式:

$$F_m^i = \begin{cases} L_{\text{car}} \cdot o & \text{if } m = \text{私家车}, \\ f & \text{if } m = \text{公共交通}, \\ 0 & \text{if } m = \text{自行车} \end{cases} \quad (4.10)$$

其中, f 可以表示为:

$$f = \mathbb{I}(\text{bus}) \cdot f_{\text{bus}} + \mathbb{I}(\text{subway}) \cdot f_{\text{subway}}, \quad (4.11)$$

这里的 $\mathbb{I}(\cdot)$ 是一个指示函数, 当选择的交通工具是公共汽车或地铁时, 它返回 1, 否则返回 0。公交车费用 f_{bus} 是固定的, 而地铁费用 f_{subway} 则随着行驶距离的增加而增加。这些费用可以用于计算奖励函数中的其他出行相关成本项。

4.2 基于深度 Q 网络算法的模式与出发时间选择算法

本文采用无模型基于动作的深度强化学习方法深度 Q 网络作为基础算法解决出行模式与时间选择问题。它学习与 MDP 相关的状态动作值, 基于此隐含地推导出最优策略, 即始终选择导致最大状态动作值的动作。根据式2.5与式2.6可知, 它使用神经网络来近似最优状态动作值函数, 从而将传统的 Q-learning 应用于高维或连续空间问题。

为了设计一个优秀的 DQN 算法, 本节需要对神经网络结构进行设计, 并对超参数进行选择和标定, 然后对模型进行优化。神经网络结构设计应考虑输入和输出的特征, 并充分考虑网络的深度和宽度。超参数包括学习率、批量大小、折扣因子和经验回放的容量。优化算法包括随机梯度下降、Adam 和 RMSProp 等。通过选择适当的超参数和优化算法, 可以提高 DQN 算法的收敛速度和性能。

4.2.1 神经网络结构设计

神经网络在强化学习中扮演着重要的角色, 它可以作为函数逼近器来估计 Q 值函数, 从而得到最优的策略。在 DQN 算法中, 神经网络被用来近似状态-动作值函数, 因此它的结构对算法的性能有着很大的影响。在 DQN 算法中的神经网络结构设计可以通过以下步骤进行:

1. 确定输入和输出层的维度。神经网络的输入层节点通常接收标量值, 每个输入节点代表输入数据中的一个特征, 输出层的输出数据同理。当有多个特征时, 输入层会有

多个节点，每个节点对应一个特征。在本文的研究中，输入层的维度应该与状态向量的维度相同，而输出层的维度应该等于动作的数量。

2. 选择合适的隐藏层数和节点数。神经网络的隐藏层节点数量和层数决定了网络的复杂性和表示能力。一个具有更多节点和层数的网络可能具有更强大的表示能力，但也可能导致过拟合，特别是在训练数据有限的情况下。相反，一个具有较少节点和层数的网络可能会降低过拟合的风险，但也可能导致模型的表达能力不足，从而导致欠拟合。在本研究的神经网络设计中，选择 32、64、64 这些值作为隐藏层的节点数量，这些值在实践中能够提供一个合理的平衡，既不会导致过拟合，也不会导致欠拟合。

3. 选择适当的激活函数。激活函数对神经网络的性能有着重要的影响。在本文的深度 Q 网络中，神经网络使用 ReLU 作为激活函数（参考式2.16）。ReLU 函数的优点是计算简单，同时能够引入非线性特性。在许多深度学习任务中，ReLU 激活函数表现良好，因此在 DQN 中也是一个常用的选择。

4. 选择适当的优化器和损失函数。在训练过程中，将使用经验回放的技术来解决深度 Q 网络算法中的样本问题。通过将每个状态-动作转换存储在一个固定大小的缓存器中，使模型可以从以前的经验中学习。这种方法允许从整个经验集合中随机抽取样本进行训练，以减少梯度下降的样本相关性，并增加学习的稳定性。在优化深度 Q 网络算法时，使用均方误差损失函数 (参考式2.18) 作为优化目标，其中 y_i 是目标 Q 值，可以通过式2.6计算得到。最后，使用优化器（如 Adam）来调整神经网络的权重，以最小化损失函数。

图4-2是本文深度 Q 网络中的神经网络结构示意图。在代码实现中，使用 PyTorch 框架来定义神经网络。在 DQN 类的初始化函数中，定义了前馈神经网络的结构。它包括四个线性层，分别包含 32、64、64 和 n 个节点，其中 n 是动作的数量。这个结构相对简单，在实际应用中已经被证明可以取得不错的效果。

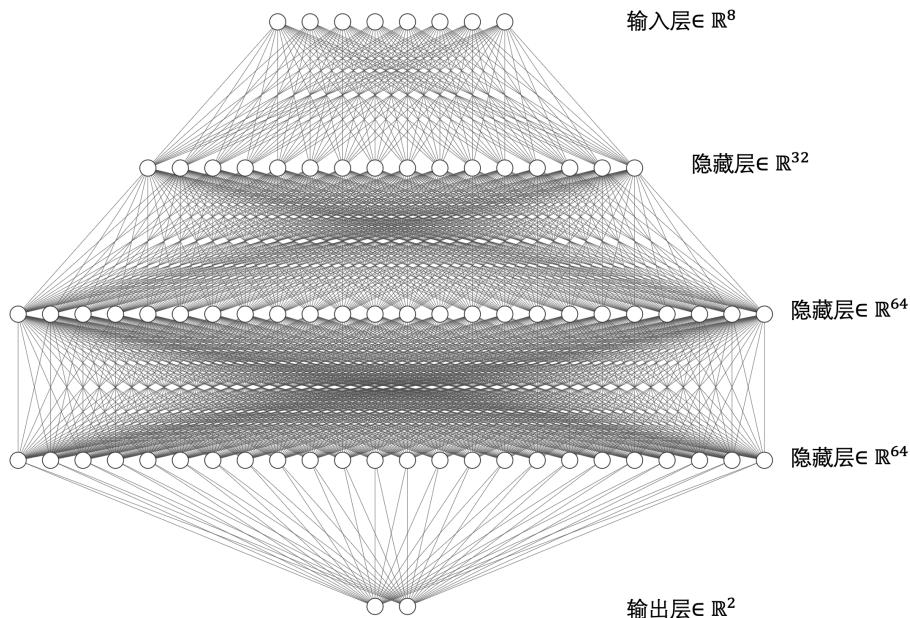


图 4-2 神经网络结构示意图

4.2.2 超参数的选择

超参数是在训练过程中固定不变的参数，它们对模型的性能和训练速度有很大影响。合适的超参数选择对于训练一个高效且鲁棒的模型至关重要。优化器的学习率等超参数的选择和标定也会对 DQN 算法的性能有很大的影响。因此，需要进行仔细的超参数调优，以找到最优的超参数组合，以提高 DQN 算法的收敛速度和性能。可以将不同超参数在深度 Q 网络中的作用将其分为四大类：学习与优化相关参数、经验回放相关参数、神经网络结构与同步相关参数和探索与利用相关参数。

学习与优化相关参数主要负责控制神经网络的学习过程，包括奖励折扣、学习速率和每次权重更新的样本数量。这些参数可以影响模型收敛速度和学到的策略质量。为了限制模型训练的时间和资源消耗，通常也会设置最大训练轮数，根据实际需求和计算资源，可以调整这个参数来平衡训练时间和模型性能。学习与优化相关参数包括 GAMMA, BATCH_SIZE, LEARNING_RATE 和 MAX_EPISODE。

GAMMA: 折扣因子用于调整未来奖励的重要性。较高的值表示更关注未来的回报，而较低的值表示更关注短期回报。在本实验中，选择了 0.99 的折扣因子，以平衡短期和长期回报的关注度。

BATCH_SIZE: 每个训练步骤中使用的样本数。较大的批量大小可能会导致梯度更新更稳定，但同时也会增加计算成本。选择了 5 作为批量大小，以在计算效率和梯度稳定性之间取得平衡。

LEARNING_RATE: 模型在训练过程中更新权重的速度。较大的学习速率可能会导致模型收敛得更快，但也可能导致不稳定的训练过程。较小的学习速率可能会使训练过程更稳定，但需要更长的时间才能收敛。选择了 1e-4 作为学习速率，以在收敛速度和训

练稳定性之间取得平衡。

MAX_EPISODE: 最大的训练轮数，用于限制训练时间。在实际应用中，训练时间可能受到硬件和计算资源的限制。我们选择了 800 作为最大训练轮数，以在保证模型性能和限制训练时间之间达到平衡。

经验回放是深度 Q 网络算法的一个关键组成部分，可以有效地降低数据相关性、提高样本利用率，从而加速学习过程。这类参数主要调整回放缓冲区的容量和在开始训练前填充缓冲区的样本数量，以确保训练过程中有足够的经验样本可以使用。经验回放相关参数包括 REPLAY_SIZE 和 REPLAY_START_SIZE。

REPLAY_SIZE: 回放缓冲区的最大容量。回放缓冲区用于存储经验样本，以便在训练过程中进行随机抽样。较大的回放缓冲区可以提高样本的多样性，但也会增加内存计算压力。经过多次实验测试后，最终选择了 40 作为回放缓冲区大小。

REPLAY_START_SIZE: 开始训练前等待填充回放缓冲区的帧数。这个参数确保在训练开始时，回放缓冲区已经存储了足够的样本。选择 40 作为开始训练前填充回放缓冲区的帧数，以确保有足够的样本用于训练。

深度 Q 网络算法使用了两个相同结构的神经网络，一个用于训练，另一个作为目标网络计算目标值。这类参数主要负责控制训练网络和目标网络之间权重同步的频率。同步过程可以保证目标网络的稳定性，从而提高训练的稳定性。神经网络结构与同步相关参数包括 SYNC_TARGET_FRAMES。

SYNC_TARGET_FRAMES: 将模型权重从训练模型同步到目标模型的频率。定期将训练模型的权重同步到目标模型有助于提高训练过程的稳定性。经过多次实验测试后，选择了 5 作为同步频率，以在训练速度和稳定性之间达到平衡。

深度 Q 网络算法通过 ϵ -greedy 策略平衡探索（尝试新动作）和利用（采用当前最优策略）。这类参数主要用于调整探索程度，包括 ϵ 值的初始值、最终值以及从初始值到最终值所需的帧数。合适的探索与利用平衡有助于模型找到全局最优策略。探索与利用相关参数包括 EPSILON_START、EPSILON_FINAL 和 EPSILON_DECAY_LAST_FRAME。

EPSILON_START: ϵ -greedy 策略中的初始 ϵ 值。较高的初始 ϵ 值意味着在训练初期，智能体更倾向于探索环境而非利用已知的策略。选择 1.0 作为初始 ϵ 值，以鼓励智能体在训练初期进行充分的探索。

EPSILON_FINAL: ϵ -greedy 策略中的最终 ϵ 值。较低的最终 ϵ 值表示在训练后期，智能体更倾向于利用已知的策略而非进行探索。选择 0.01 作为最终 ϵ 值，以在训练后期降低探索频率并优化已知策略。

EPSILON_DECAY_LAST_FRAME: ϵ 值从初始值到最终值所需的帧数。较快的衰减速率可能会导致智能体在训练初期就过于关注已知策略，而较慢的衰减速率可能会导致智能体在训练过程中持续进行过多的探索，最终选择 800 作为衰减帧数。

