◆人工智能算法与应用◆

基于逆强化学习的电动汽车出行规划方法研究*

李繁宛,张　莹**,华云鹏,李沐阳,陈元畅
(华北电力大学控制与计算机工程学院,北京　102206)

摘要:随着电动汽车的普及,对电动汽车出行规划问题的研究显得尤为重要。有别于路径规划,出行规划既需要考虑路径问题又需要考虑充电问题。本文提出了一种基于逆强化学习(Inverse Reinforcement Learning,IRL)的电动汽车出行规划(Electric Vehicle Travel Planning,EVTP)方法,有效地为电动汽车用户规划一条兼顾行驶路径短以及充电时间短的可达路径。将 Dijkstra 算法进行改进得到考虑充电行为的最短路径作为专家示例输入到逆强化学习算法中;利用逆强化学习算法得到兼顾行走与充电的奖励;在学习策略上,采用 Dueling DQN 算法高效更新 Q 值,提升学习性能;采用部分充电策略以及分段充电策略,提升充电效率并使研究更接近真实情况。通过对模型的工作性能和结果进行详细分析,并结合基准方法进行对比,结果表明,基于逆强化学习的电动汽车出行规划方法在行驶时间与充电时间两方面都有较好的性能,且具备很好的迁移性。

关键词:逆强化学习　电动汽车　出行规划　Dueling DQN　部分充电策略

中图分类号:U495　　**文献标识码**:A　　**文章编号**:1005-9164(2022)04-0668-13
DOI:10.13656/j.cnki.gxkx.20220919.007

随着社会经济的发展和生活水平的提高,人们对汽车的需求量逐年大幅度上涨,环境问题也随之而来,汽车尾气的排放对环境造成极其严重的影响。面对严峻的环境问题,我国大力推动碳达峰、碳中和各项工作,电动汽车产业普及率不断提升[1]。电动汽车一方面满足人们的短距离出行需求,另一方面它以电作为主要动力源,相比以燃油为驱动的汽车更加节能和环保。目前对于路径规划的研究主要针对燃油汽车,而对燃油汽车的路径规划无须考虑电量的因素,该场景不适用于电动汽车。随着电动汽车的普及,由于电动汽车的特殊属性以及"里程焦虑"[2],电动汽车的出行规划显得尤为重要。若能有效引导电动汽车用户选择可达目的地、能及时充电、行驶时间短、充电时间短的路线行走,出行效率就能有效提高。

本文提出了一种基于逆强化学习(Inverse Reinforcement Learning,IRL)的电动汽车出行规划(Electric Vehicle Travel Planning,EVTP)方法,将考虑充电行为的 Dijkstra 算法作为专家示例以保证专家

收稿日期:2022-03-30
* 国家自然科学基金项目(52078212)资助。

【作者简介】
李繁宛(1998-),女,在读硕士研究生,主要从事强化学习、路径规划研究,E-mail:18810733711@163.com。
【**通信作者】
张　莹(1982-),女,教授,主要从事智能交通、城市计算研究,E-mail:dearzppzpp@163.com。

【引用本文】
李繁宛,张莹,华云鹏,等. 基于逆强化学习的电动汽车出行规划方法研究[J]. 广西科学,2022,29(4):668-680.
LI F Y,ZHANG Y,HUA Y P,et al. Research on Electric Vehicle Travel Planning Based on Inverse Reinforcement Learning [J]. Guangxi Sciences,2022,29(4):668-680.

策略的优越性,在对电动汽车进行出行规划时须同时考虑电量和路径两方面因素,从而得到一条满足电量可达条件的最优路径。此外,本文还对充电策略进行考虑,在充电时采用部分充电策略以提高充电效率,使充电时间尽可能短。逆强化学习算法旨在对寻找路径的策略进行学习,因此有别于传统路径规划算法依赖于路网结构的特性。

1 相关工作

电动汽车出行规划(EVTP)问题可以拆分为电动汽车路径规划问题(Electric Vehicle Routing Problem,EVRP)以及充电策略问题。电动汽车的路径规划问题会受到电动汽车电池容量的限制,即电动汽车的电量不能低于某个限定最小值。因此,EVRP也可以被建模为受约束的最短路径规划问题,此类问题与经典的最短路径问题的不同之处在于解决方案必须满足附加约束,这是典型的 NP-hard 问题[3]。目前,求解受约束的最短路径问题大体上可分为两类方法:精确方法和启发式方法。

精确方法通过严谨的数学模型或数据结构规划问题,利用数学法则或数据结构搜寻的方式求得问题最优解[4]。Lu 等[5]提出了一种运用整数线性规划来求解电动汽车最优路线问题的方法,该方法以最小化电动汽车行驶时间为优化目标,然而线性规划的方法计算复杂度较高,仅在小型实例中有很好的表现,一旦应用在大型路网中往往需要很大的存储空间而导致方法不可用,因而其效率取决于实际路网的规模且不可直接迁移到其他未经预处理的路网中。

由于精确方法有较高的计算复杂度而不便应用于实际情况,在此方面有一定改善的启发式方法是EVRP中更为常用的一种求解方法。启发式方法[4]的主要思想是以当前解为基点,在当前解的邻域中寻找较优解赋给当前解,并继续寻找,直到没有更优解为止。Lebeau 等[6]对每条路径计算合并一条新路径的成本,对可行路径按成本量排序,再以贪心方式进行归并得到路径集。Felipe 等[7]采用两阶段启发法,在得到初始可行解的基础上不断对路径进行调整,以逐渐靠近最优目标,并在每一步更新目标函数的值,反复迭代直到目标函数的值收敛为止。类似地,Dijkstra 算法和 A* 算法[8]在路径规划问题中也被广泛运用,对其稍加修改也可以运用到 EVRP 中。Cuchý 等[3]提出了一种基于 A* 算法计算电动汽车最优路线的方法,并将减少充电行为中的等待时间作为优化的一部分。Kobayashi 等[9]提出了一种基于Dijkstra 算法的电动汽车路线搜索方法,这种方法对于充电行为也有一定的考虑,但未对充电时间作出优化。虽然启发式方法能有效减少运行时间,但是此类方法仍然不能独立于实际路网,即其运行效率仍与路网规模相关,且应用在不同路网时需要重新进行生成邻接矩阵等数据预处理工作。而本文提出的基于逆强化学习的方法与具体路网结构无关,即在某个路网中训练得到的模型也能在其他路网中有较优的性能。

对于 EVTP 问题,为得到一条最优路径,除了路径最短之外,充电策略也是需要重点考虑的因素。Wang 等[10]提出了一个基于电量驱动的上下文感知路径规划框架,该算法通过电量的约束限制 A* 算法对路径进行搜索,并且对于需要充电而绕路的情况给出了最佳方案。然而,此类方法在到达充电站时仍假设电动汽车会将电量充满,这并不适用于实际情况。在现实生活中,电动汽车通常只需要使其电量在接下来的路程中不会低于限定值即可,这样既可以减少充电时间又能满足可达性。此外,电动汽车的充电函数并不是线性的,电动汽车的充电效率会随着电量的变化而变化[11]。因此,本文提出了一种更接近于真实情况的充电策略,即分段充电以及部分充电策略。

2 方法

2.1 问题定义

对于 EVTP 问题,电动汽车车主通常希望得到一条行驶时间短、充电时间短、充电次数少的可达路径。因此,目标可形式化表达为

$$\min \sum_{i,j,(i,j)\in E} \frac{d_{ij}}{\bar{v}} x_{ij} + \sum_{j\in C} (s_j + c_j) y_j \ ,$$

式中,E 表示路网中边的集合;d_{ij} 表示路径 (i,j) 的距离,$\bar{v}$ 表示电动汽车在一条路径中的平均速度,则 $\frac{d_{ij}}{\bar{v}}$ 表示电动汽车在路径 (i,j) 上的行驶时间;x_{ij} 为一个指示变量,当电动汽车选择了路径 (i,j) 时,$x_{ij}=1$,否则 $x_{ij}=0$;C 表示充电桩节点的集合;s_j 表示在充电桩充电时的服务时间,在本文中设置 $s_j=10$;c_j 表示在 j 点的充电时间;y_j 为一个指示变量,当 j 点在充电桩节点的集合中且电动汽车选择在该充电桩充电时,$y_j=1$,否则 $y_j=0$。