表4.1列出了经过测试与参考后确定的在深度 Q 网络中使用的超参数及其取值。

表 4.1 参数的取值与描述

参数	取值	描述
GAMMA	0.99	未来奖励的折扣因子
BATCH_SIZE	5	每个训练步骤中使用的样本数
REPLAY_SIZE	40	回放缓冲区的最大容量
LEARNING_RATE	10^{-4}	模型在训练过程中更新权重的速度
SYNC_TARGET_FRAMES	5	将模型权重从训练模型同步到目标模型的频率
REPLAY_START_SIZE	40	开始训练前等待填充回放缓冲区的帧数
EPSILON_START	1.0	ϵ -greedy 策略中的初始 ϵ 值
EPSILON_FINAL	0.01	ϵ -greedy 策略中的最终 ϵ 值
EPSILON_DECAY_LAST_FRAME	600	ϵ 值从初始值到最终值所需的帧数
MAX_EPISODE	800	最大的训练轮数，用于限制训练时间

4.2.3 模型的优化

为了提高本研究中深度强化学习模型的性能，引入经验回放的技术。经验回放是一种重要的优化技术，它解决了 DQN 算法中的样本相关性问题。在传统的在线学习中，神经网络每次只更新一个样本的参数，因此相邻样本的训练数据之间存在强相关性，容易导致模型的过拟合。经验回放的核心思想是将所有样本存储到经验池中，并从中随机采样一批样本进行训练。这样可以打破样本之间的相关性，减少过拟合的风险，提高模型的泛化能力。

图4-3展示了经验回放缓冲区和目标网络的深度 Q 网络的工作流程。图中描绘了状态输入到主网络和目标网络的过程，以及如何将经验元组存储到经验回放缓冲区中。同时，从缓冲区中随机抽样的经验被用于计算损失并更新主网络。此外，通过某种策略（例如，每隔固定步数）更新目标网络，从而提高模型的稳定性。

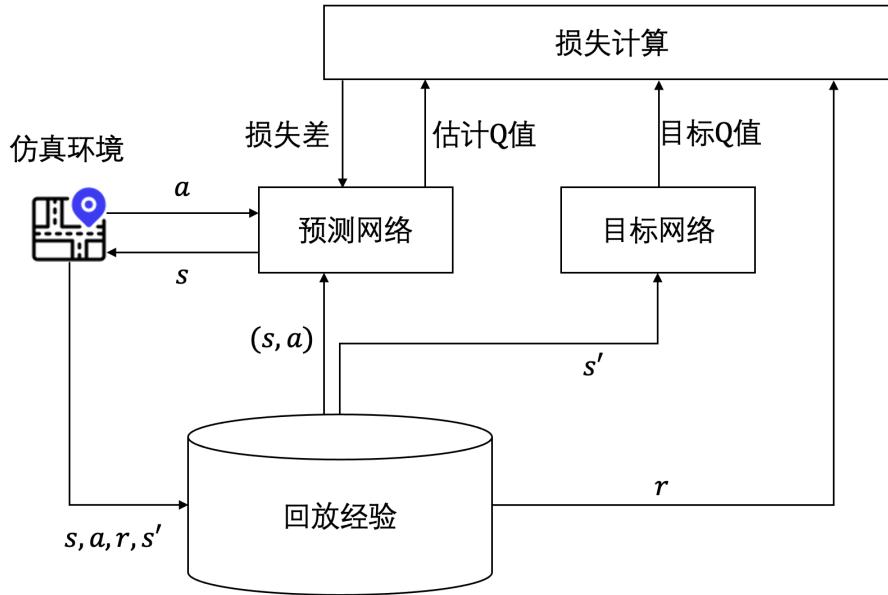


图 4-3 经验回放缓冲区的工作流程

其中，目标网络是一种防止算法中目标值剧烈变动的技术。在传统的强化学习算法中，目标值是使用当前网络计算得出的，因此目标值随着网络参数的更新而不断变化，容易导致模型不稳定。为了解决这个问题，目标网络的主要思想是使用一个独立的神经网络来计算目标值，该网络的参数较为稳定，不随着训练过程中的更新而变化。目标网络的参数定期地从当前网络中复制而来，以减缓目标值的变化速度，提高模型的稳定性。

本文的算法使用均匀随机抽样从缓冲区中选取经验元组，主要原因是打破数据间的时间相关性。在强化学习中，智能体与环境交互产生的数据序列通常具有很强的时间相关性。通过从缓冲区中均匀随机抽样，可以减小训练数据的相关性，从而减少训练过程中的不稳定性，提高模型的泛化能力和学习效率。

另一个原因是简单性和计算效率。均匀随机抽样易于实现，计算成本较低，同时可以平衡各个经验元组在训练中的使用频率。尽管优先级经验回放可能在某些情况下带来更好的性能，但其实现更复杂，涉及优先级计算、更新以及在抽样过程中的加权抽样等。因此，均匀随机抽样在许多应用中仍然是一种实用且有效的方法。在训练数据存在很大差异、关键经验对学习过程更为重要、模型收敛速度较慢或面临稀疏奖励任务时，可以考虑将均匀随机抽样改为优先级抽样。

算法4.1是本文建立的基于深度强化学习的联合出行模式与出发时间选择建模和训练方法。算法首先初始化经验回放缓冲区和策略网络参数，然后在多个训练回合中进行迭代。在每个回合中，智能体观察初始状态，然后根据不同的时间步长确定最佳的出行模式与出发时间。接着，智能体计算总出行成本并获得相应的奖励。将观测到的经验元组存储到经验回放缓冲区，以便进行网络参数更新。最后，智能体从缓冲区中随机抽取一批样本进行学习，并定期更新目标网络参数，以提高模型的稳定性和学习效率。

算法 4.1 基于深度强化学习的联合出行模式与出发时间选择建模和训练方法

初始化经验回放缓冲区 D , 策略网络参数 \mathbf{w} 和 \mathbf{v}

for $episode = 1$ to M **do**

 从环境中观测初始状态 s_0

for $t = 1$ to T **do**

 使用式2.22确定时间步长为 t 时的出行模式与出发时间

 根据4.1.3小节计算相应的总出行成本, 使用式4.7获得奖励 r_t

 记录出行和环境信息, 建立下一个状态 s_{t+1}

 将元组 $(\mathbf{s}_t, \mathbf{a}_t, r_t, \mathbf{s}_{t+1})$ 存储到经验回放缓冲区 D

 从 D 中随机抽取小批量样本

$$Q \leftarrow r_t + \gamma \underset{\mathbf{a} \in A}{\operatorname{argmax}} \hat{q}(\mathbf{s}_{t+1}, \mathbf{a}, \mathbf{v})$$

$$\hat{Q} \leftarrow \hat{q}(\mathbf{s}_t, \mathbf{a}, \mathbf{w}_t)$$

$$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \alpha \nabla \frac{1}{2}(Q - \hat{Q})^2$$

 每隔固定数量 c 个时间步, 设置 $\mathbf{v} \leftarrow \mathbf{w}_t$

end for

end for

4.3 基于聚类的深度强化学习方法

在前一节中, 对深度 Q 网络算法作为解决出行模式和时间选择问题的基本解决方案进行了详细阐述。然而, 将该算法扩展为解决具有大量个体的类似问题仍然是一个挑战。正如先前讨论过的, 将所有个体的经验存储并用于训练是计算上不可行且低效的, 而随机选择一个或几个个体则不够准确且不可靠。因此, 本节中提出了一种有效的方法来选取具有代表性的个体以进行高效的模型训练, 即基于个体的出行特征进行聚类。对于处于同一聚类中的个体, 可以认为它们的出行特征相似。因此, 它们中的每一个都可以被视为该聚类的代表, 其经验可以代表其余个体来训练深度 Q 网络。通过这种方式, 不仅避免了部署与个体数量相同数量的代理, 还有效地利用代表性个体的经验进行充分的模型训练。实际上, 采用这种方法可以有效地解决具有许多个体的出行模式与时间选择问题, 而在决策制定过程中不会牺牲太多的最优性。

4.3.1 DBSCAN 聚类方法

密度基于空间聚类应用 (Density-Based Spatial Clustering of Applications with Noise, 简称 DBSCAN) 是一种基于密度的聚类方法, 旨在识别高密度区域并将其作为簇的核心。DBSCAN 的一个显著优势是无需预先确定簇的数量, 因此算法本身是非参数化的。这种算法可以自适应地调整簇的大小和形状, 即使在数据中存在噪声的情况下, 它的性能仍然可靠。

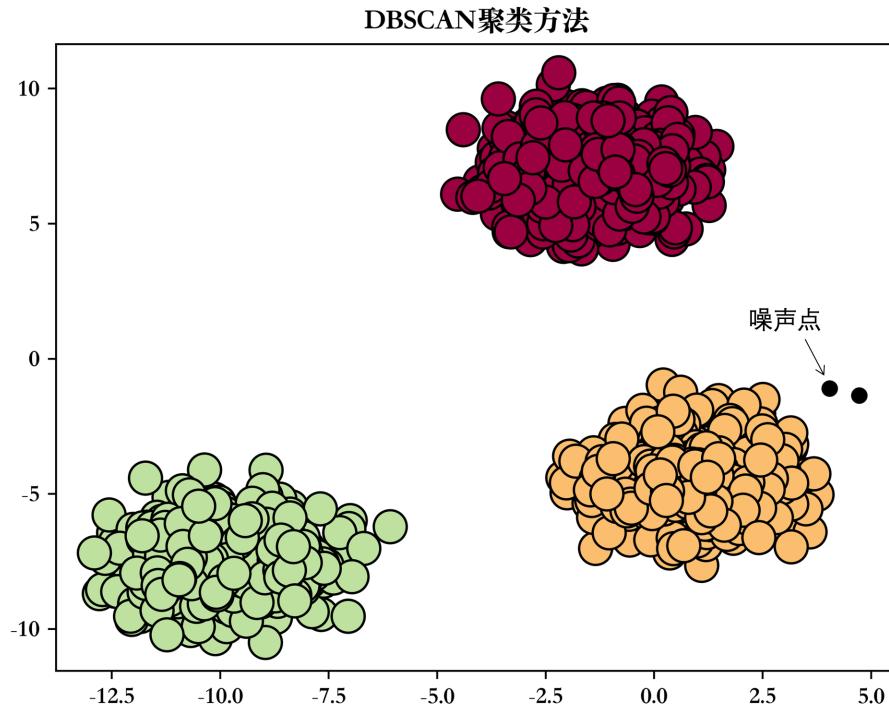


图 4-4 DBSCAN 聚类方法聚类效果图

DBSCAN 算法的核心概念是将数据点划分为三类：核心点、边界点和噪声点。核心点是一个密度可达的点集，即其周围的密度大于等于指定的阈值。边界点是密度可达的点，但其周围的密度低于指定阈值。噪声点则是那些不属于任何簇的点，它们周围的密度也低于指定阈值。除了 DBSCAN 算法，还有其他一些聚类方法，例如 k-means 和层次聚类。k-means 算法是一种广泛使用的聚类方法，通过将数据集划分为 k 个簇来实现聚类。其核心思想是将数据点分配到距离最近的质心（簇的中心），然后将质心更新为簇中所有点的平均值。接着，这一过程会不断重复，直到质心不再发生变化或达到预定的最大迭代次数为止。图 4-4 展示了使用 DBSCAN 算法的聚类效果图。

层次聚类则是一种自下而上的聚类方法，通过递归地将最相似的数据点组合成更大的簇，最终形成一个完整的聚类树。层次聚类可以是聚合的（自底向上）或分裂的（自顶向下）。在聚合层次聚类中，每个数据点起初都是一个单独的簇，然后将最相似的簇合并在一起，直到所有数据点都被分配到一个簇中。在分裂层次聚类中，开始时将所有数据点都分配到一个大簇中，然后逐步将其分裂成较小的簇，直到达到预定的聚类数目。