此外,在寻找路径的过程中还应满足以下条件:(1)电动汽车选择的每一个节点都在路网的点集 G 中,选择行走的每一条路径都在边集 E 中;(2)路网

中的边为无向图,选择边 (i,j) 等价于选择边(j,i),即 $x_{ij}=x_{ji}$;(3)选择的任意两个连续的充电桩之间的电量消耗要小于限制条件,即到达充电桩时的电量大于限定的最小电量;(4)充电次数尽可能少。充电次数越多,充电桩总计服务时间越长[12],从而会降低电动汽车的出行效率。

2.2 马尔可夫决策过程

为将逆强化学习(IRL)算法应用到 EVTP 问题中,EVTP 问题可被建模为马尔可夫决策过程[13,14]。在此场景中的马尔可夫决策过程可由以下 5 元组表示。

(1)状态集 S。在 EVTP 问题中,状态集 S 中的每一个状态 s 由当前电动汽车的电量(pow)、与终点的距离比例(distance_ratio)、行走的角度(degree)、绕路情况(loop)组成。通过 pow 可以判断当前状态是否满足可达性条件,即行走时电动汽车的电量是否在限定值以上。另外,状态值的其他组成部分也是电动汽车选择路径过程中需要特别关注的因素,状态值中不包含具体位置信息,使得训练得到的模型在新的路网中也能有较好的表现。

(2)动作集 A。由于电动汽车的特性,动作集 A 中的动作 a 包含两类:行走行为和充电行为。在本文的设定中,动作集中共包含 10 个动作,前 8 个动作表示行走行为(假设所有的节点最多有 4 个邻接节点),后 2 个动作表示充电行为,集中离散化的动作用 one-hot 编码[15]表示。在行走行为中,电动汽车的目标是向更接近终点的方向行驶,且需要为之后可能出现的充电行为做准备。因此,行走行为中的前 4 个动作对某一位置的相邻点按照(degree,distance_ratio)升序排列,即先按照 degree 升序排列,若 degree 相同再按照 distance_ratio 升序排列;行走行为中的后 4 个动作对某一位置的相邻充电站按照上述规则排序,以便之后选择充电行为。若选择动作集中后 2 个动作,即电动汽车选择充电行为,在充电行为中本文对于充电量进行了区分,其中一个动作表示电量充到 80%,而另一个动作表示电量充到 100%,这种部分充电的策略有利于节省充电时间。

(3)奖励 r。传统强化学习算法中的即时奖励 r 往往是人为设定的,然而在复杂问题中仅凭经验通常无法设置最优的奖励值。例如,本文的场景需要同时考虑行走行为以及充电行为,难以人为设置兼顾二者的最优奖励。奖励是影响强化学习算法性能的重要因素,因此,本文采用基于逆强化学习的方法对奖励值进行设定[16],通过专家示例以及学习策略过程中得到的轨迹间的交互,求得最优的奖励值。

(4)动作价值 Q。$Q(s,a)$ 表示选择了动作 a 之后直到到达终点的奖励总和的期望,即

$$Q(s,a)=E[\sum_{t=0}^{T}\gamma^t r(s_t,a_t)\mid s=s_0,a=a_0]。$$

本文用 Dueling DQN 算法[17]近似动作价值 Q,即将某一时刻的状态 s_t 作为输入,输出对每个动作 a 的价值预测。此时 Q 值被定义为 $Q(s,a;\theta)$,其中 θ 为神经网络中的参数。

(5)策略 π。策略 $\pi(s_t)$ 依据当前状态对动作进行选择。在本文中使用 ε-greedy 以一定的概率进行探索,即大概率选择动作价值最大的动作,以极小的概率对动作进行随机选择来避免陷入局部最优。ε-greedy 的计算公式如下:

$$\pi(s_t)=\begin{cases}\text{argmax}_{a\in A}Q(s,a),\text{if random}[0,1]>\varepsilon\\ \text{random } a\in A,\text{otherwise}\end{cases}。$$

根据上述 5 元组,对算法整体流程表示如图 1 所示。

2.3 状态特征选择

状态是马尔可夫决策过程中的重要组成部分,状态的选择尤为重要。特征是状态到真实值的映射,包含一些重要的属性,且与奖励值有密切关系。在本文中将 EVTP 问题的特征定义为 5 个方面。