尽管 k-means 和层次聚类方法也可以用于选择代表性个体，但它们对噪声点的处理能力不如 DBSCAN，可能会将噪声点误分类为一个簇或将它们分配到多个簇中。此外，DBSCAN 算法可以自动确定簇的数量，并且不需要提前指定 k 值或层次聚类的高度。

在这项研究中，DBSCAN 算法被选作选择代表性个体的方法，因为它在处理数据噪声和密度不均匀的情况下表现出色，并且无需预先设定簇的数量。通过聚类选择代表性个体，可以显著降低强化学习算法中状态和动作空间的规模，从而提高训练效率。

总的来说，DBSCAN 作为一种基于密度的聚类方法，具有很多优点。首先，它不需

要预先确定簇的数量，因此更具自适应性。其次，它可以在数据中存在噪声的情况下保持较好的性能。此外，DBSCAN 算法能够自动调整簇的大小和形状，使其适应不同密度的数据分布。

4.3.2 聚类参数的选择

在研究出行特征个体数据集 D 中，为了对出行数据进行有效的分类和分析，选择两个关键特征作为聚类的输入，分别是出行距离和公共交通的可达性。这两个特征在很大程度上反映了出行者的出行需求和出行方式的选择。

首先，出行距离是一个重要的出行特征，因为它与出行者的出行成本和时间成本密切相关。在现实生活中，出行距离通常会影响出行者在时间和金钱上的投入。例如，较长的出行距离可能意味着更高的燃油消耗、更多的过路费以及更长的行驶时间。此外，出行距离还与环境污染有关，较长的距离可能会导致更多的温室气体排放。因此，在对出行特征进行聚类时，出行距离是一个重要的度量标准，有助于我们了解出行者在不同距离范围内的需求和行为特征。公共交通的可达性是另一个关键特征，因为它体现了出行者对公共交通设施的依赖程度和使用意愿。在城市规划和交通管理中，公共交通的可达性对于评估城市的可持续性和交通拥堵状况具有重要意义。具有较高公共交通可达性的区域通常会吸引更多的出行者使用公共交通工具，从而减轻道路拥堵和环境污染。选择这两个特征的原因还在于它们之间可能存在的相互关系。例如，当公共交通可达性较好时，出行者可能会倾向于选择公共交通工具进行较长距离的出行，而在公共交通可达性较差的情况下，他们可能会选择私人交通工具或其他出行方式。因此，通过对这两个特征进行聚类，可以揭示出行者在不同场景下的出行偏好和需求。此外，在实际应用中，出行距离和公共交通可达性还可以与其他出行特征相结合，以提供更全面的出行数据分析。例如，可以结合出行者的年龄、职业、家庭状况等社会经济特征，以深入了解不同人群在出行距离和公共交通可达性方面的需求和偏好。这将有助于城市规划者和交通管理者制定更有效的交通政策和措施，以满足不同人群的出行需求。

4.3.3 深度强化学习模型的改进

通过将深度强化学习方法与个体聚类和获取代表智能体的过程相结合，该集成算法旨在解决具有多个体的出行模式与时间选择问题。在这种方法中，所需训练的代理数量等于选定的代表智能体数量，从而降低了计算复杂度和训练时间。在整个训练过程中，这些代表智能体与它们各自的经验存储池同时进行训练。在训练的过程中，代表智能体学习并优化它们在出行场景中的决策。当这些代表智能体经过充分训练后联合起来，为不同的个体提供出行选择决策，从而提高整体系统的效率和性能。这种集成算法的优势在于一旦代表智能体训练完成，就无需重新进行聚类，因为它们已经具备了为聚类中各个个体做出决策的能力。在实际应用中，这些代表智能体将评估可用的出行选项，并选择具有最高奖励的行动来实施。这种方法有助于确保个体能够在各种出行场景中作出合

理且有效的决策，从而提高整体系统的性能。

在提出的方法中，目标是训练多个集群的代表性智能体，共同形成一个集成网络。为此，同时训练不同的网络模型，每个模型代表一个不同的集群集合。这些集合被训练来评估问题的不同方面，并根据它们自己的策略提供建议。在训练过程中，每个网络都会接触到相同的输入数据，并学会优化自己的推荐策略。训练过程是迭代的，每个网络根据从环境中获得的反馈更新其参数。通过同时训练所有网络，目标是提高集成网络的整体性能。一旦训练完成，新代理的决策将由所有网络同时评估最高奖励结果，根据它们自己训练过的策略来进行。代理的最终选择是根据所有网络评估获得最高奖励的那个。这种方法使得能够利用每个网络的优势，做出更准确和有效的建议。

算法4.2用于聚类个体并获取代表智能体。给定一组具有出行特征的个体数据集 D 、距离阈值 μ 和形成聚类所需的最小点数 m ，算法首先对数据进行归一化，然后遍历每个非聚类个体。对于每个个体 \mathbf{x} ，算法计算其在阈值 μ 内的邻居集 N 。如果邻居集的大小小于 m ，则将个体 \mathbf{x} 标记为噪声；否则，将其设置为新聚类 C 的核心点。接着，算法遍历 N 中的每个个体，并更新邻居集和聚类。最终，算法返回用于获取代表的个体群集。算法4.2实现了 DBSCAN 聚类过程，用于处理具有出行特征的个体数据集。通过这种方式，算法能够将个体划分为不同的聚类，同时识别并处理噪声点，从而在聚类中找到具有代表性的个体。

算法 4.2 聚类个体并获取代表智能体

输入: 所有个体的出行特征 $D = \{\mathbf{x}_1 = [L_1, p_1]^T, \mathbf{x}_2, \dots, \mathbf{x}_m\}$, 距离阈值 μ , 形成聚类所需的最小点数 m

输出: 代表智能体的集群

```

数据归一化:  $x_{\text{normalized}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$ 
for 每个非聚类个体  $\mathbf{x} \in D_{\text{normalized}}$  do
    将个体  $\mathbf{x}$  标记为归入某个群组
     $N \leftarrow GetNeighbors(\mathbf{x}, \mu)$ 
    if  $|N| < m$  then
        将个体  $\mathbf{x}$  标记为噪声
    else
        设置新聚类  $C \leftarrow \mathbf{x}$ 
        for 每个个体  $\mathbf{x}' \in N$  do
             $N \leftarrow N \setminus \mathbf{x}'$ 
            if 个体  $\mathbf{x}'$  是非聚类的 then
                将个体  $\mathbf{x}'$  标记为已聚类
                 $N' \leftarrow GetNeighbors(\mathbf{x}', \mu)$ 
                if  $|N'| \geq m$  then
                     $N \leftarrow N \cup N'$ 
                end if
                if  $\mathbf{x}'$  是噪声 then
                     $C \leftarrow C \cup \{\mathbf{x}'\}$ 
                end if
            end if
        end for
    end if
end for

```

4.4 本章小结

本章主要介绍了基于深度强化学习的出行模式与时间选择方法。首先，在4.1节中建立了马尔可夫决策过程框架，定义动作空间、状态空间和奖励函数。接下来，在4.2节中采用基于深度神经网络的算法，设计了神经网络结构，并通过选择合适的超参数和优化模型来改善出行模式与出发时间选择的准确性。为了进一步提高算法性能，本章在4.3节中还引入了基于聚类的深度强化学习方法。通过使用合适的聚类方法和参数选择，对出行数据进行了预处理，以便在深度强化学习模型中实现更好的特征提取和泛化能力。最后，通过对深度强化学习模型进行改进，提高了出行模式与时间选择的准确性和稳定性。

第五章 模式与出发时间选择方法的训练与评估

建立好深度强化学习模型后，进行训练和评价对模型的发展和应用具有重大意义。这一过程可以从多方面优化模型性能，为模型在实际应用中的选用提供依据。首先，训练和评价过程可以验证模型的有效性，确保模型能够有效解决所面临的问题。训练过程使智能体学会最优策略，而评价过程则有助于了解模型在实际应用场景中的性能表现。其次，训练和评价过程有助于提高模型性能。智能体在训练过程中持续优化神经网络参数，提高任务表现。评价过程揭示了模型在某些方面的不足，为进一步优化提供线索。训练和评价过程还可以揭示不同算法在特定任务上的优劣，为实际应用中的算法选择提供依据。同时，对现有模型的训练和评价可以发现算法存在的问题和不足，激发新的算法研究和改进，进而推动深度强化学习领域的发展。总之，建立好模型后的训练和评价过程不仅能全面了解模型性能表现，提高泛化能力，优化参数设置，还能为算法选择和领域研究提供重要支持。

在本章中，将对上一章中提出的基于深度强化学习的出行模式与时间选择模型进行训练，并对训练结果进行详细分析，以了解模型的收敛速度和稳定性。此外，本章还将对模型进行评估，包括与传统方法的对比和模型参数的灵敏性分析，以验证模型的有效性。

5.1 模型的训练与分析

在本节中，将对上一章提出的算法模型进行训练，并对训练结果进行详细分析。通过观察训练过程中的学习曲线和计算累积奖励，可以深入了解模型在训练过程中的性能变化和整体表现。学习曲线和累积奖励作为评估深度强化学习模型表现的关键指标，能够从多方面反映模型性能。学习曲线揭示了模型在训练过程中的性能变化，有助于判断模型收敛速度和稳定性；累积奖励则反映智能体在任务中的整体表现，可以用于比较不同模型或算法的相对性能优劣。此外，这些指标还有助于评估模型的泛化能力和算法效率。观察学习曲线和累积奖励在训练集与测试集上的表现，可以判断模型是否适应新环境；同时，学习曲线还能反映算法在训练过程中的效率差异。因此，这些指标为模型优化、算法选择和参数调整提供了重要依据。

5.1.1 实验场景的设置

在训练之前，先确定一系列仿真参数来构建训练模型的场景。表5.1列出了数值实验中使用的各个参数及其描述和取值。这些参数反映了不同出行方式的特点，如公交车和地铁的票价、运行频率和停留时间等。此外，还包括了如时间价值、提前到达和迟到

的时间表延误成本等与个体出行决策相关的因素。

表 5.1 实验中使用的仿真参数取值

参数	描述	数值
t_{\min}	最早出发时间	07:00
t_{\max}	最晚出发时间	09:00
t_{unit}	出发时间选择的单位间隔（分钟）	30
E_1	将成本映射到奖励的常数	100
E_2	将成本映射到奖励的常数	0.1
α	时间价值（元/分钟）	0.5
β	提前到达的时间表延误的单位成本（元/分钟）	0.05
γ	迟到的时间表延误的单位成本（元/分钟）	0.3
o	燃油价格（元/公里）	0.56
\	公交票价（元）	2
\	地铁起步票价（元）	1
\	地铁每公里递增票价（元/公里）	0.2
\	公交高峰频率（辆/小时）	10
\	地铁高峰频率（辆/小时）	14
\	公交非高峰频率（辆/小时）	6
\	地铁非高峰频率（辆/小时）	8
\	公交停留时间（秒）	40
\	地铁停留时间（秒）	30

5.1.2 智能体的聚类与选取

选择 60 个具有时间依赖性的起点-终点（OD）出行旅程作为智能体选取的样本集，并将它们的旅行特征输入到聚类方法中。表 5.2 为部分出行的相关信息，包含出行 ID，出发时间，出发路段，到达路段，与公共交通的可达性。

在实施 DBSCAN 算法的过程中，需要确定两个关键参数：邻域半径 (μ) 和最小样本数 (m)。邻域半径的选取至关重要，因为它决定了算法认定哪些点属于同一个簇。如果两个点之间的距离小于邻域半径，那么就认为这两个点属于同一个簇。另一个关键参数是最小样本数，它表示一个簇中至少需要包含 m 个样本才能被认定为有效簇。这两

表 5.2 智能体部分样本数据示例

出行 ID	出发时间 [s]	出发路段	到达路段	行程距离 [km]	可达性 [km]
1	76	912941#0	922772#0	6.37	1.48
2	267	109758#0	673925#0	2.89	0.89
3	331	76760#0	98918#0	5.56	1.89
4	0	818890#0	27601#0	4.21	1.93
...
60	18	818890#0	81924#0	3.41	0.32

个参数的选择会直接影响到聚类结果的质量，因此需要慎重考虑。

在本研究中，选择了经验法来确定 DBSCAN 算法的参数。经验法是一种依赖于经验或常识的方法，通过手动调整邻域半径和最小样本数的值来寻找最佳参数组合。首先尝试了不同的邻域半径和最小样本数组合，同时记录每种组合的轮廓系数得分。轮廓系数是一种用于评估聚类结果质量的指标，其值介于 -1 和 1 之间。计算轮廓系数的方法是将一个样本的簇内平均距离 (a) 与与其最近簇的所有样本的簇内平均距离 (b) 进行比较，计算得出该样本的轮廓系数为 $(b - a)/\max(a, b)$ ，然后计算所有样本的轮廓系数的平均值。轮廓系数越接近 1，说明聚类效果越好；越接近-1，说明聚类效果较差。

为了展示这一过程，表 5.3 呈现了尝试过的不同参数组合及其对应的轮廓系数得分。这有助于选择出最佳参数组合，从而提高聚类结果的质量。通过这种方法，可以确保在处理具有代表性的个体时，采用了较为合适的聚类参数。

表 5.3 参数组合及轮廓系数得分表

邻域半径 (μ)	最小样本数 (m)	轮廓系数得分
0.3	2	0.75
0.3	3	0.72
0.5	2	0.82
0.5	3	0.79
0.7	2	0.76
0.7	3	0.74

通过实验，发现最佳参数组合为邻域半径为 0.5，最小样本数为 2。在这个参数组合下，得到的轮廓系数为 0.82，表明聚类效果良好。将最小点数设置为 10 和最小距离阈值设置为 0.05，共得到了 4 个类别以及 2 个噪声点，聚类的结果如图 5-1 所示，图中标

识的 4 个代表点将作为代表智能体，在训练时其经验被放入公共记忆池中。

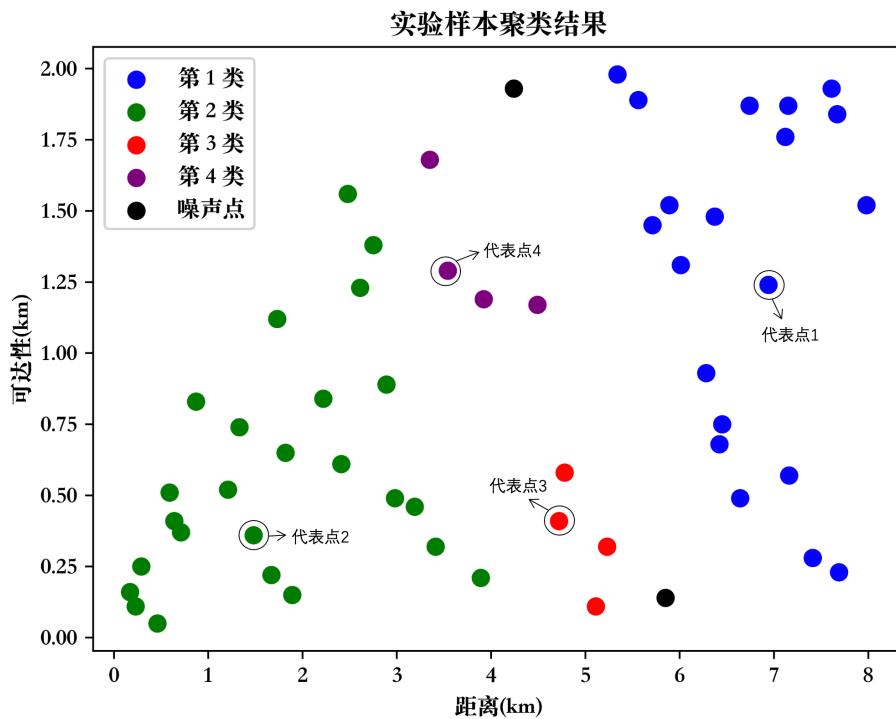


图 5-1 代表智能体的位置示意图

5.1.3 训练及结果分析

在本文中，采用深度 Q 网络算法来优化出行模式和出发时间的选择，以获得最佳的出行体验。有数值实验均在标准计算机上进行，配置为 Intel Core (TM) i5-9400F 2.90 GHz CPU 和 8 GB RAM。将训练过程设置为 800 个 episode。在每个时间步长内，智能体根据当前状态选择一个动作，然后接收到环境返回的奖励，并转移到下一个状态。我们使用 ε -贪婪策略来探索动作空间，其中 ε 在前 400 个 episode 中线性下降到 0.1，然后保持不变。在探索时，智能体将以 ε 的概率选择一个随机动作，否则将根据当前 Q 值估计选择一个最佳动作。

使用 SUMO 仿真工具来模拟城市交通网络，四个代表智能体的 OD 在网络中的位置如图 5-9 所示，将智能体的行为应用于仿真中的出行模式选择和出发时间选择中。根据仿真结果，可以获得出行时间、路线和交通方式等方面的信息，以评估该算法的性能。

首先，在训练过程中设置了一个经验回放机制。该机制用于存储代理在仿真环境中所经历的状态、动作、奖励和下一状态等信息，并且按照一定的概率进行抽样，以保证数据的独立性和随机性。其次，采用目标网络的方法来减小估计误差的影响。在训练过程中使用了两个神经网络，即一个本地网络和一个目标网络。本地网络用于根据当前状态计算 Q 值，而目标网络则用于计算目标 Q 值。在一定的时间间隔内，目标网络的参



图 5-2 代表智能体的位置示意图

数会从本地网络中更新，以缓解估计误差的影响。最后，将使用 Adam 优化器来对网络进行训练，以提高算法的收敛速度和性能。

图 5-3 显示了在训练 800 个周期中损失值和奖励值的收敛模式变化。在损失值变化图像中，整体呈现出明显的下降趋势，这与期望相符。在训练初期，由于智能体在训练开始阶段需要对环境进行行动探索，因此损失函数值波动较大。然而，这种波动主要集中在前 300 个周期，并且持续时间相对较短。实际上，从大约第 500 个周期开始，损失函数值的变化微小，几乎呈直线趋势。这一观察结果明确地证实了算法的收敛性。从图中可以看出，本文提出的深度强化学习算法在训练过程中表现良好，并已在此任务上达到收敛。为了评估算法在此任务上的性能，可以计算平均奖励值。在训练过程中，每个代表个体通过遵循深度强化学习推荐的行动所获得的奖励的变化结果如图 5-3 (b) 所示。考虑到每个代表个体的旅行都具有不同的 OD，相关的奖励大小也不同。事实上，可以轻易地发现，智能体 1 可能会有最长的旅行，而智能体 3 和 4 可能会有较短的旅行。然而，无论奖励的绝对值如何，所有代表个体的曲线都呈现出增长的趋势，这意味着他们通过与环境的交互并学习不断改进他们的旅行选择。在大约 700 个阶段内，每个代表的奖励基本稳定，并且此后变化不大，这一观察结果与图 5-3 (a) 一致。图 5-3 的结果表明，通过使用深度强化学习，智能体能够有效地学习并逐渐改善其决策。在训练过程中，智能体能够对环境进行探索，并通过与环境的交互学习如何选择最优的行动，从而

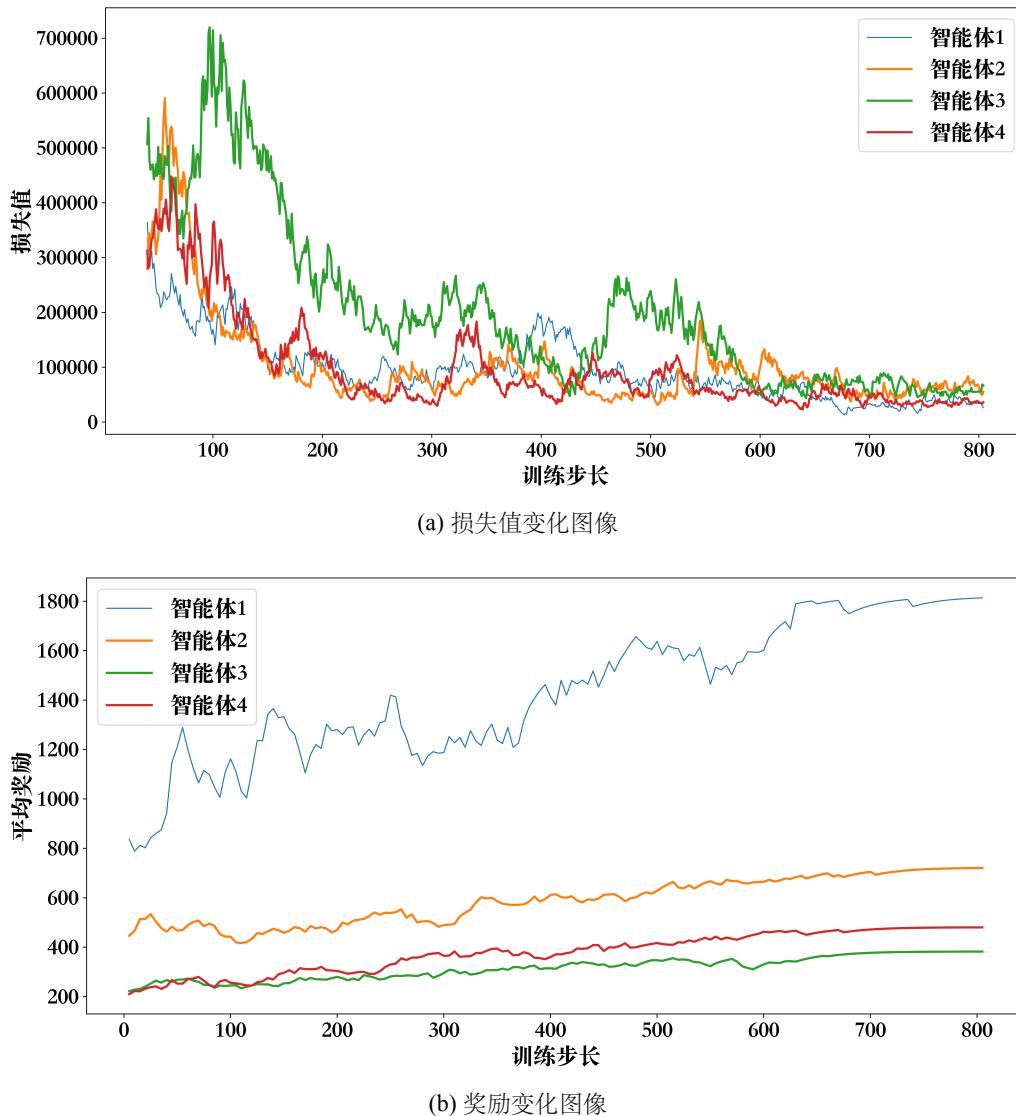


图 5-3 训练过程中各个代表智能体损失值与奖励值的变化图像

最小化旅行时间和出行成本。此外，随着训练周期的增加，智能体的学习能力逐渐提高，同时收敛速度也变得更快。

图5-6中展示了四个代表性个体在训练过程中遵循的深度强化学习建议行动。初始阶段，智能体的行动变化频繁，这是由于行动探索阶段所导致的。在这个阶段，智能体个体主要通过不断地尝试来了解环境并寻找最佳策略。由于环境的不确定性，智能体可能会尝试多种不同的行动，以了解每种行动的结果和潜在收益。但是，一旦智能体个体进入利用阶段，它们开始依赖其已经学习到的知识和经验，从而选择能够带来最大收益的行动。在这个阶段，由于智能体个体已经建立了一个比较准确的环境模型和策略，所以行动变化的波动性逐渐减小，每个智能体个体似乎已经找到了其自身出行模式与出发时间的最优解。在利用阶段期间，智能体个体偶尔会进行一些行动更改，这可能是由于 ε -贪心策略所导致的。 ε -贪心策略是一种在训练初期使用的策略，其目的是在探索阶段

时增加行动的随机性，从而帮助智能体个体更好地了解环境。随着训练的进行， ε 的值逐渐降低到一个非常小的值，从而逐渐减小了随机性，使得智能体个体的行动更加稳定和准确。总的来说，这些观察结果表明，深度强化学习算法在训练过程中，智能体个体通过探索与利用阶段的交替学习，最终学会了最优策略，并且能够在不同的情况下做出正确的决策。

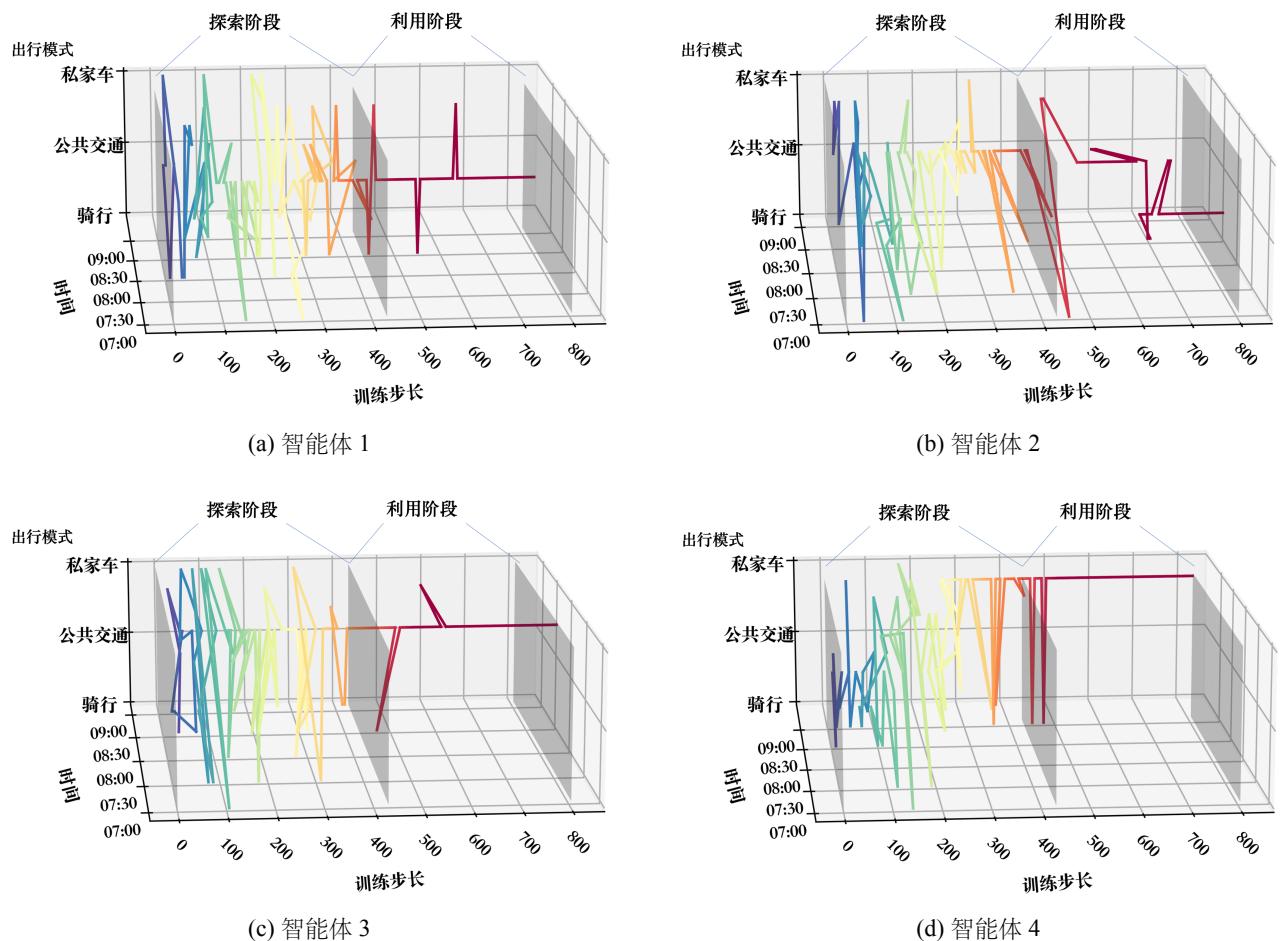


图 5-4 训练过程中代表智能体的动作选择变化

5.2 模型的评估

在本节中，将对基于深度强化学习的出行模式与时间选择模型进行评估，其目的在于验证模型的有效性和泛化能力。本节将分别从模型泛化能力检验、与传统方法的对比和模型参数的灵敏度分析三个方面对模型进行评估。在泛化能力检验中，使用测试集对模型进行测试，以验证其在不同数据集上的表现；在对比传统方法方面，将使用传统的离散选择方法与深度强化学习方法进行对比，并分析两者的优劣势；最后，在参数灵敏度分析中，将探讨模型参数对模型表现的影响，以确定模型最优参数。

5.2.1 模型泛化能力检验

为了对模型的泛化能力进行了测试，从训练数据中随机选取一部分个体，并将其作为测试集。然后对这些测试集中的个体进行测试，并计算其平均奖励值。为了检验所提出方法得到的解决方案的优劣或最优化，将所提出的方法与传统 DQN 方法和朴素法进行比较。传统 DQN 方法是一种已经被广泛研究和应用的深度强化学习方法，因此可以作为一个有代表性的基准。而朴素法是一种简单的贪心算法，它的性能不如传统 DQN 和所提出的方法，但它的计算复杂度很低，因此可以作为一种基准来比较所提出方法的性能。通过与传统 DQN 和朴素法的比较，可以验证本文方法的优越性和泛化能力。传统 DQN 方法与本文方法不同的是将只选择一个智能体样本进行训练。朴素法是在每个仿真中随机选择一个动作，为了充分探索动作空间，朴素法将运行 1000 个仿真次数。在这次比较分析中，测试数据集选择了网络中 50 个不同 OD 的出行个体，这些个体在训练中均没有被考虑。为了进行比较，让每个测试个体根据这些相互比较的决策者之一分别采取行动，并收集所得到的奖励。

图5-5(a)、5-5(b) 和5-5(c) 展示了三个选定的测试个体的比较结果。结果表明，朴素法的奖励存在显著波动，因为随机选择的行动不能保证总是获得良好的结果。相比之下，所提出方法和普通 DQN 获得的奖励是与 JTMDTC 问题的最优解相关的单一值，如图中的直线所示。在这三个测试个体中，所提出方法给出的解决方案优于普通 DQN，更优于大部分朴素法给出的解决方案，这一结果明显可见于图5-5(a) 和5-5(c)。

为了更全面地了解所有 50 个测试个体的比较情况，找到了朴素法为每个测试个体获得的最大奖励，将其视为该个体出行模式和出发时间问题的近似最优解。将所提出方法和传统 DQN 给出的解决方案与这个参考值进行比较，可以从全局角度观察性能差异。图5-5(d) 显示，本文方法的大部分解决方案（超过 40 个）超过了参考奖励值的 95%，这意味着这些解决方案接近最优。然而，对于普通的 DQN 来说，情况并非如此。只有大约 30% 的解决方案超过了同一参考值的 95%，还有相当多的解决方案甚至低于参考值的 60%。这些比较结果表明，所提出的方法在解决许多个体的出行模式和出发时间问题时具有有效性，并且代表在完成这个任务中所发挥的重要作用。由于测试个体并不是训练的一部分，结果表明所提出方法具有良好的可迁移性。

在本次实验中，随机选取了一部分个体，并将其用作测试集。然后，对这些测试集中的个体进行了测试，并计算了它们的平均奖励值。结果表明，测试集中的个体的表现略低于训练集中的个体，但仍表现出较好的性能。这表明所提出的方法具有一定的泛化能力，即能够适应一些新的情况并产生有意义的结果。

5.2.2 部分信息感知的模型检验

在现实世界中，人们在做出决策时，通常不会拥有完全的信息，例如，可能无法准确地了解周围环境、其他人的行为、交通状况等。因此，仅仅依靠部分信息进行决策，可能会导致更差的行动选择。为了比较所提出的深度强化学习方法的性能，此小节引入

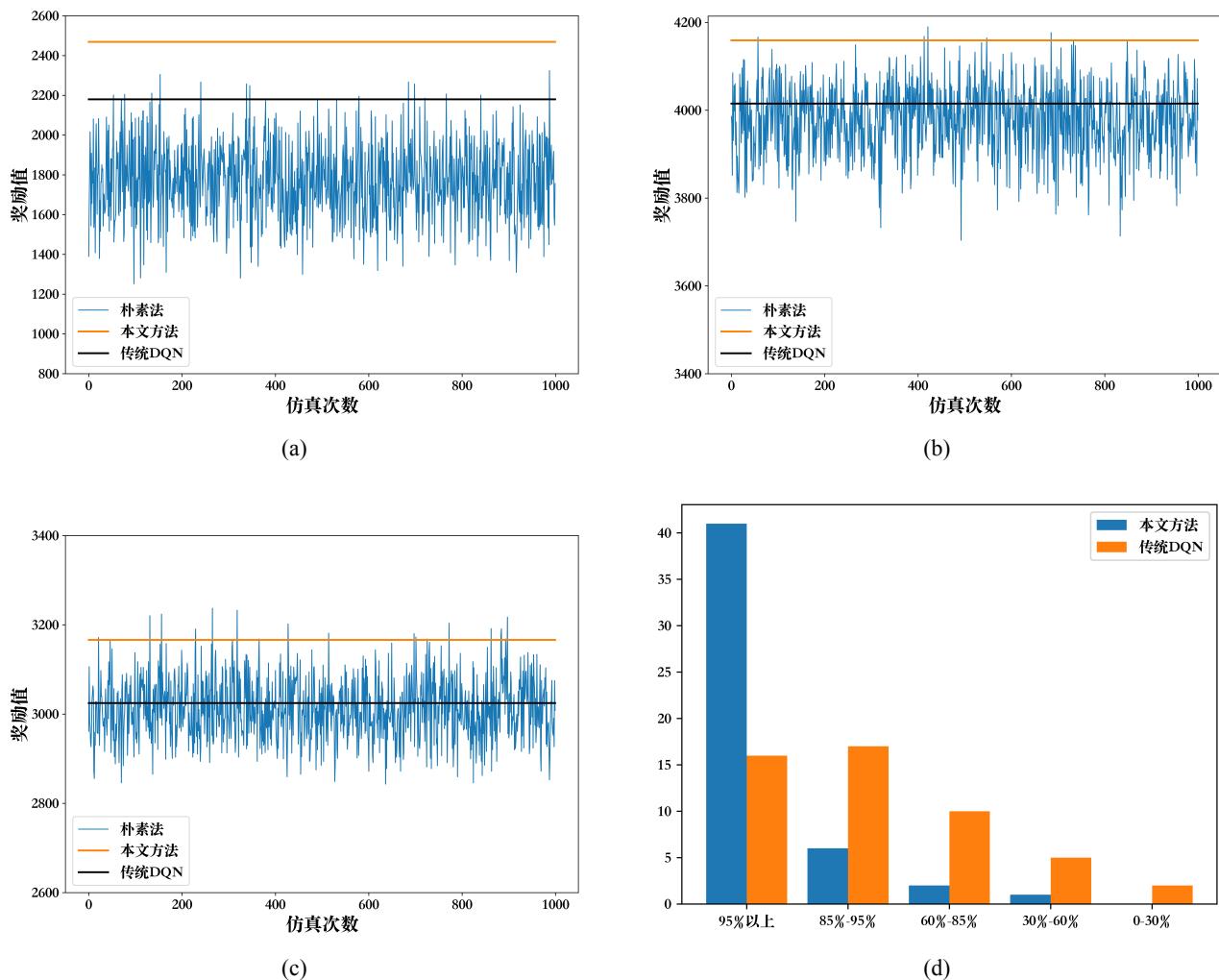


图 5-5 选定测试个体的性能比较以及与朴素法最大奖励值的比较

了另外两种模型，以考虑在部分信息的情况下，深度强化学习方法是否仍然是有效的。

第一个是一阶马尔可夫链（MC）模型，仅使用当前旅行距离和出发时间差来决定下一个选择。其状态、转移和初始状态概率以及奖励函数均源自 DRL 模型。两者的主要区别在于决策过程。MC 模型使用转移和初始状态概率来模拟随时间变化的行为，而深度强化学习模型使用迭代试错过程。在一阶马尔可夫链（MC）模型中，每个时间步都可以看作是一个状态，用 s_t 表示。模型假设，行为只取决于当前状态 s_t ，并且在状态 s_t 下，采取行动 a_t 会以概率 $p(s_{t+1}|s_t, a_t)$ 转移到下一个状态 s_{t+1} ，其中 $p(s_{t+1}|s_t, a_t)$ 表示在状态 s_t 下采取行动 a_t 并转移到状态 s_{t+1} 的概率。这个概率可以通过训练数据中的频率来估计。另外，MC 模型还需要一个初始状态分布 μ ，表示在开始时每个状态的出现概率。这个分布也可以通过训练数据中的频率来估计。在决策过程中，MC 模型首先从初始状态分布 μ 中随机选取一个初始状态 s_0 。然后，在每个时间步 t ，根据当前状态 s_t 和行动 a_t ，使用转移概率 $p(s_{t+1}|s_t, a_t)$ 随机转移到下一个状态 s_{t+1} 。在每个时间步，根据当前状态 s_t 和一些其他信息（例如旅行时间，距离等），MC 模型使用某种规则（例如贪心算法）选择下一个行动 a_t ，直到终止状态被达到。相比之下，深度强化学习方法不仅使用状态和行动，还使用神经网络来学习状态和行动之间的复杂映射关系，并使用 Q 值函数来指导行动选择。因此，深度强化学习方法具有更好的表达能力和泛化能力，可以更好地处理具有大量状态和行动的环境。

第二个是传统的 MNL 模型，传统的 MNL 模型是基于多项式逻辑回归的经典模型，它用于预测个体选择某个行动的概率。MNL 模型通常使用可解释的特征来描述选择行动的动机。这些特征通常是人为选择的，而不是通过学习过程得到的。例如，特征可能包括行动的属性（如价格、距离、时间等），个体的属性（如性别、年龄、收入等）等。在本文中，用于对比的 MNL 模型使用深度强化学习的奖励函数作为效用函数。这里的奖励函数（式（4.7））考虑了行动的成本和收益，通过计算收益与成本的比值得到奖励。

使用前文中在信息完全的情况下，由提出的方法得到的个体出行选择作为基准线，根据其来比较其他模型的性能。考虑了三个性能评估和比较指标。除了奖励值之外，另外两个是负对数损失（NLL）和 Jaccard 指数。NLL 通常用于评估分类任务中的模型性能，特别是在类别不平衡的情况下。在本文中，使用 NLL 作为评价指标，可以测量每种模型的预测能力，即它们在给定历史信息的情况下是否能够正确地预测个体的行动选择。NLL 的较低值表示模型更准确地预测了行动选择，因此较低的 NLL 值是期望的。Jaccard 指数是一种测量相似性的指标。在本文中，它可以测量每个模型产生的行动选择与基准线之间的相似程度。基准线是在信息完全的情况下，由提出的方法得到的个体出行选择。因此，Jaccard 指数可以反映其他模型相对于基线的优化水平。值越高表示其他模型的行动选择越接近基线，因此较高的 Jaccard 指数是期望的。这两个指标都衡量其他模型产生的行动与基线获得的行动之间的接近程度或相似程度。因此，它们可以反映其他模型相对于基线的优化水平。

表??展示了三个性能指标的比较结果。正如预期的那样，在信息完全的情况下，

所提出的方法表现出了最佳性能，可以获得尽可能多的奖励。相比之下，其他所有模型的性能都较差，这可以通过比较平均奖励值来看出，这些值都低于基线的奖励值，类似的趋势也反映在 NLL 和 Jaccard 指数上。然而，即使在部分信息的情况下，所提出的方法仍然表现出比一阶 MC 模型和 MNL 模型略好的性能，这表明所提出的方法即使在存在部分信息的情况下，仍然是有效的。

表 5.4 不同替代模型的性能评估结果

模型	NLL	平均回报	准确度
一阶 MC 模型	19.23	1,354	0.24
MNL 模型	24.17	1,217	0.23
部分信息下的 DRL 方法	17.31	1,449	0.29
完全信息下的 DRL 方法（基线）	\	1,735	\

5.2.3 模型参数的灵敏性分析

为了探究本研究所提出的方法在模型参数变化时的性能变化，将进行两个敏感性分析。第一个参数是训练智能体数量的变化，第二个参数是训练个体的群组的选择。

为了探究前者的影响，在进一步的实验中，分别使用 1、10、20 和 40 个智能体进行训练。保持相同的实验设置，将 60 个个体的 OD 行程聚类成上述数字，并用另外 50 个测试个体则用于评估和比较。同样地，将朴素法作为参考进行比较。比较结果总结在表 5.5 中。由于内存溢出，训练 40 个智能体的实验无法在本实验计算机上完成，因此没有报告结果。从表中可以看出，随着训练智能体数量的增加，所需的训练或计算时间增加，并且智能体数量的增加确实会导致更好的奖励。将一个智能体转变为四个智能体，奖励得到了最大的提高。进一步将该数字增加到 10 或 20 并不能显著提高奖励。这个结果表明，增加代表智能体数量相较于计算成本不一定合理。实际上，少量智能体已经可以在合理的计算时间内产生相当好的结果。使用朴素法得到的最大奖励作为参考值，比较所提出的方法所给出的奖励高于参考值 95% 以上的测试个体数量。如预期的那样，对于 4、10 和 20 个智能体，这个数字保持较大且变化很小。图 5-6 展示了对于不同数量的智能体，八个选定测试个体结果的比较。只有一个智能体显然不足以超过朴素法，而四个或更多智能体则有更良好的结果。

为了显示所提出方法的性能并不因每个类别中挑选训练个体的不同而发生显著变化，使用不同的训练个体集合进行另一组实验，挑选在每个类别中挑选四个智能体进行训练，其余的实验设置保持不变。表 5.6 列出了这样四个实验的结果。由于不同的训练个体集合不会改变计算时间（对于四个代表，计算时间为 22 小时），因此不再报告这个度量。从结果中可以看出，所提出方法的性能是稳定的，不会因为用于训练的代表个体不同而表现出显著的变化。类似于图 5-6，图 5-8 显示了当使用不同的训练个体集合时，八

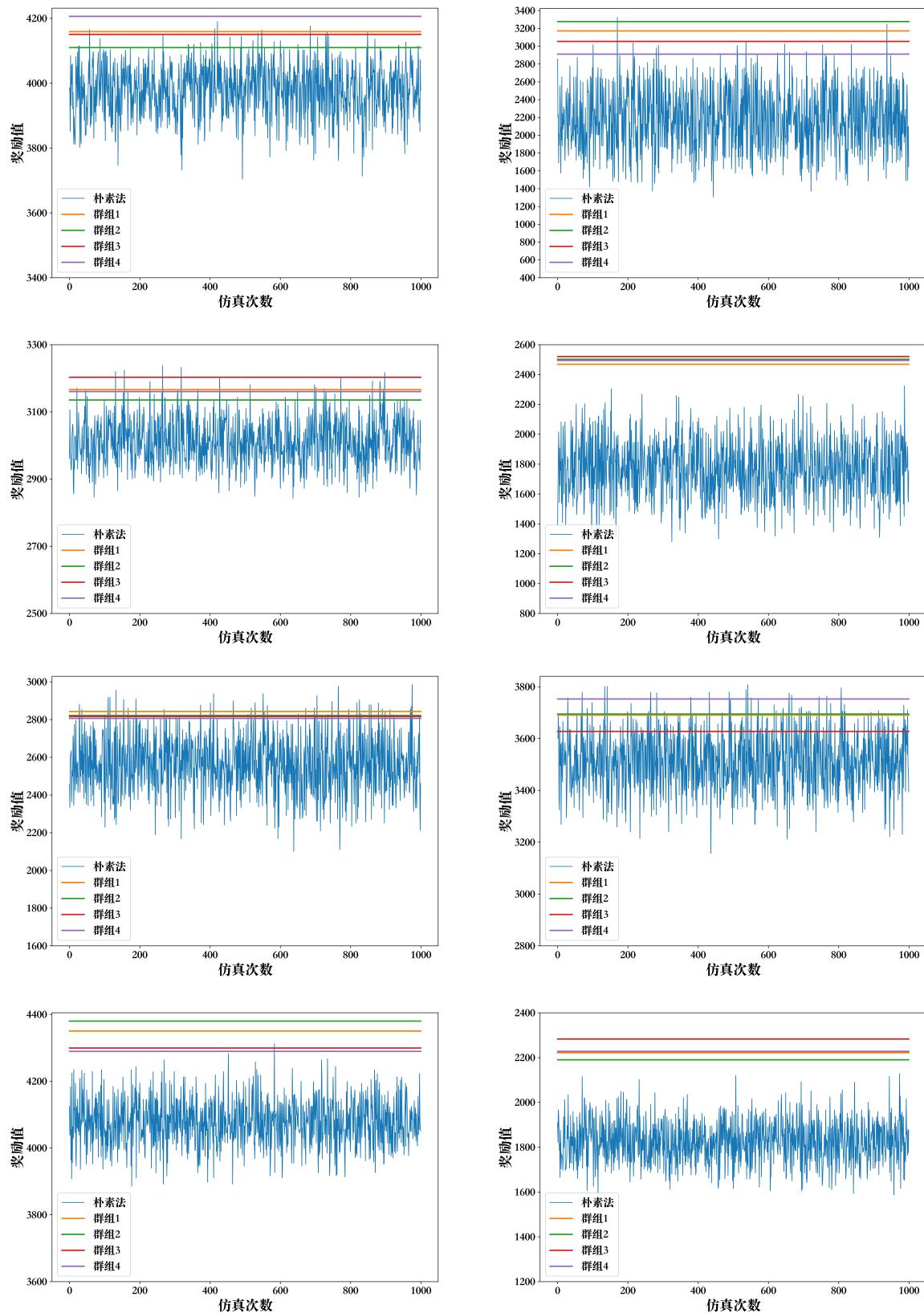


图 5-6 不同智能体数量训练下的八个测试个体测试结果

表 5.5 不同智能体数量训练下所提出方法的性能变化

代表点数量	1	4	10	20	40
训练时间 (小时)	5	22	68	140	\
平均奖励值	2,684	3,203	3,351	3,394	\
超过 95 (共 50 次) 的次数	16	41	45	46	\

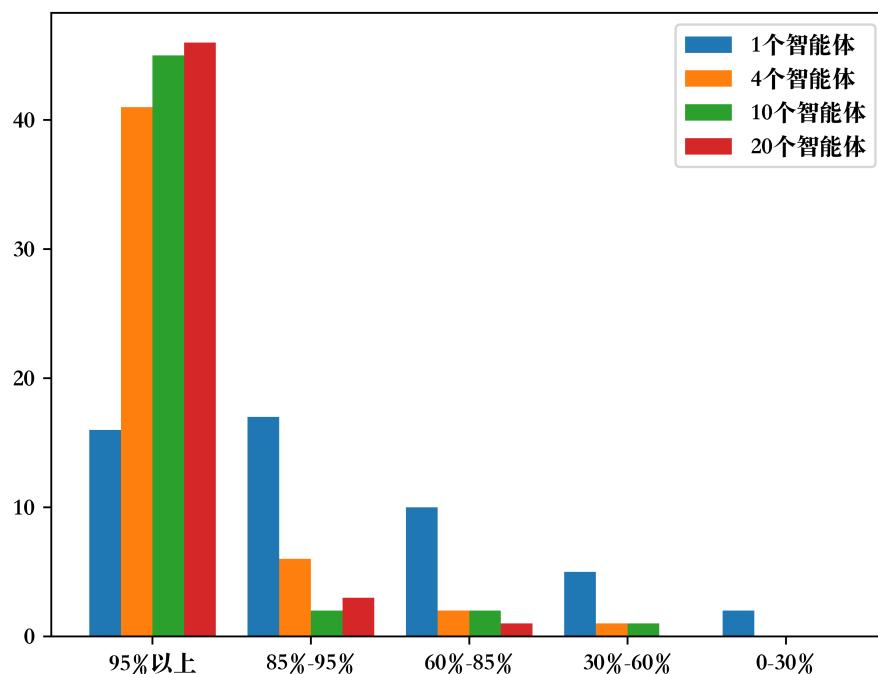


图 5-7 不同智能体数量训练下与朴素法最大奖励值的比较

个选定测试个体结果的比较，表明所提出的方法对代表个体的选择具有鲁棒性。这些结果表明，所提出的方法是有效的，不会受到训练代表个体集合的影响。从这些敏感性分析中可以得出结论，所提出的方法在实际应用中具有很强的可操作性和鲁棒性。通过将模型应用于新的时间依赖 OD 数据集并比较与传统模型和暴力方法的性能，证明了该方法在解决模式选择与出发时间问题方面的有效性。

表 5.6 不同智能体群组训练下所提出方法的性能变化

	集合 1	集合 2	集合 3	集合 4
平均奖励	3,203	3,179	3,280	3,248
超过 95% (共 50 个)	41	39	43	43

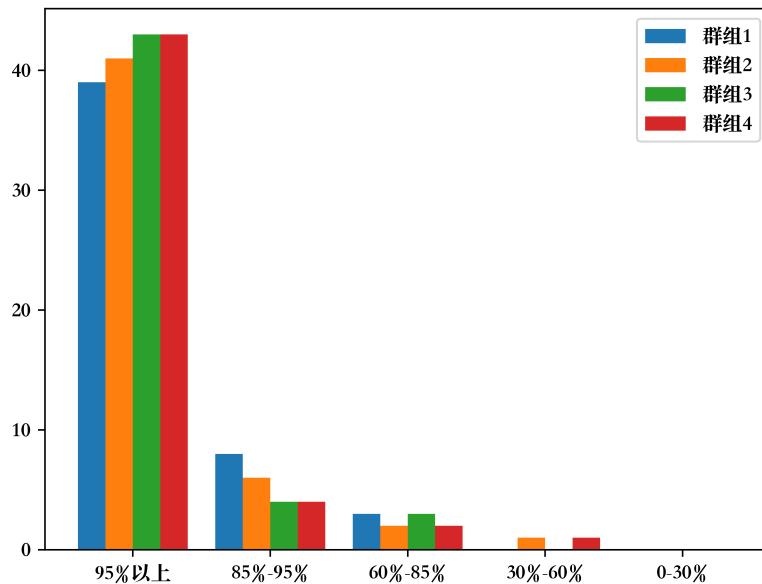


图 5-9 不同智能体群组训练下与朴素法最大奖励值的比较

5.3 本章小结

本章主要介绍了针对所提出的基于深度强化学习方法的训练以及评估。在小节 5.1 中介绍了实验场景的设置和智能体的聚类与选取方法，然后详细讨论了模型的训练过程及结果分析，在小节 5.2 中对实验结果进行了全面的模型评估，包括与传统 DQN 方法的对比、各测试个体的性能分析以及模型的泛化能力和灵敏性分析，也包含了模型的最优解、部分信息感知模型检验等多个方面。通过实验得出结论，所提出的基于深度强化学习的模式与出发时间选择方法在交通出行领域中具有很好的应用价值和实际意义。在实验中，该方法在解决许多测试个体的出行模式和出发时间问题时都具有较好的效果，并且相较于传统 DQN 方法和朴素法，能够更快地收敛到最优解。同时，模型的泛化能力

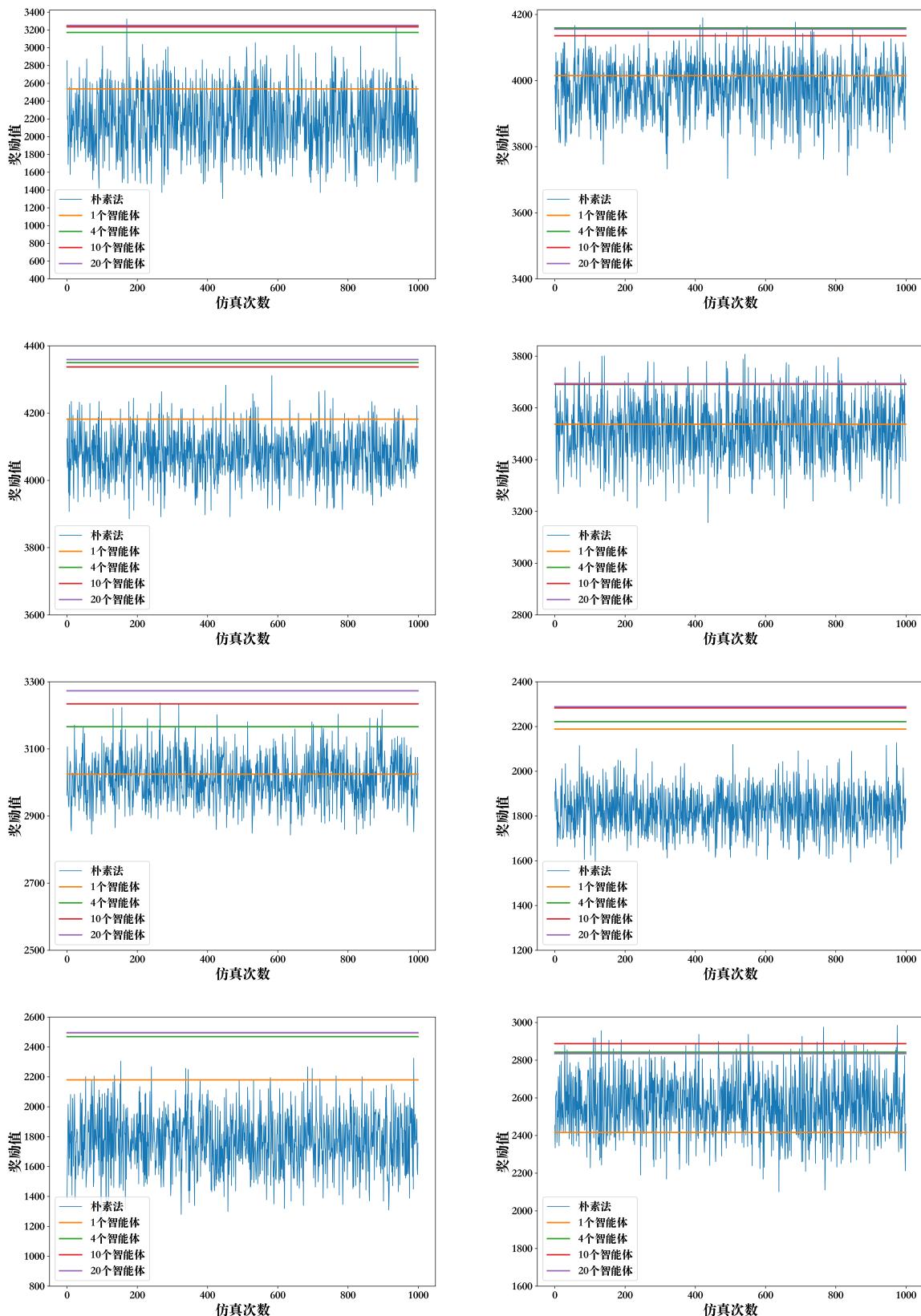


图 5-8 不同智能体群组训练下的八个测试个体测试结果

和灵敏性也在一定程度上得到了验证。在未来的研究中，可以进一步优化模型的结构和参数，提高模型的性能和适用范围，以更好地应对实际问题的挑战。

第六章 总结与展望

6.1 工作总结

在现代社会，人们对出行效率和出行体验的要求越来越高，交通问题也日益突出。传统的交通管理方法已经难以应对日益增长的交通需求和不断变化的交通状况。因此，研究如何更好地优化交通流动，提高个体出行效用，是当前交通领域的重要研究方向之一。传统的出行选择模型主要基于随机效用理论，对人们的出行行为进行描述和预测。然而，这些模型忽略了个体对交通环境的实时感知和对环境变化的适应能力，因此其预测准确性有限。而基于深度强化学习的出行选择模型则可以在动态交通环境中实时调整个体的决策策略，提高出行效用，并且具有更高的预测准确性和适应能力。本文提出了一种基于深度强化学习的新型出行模式与时间选择模型，旨在解决传统模型的局限性，并实现更高效、更智能的出行决策。该模型能够适应复杂的交通环境，并且可以处理许多具有出行决策请求的个体，具有较高的计算效率和优秀的性能表现。该研究成果不仅可以为交通管理提供新的思路和方法，还可以为人们提供更高效、更个性化的出行选择建议，提高人们的出行体验和生活质量。论文的主要研究如下：

(1) 基于 SUMO 的城市多模式路网场景

本研究探讨了城市交通多模式仿真环境的建立，主要包括路网编辑与生成、出行模式设计和流量生成三个方面。通过对 SUMO 搭建仿真路网存在的缺陷和不足进行分析，确保了实验场景能够满足不同研究需求。在出行模式设计部分，考虑了不同交通需求和策略，包括私家车、地铁、公交、自行车等，以便更好地反映城市交通的多样性。在流量生成部分，根据实际数据生成合理的交通流量，提供了准确的交通分析。通过该多模式仿真环境的建立，为交通规划、智慧交通等相关领域的研究和实践提供了可靠的仿真平台。

(2) 出行模式与时间选择问题特定的马尔可夫决策过程

马尔可夫决策模型可以充分考虑不同状态之间的转移概率。在出行模式和时间选择问题中，个体每天的状态可能受到多种因素的影响，例如天气、交通状况等，而这些因素的变化可能会影响个体决策。因此，马尔可夫决策模型可以对这些状态进行建模，并考虑它们之间的转移概率，以更好地理解和解决这些问题。其将整个决策过程形式化为一个数学问题，通过定义状态空间、动作空间和奖励函数，马尔可夫决策框架将出行模式和时间选择问题转化为一个数学问题，从而方便进一步的理论研究和算法优化。

(3) 基于深度强化学习的出行模式与时间选择模型

为了在处理出行数据时能够更好地提取和泛化特征，以提高模型的准确性和稳定性。本研究使用了基于聚类的深度强化学习方法，该方法使用聚类算法将出行数据集分

成不同的子集，然后对每个子集进行深度强化学习模型的训练。这样做的好处是可以在每个子集中学习更具体和相关的特征，从而提高模型的预测能力。此外，该方法还通过改进深度强化学习模型来进一步提高其准确性和稳定性。整个模型的目的是为了能够更好地解决出行模式与时间选择问题，提高个体的出行效用。

(4) 针对改进的深度强化学习方法的训练和评估

深度强化学习方法在解决交通出行中的多目标决策问题方面具有很好的应用潜力和实际意义。但是，深度强化学习方法的训练和评估是一个复杂的过程，需要针对实际问题进行合适的实验场景设置、模型参数调优、训练过程监控和结果分析，才能得出具有参考价值的研究成果。训练和评估可以帮助研究者更好地了解模型的性能、泛化能力和灵敏性等方面的特征，以及与其他方法的比较优劣。本研究对实验结果进行了全面的模型评估，包括与传统 DQN 方法的对比、各测试个体的性能分析以及模型的泛化能力和灵敏性分析。最终，研究得出的结论是基于深度强化学习的模式与出发时间选择方法在交通出行领域中具有很好的应用价值和实际意义，可以更快地收敛到最优解，并且模型的泛化能力和灵敏性得到了验证。

6.2 论文创新点

本文的创新点如下：

(1) 联合出行模式和出发时间选择问题建模为连续多天的马尔可夫决策过程，并用深度强化学习模型求解。

传统的出行模式选择和出发时间选择问题通常被视为静态决策问题，即每次出行都是一个单独的决策过程。但实际上，人们的出行决策往往受到历史决策的影响，因此将这些决策过程建模为连续多天的马尔科夫决策过程模型是更为真实和合理的。深度强化学习是一种强化学习的方法，可以通过让模型自主学习和改进来解决复杂的决策问题。与传统的规则或手动设计模型相比，深度强化学习模型可以更好地适应实际情况和变化，从而提高模型的效果和性能。

(2) 提出一种新的深度强化学习方法，利用聚类算法选取代表性个体进行高效训练，作为解决算法。

本研究提出的基于深度强化学习方法是为了解决多模式出行模式和出发时间选择问题而设计的，而传统的深度强化学习方法的训练过程通常需要大量的训练数据和计算资源。因此，本研究引入了一种新的深度强化学习方法，该方法使用聚类算法对大量个体进行聚类，并从每个聚类中选取代表性个体进行训练。这种方法可以大大减少训练数据和计算资源的需求，提高了训练效率和训练效果。这是本研究提出的一项创新点，也是本研究能够成功解决多模式出行模式和出发时间选择问题的关键因素之一。

(3) 在真实城市交通网络上进行多模式微观仿真实验，以展示与验证所提出的方法的有效性。

在交通规划和智能交通领域，很多方法和算法都是基于理论或简化的仿真场景进行

设计和验证的。然而，在真实世界的城市网络中，存在许多复杂的因素，如道路拓扑结构、交通信号控制、出行行为等等，这些因素对交通流和出行模式的产生和变化都具有重要的影响。因此，对于交通领域的研究来说，在真实世界网络上进行多模式微观仿真实验是非常重要的，可以更准确地反映出真实的交通情况，并验证所提出的方法的有效性和可行性。本文在真实世界网络上进行了多模式微观仿真实验，并与其他方法进行比较和敏感性分析，验证了所提出方法的优越性和鲁棒性。

6.3 展望

在这项研究中提出了一种深度强化学习方法，为在高峰时段（早上 7 点至 9 点）出行的个体提供更好的出行选择。该方法旨在在考虑不同出行方式的同时，最大程度地减少用户的出行时间和成本。研究结果表明，该方法在提供准确、高效的出行建议方面是有效的。这种方法的一个关键优势是其灵活性，因为奖励函数可以根据不同用户或情况的特定需求和要求进行调整。但由于学术水平和时间精力有限，仍存在一些问题值得进一步思考与完善，主要包括：

(1) 当前模型是为单个旅行者出行选择推荐而设计的，并没有考虑多个旅行者对交通系统的潜在影响。当前模型的设计是出于简化模型和降低计算复杂度的考虑，因此只考虑了单个旅行者的出行选择。但是，在实际交通系统中，多个旅行者的出行选择会相互影响，可能会导致交通系统的拥堵或效率低下。因此，未来研究可以考虑将模型扩展到多智能体框架，以考虑多个旅行者对交通系统的潜在影响。这种扩展可以提高模型的现实性和适用性，进一步优化交通系统的性能，并对城市交通规划和管理提供更好的指导和支持。

(2) 该研究使用了一天的出行需求数据作为输入，以提供第二天的出行建议。这意味着该模型无法提供实时的推荐决策，因为需要等待一天的数据输入。未来的研究可以探索如何整合实时数据以建立实时推荐系统，以便在实时交通拥堵或其他情况下，及时为旅客提供最优的出行建议。这可以通过结合实时交通数据、用户实时出行意向以及其他相关信息来实现。

(3) 未来研究可以探索如何将个体的行为和社会人口特征纳入到建模框架中，并考虑将换乘作为一种附加的交通方式纳入到决策过程中。同时，需要研究如何在出行选择方面考虑个体之间的动态互动和合作，以提高整个系统的效率和利用交通网络资源。

致 谢

感谢。

参考文献

- [1] McFadden D, et al. Conditional logit analysis of qualitative choice behavior[C]. *Frontiers in Econometrics*. 1973. 105-142.
- [2] 刘炳恩, 隽志才, 李艳玲, 龚博文. 居民出行方式选择非集计模型的建立[J]. 公路交通科技, 2008, No.146(116-120).
- [3] CHANG M S, LU P R. A multinomial logit model of mode and arrival time choices for planned special events[J/OL]. *Journal of the Eastern Asia Society for Transportation Studies*, 2013, 10:710-727. DOI: [10.11175/easts.10.710](https://doi.org/10.11175/easts.10.710).
- [4] Train K, McFadden D. The goods/leisure tradeoff and disaggregate work trip mode choice models[J]. *Transportation Research*, 1978, 12(5):349-353.
- [5] 诸葛承祥, 邵春福, 李霞, 等. 通勤者出行时间与出行方式选择行为研究[J]. 交通运输系统工程与信息, 2012, 12(2):126-131.
- [6] Koppelman F S, Sethi V. Incorporating variance and covariance heterogeneity in the generalized nested logit model: an application to modeling long distance travel choice behavior[J/OL]. *Transportation Research Part B: Methodological*, 2005, 39(9): 825-853. <https://www.sciencedirect.com/science/article/pii/S0191261504001596>. DOI: <https://doi.org/10.1016/j.trb.2004.10.003>.
- [7] Bhat C R. Analysis of travel mode and departure time choice for urban shopping trips[J]. *Transportation Research Part B: Methodological*, 1998, 32(6):361-371.
- [8] Bhat C R, Pulugurta V. A comparison of two alternative behavioral choice mechanisms for household auto ownership decisions[J]. *Transportation Research Part B: Methodological*, 1998, 32(1):61-75.
- [9] Papola A. Some developments on the cross-nested logit model[J]. *Transportation Research Part B: Methodological*, 2004, 38(9):833-851.
- [10] 杨励雅, 李霞, 邵春福. 居住地、出行方式与出发时间联合选择的交叉巢式 Logit 模型[J]. 同济大学学报（自然科学版）, 2012, 40(11):1647-1653.
- [11] Ding C, Mishra S, Lin Y, et al. Cross-nested joint model of travel mode and departure time choice for urban commuting trips: Case study in maryland-washington, dc region [J]. *Journal of Urban Planning and Development*, 2015, 141(4):04014036.

- [12] de Jong G, Daly A, Pieters M, et al. A model for time of day and mode choice using error components logit[J]. *Transportation Research Part E: Logistics and Transportation Review*, 2003, 39(3):245-268.
- [13] 栾鑫, 邓卫, 程琳, 等. 特大城市居民出行方式选择行为的混合 Logit 模型[J]. 吉林大学学报 (工学版) , 2018, 48(4):1029-1036.
- [14] Hensher D A, Greene W H. The mixed logit model: the state of practice[J]. *Transportation*, 2003, 30:133-176.
- [15] Fukuda D, Yai T. Semiparametric specification of the utility function in a travel mode choice model[J]. *Transportation*, 2010, 37(2):221-238.
- [16] Mahesh B. Machine learning algorithms-a review[J]. *International Journal of Science and Research (IJSR).[Internet]*, 2020, 9:381-386.
- [17] Pineda-Jaramillo J D. A review of machine learning (ml) algorithms used for modeling travel mode choice[J]. *Dyna*, 2019, 86(211):32-41.
- [18] 城市经济适用房居民通勤时间模式特征研究——以中国南京为例[J]. *Promet-traffic transportation: Scientific journal on traffic and transportation research*, 2019, 31(4):432.
- [19] 石庄彬, 鄢春花, 何明卫, 等. 建成环境对老年人出行方式选择的非线性影响[J]. *交通运输工程与信息学报*, 2023, 21(1):49-63.
- [20] Arentze T, Timmermans H. Measuring the goodness-of-fit of decision-tree models of discrete and continuous activity-travel choice: methods and empirical illustration[J]. *Journal of Geographical Systems*, 2003, 5(2):185-206.
- [21] Koushik A N, Manoj M, Nezamuddin N. Machine learning applications in activity-travel behaviour research: a review[J]. *Transport reviews*, 2020, 40(3):288-311.
- [22] Singh A, Thakur N, Sharma A. A review of supervised machine learning algorithms[C]. 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACoM). Ieee, 2016. 1310-1315.
- [23] 陈学松, 杨宜民. 强化学习研究综述[J]. *计算机应用研究*, 2010(2834-2838+2844).
- [24] 高阳, 陈世福, 陆鑫. 强化学习研究综述[J/OL]. *自动化学报*, 2004(86-100). DOI: [10.16383/j.aas.2004.01.011](https://doi.org/10.16383/j.aas.2004.01.011).
- [25] 李茹杨, 彭慧民, 李仁刚, 赵坤. 强化学习算法与应用综述[J/OL]. *计算机系统应用*, 2020(13-25). DOI: [10.15888/j.cnki.csa.007701](https://doi.org/10.15888/j.cnki.csa.007701).

- [26] Small K A. The scheduling of consumer activities: work trips[J]. The American Economic Review, 1982, 72(3):467-479.

作者简介

。。

作者攻读硕士学位期间发表的论文

- [1]. ZHi X, WEI T, CHEN R, et al.
- [2]. ZHi X, WEI T, JI H, et al.
- [3]. 韦天, 知心哥哥.

作者攻读硕士学位期间参与的研究课题

- [1]. **2018.5-2019.2:**
- [2]. **2020.1-2020.3:**