(1)电量 pow。pow 表示电动汽车是否迫切需要充电。此特征用当前状态下电动汽车的剩余电量来定义,剩余电量越少越迫切需要充电。

$f_{\text{pow}}(s_t)=\text{Soc}_t$。

(2)充电时间 charge_time。充电时间是电动汽车需要重点考虑的特征之一。充电时间过长会导致旅程时间过长,从而导致效率过低;充电时间过短可能会使电量不能满足接下来旅程的需要,从而导致目的地不可达。电动汽车的充电时间需要在满足可达的条件下尽可能短。

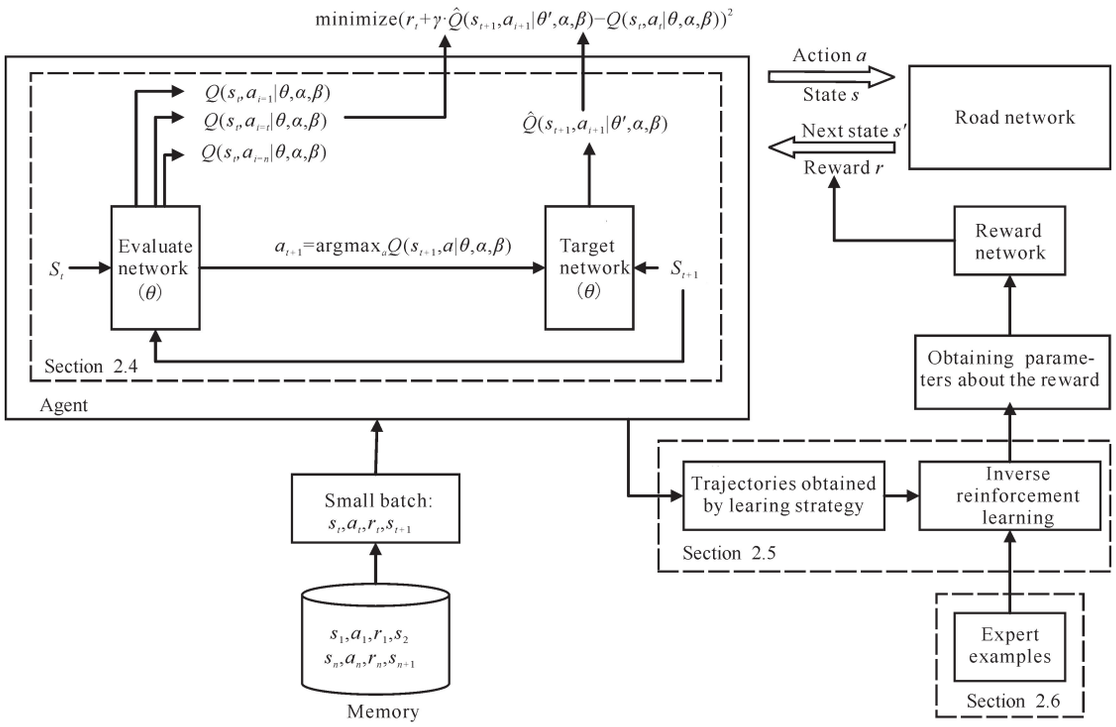

图 1　算法整体流程图

Fig. 1　Overall flow chart of algorithm

$$f_{\text{charge\_time}}(s_t)=\begin{cases}c_t,\text{if location}(t)\in C\text{ and charge\_index}(t)=1\\0,\text{otherwise}\end{cases},$$

式中,C 表示充电桩节点的集合;location(t)表示 t 步所在位置,location(t) $\in C$ 即 t 步时所在节点为充电桩节点;charge_index 表示充电指示,由 t 步时选择的动作决定,charge_index(t) = 1 即表示在 t 步时需要充电;c_t 表示计算得到的在 t 步时的充电时间。

(3)角度 degree。degree 表示当前状态所在位置与终点的连线以及起点与终点的连线之间的夹角,夹角越小说明电动汽车行驶的方向与终点方向越接近。

$$f_{\text{degree}}(s_t)=|\text{getdegree}(\text{origin},\text{destination})-\text{getdegree}(\text{location}(t),\text{destination})|,$$

$$f_{\text{degree}}(s_t)=\min(f_{\text{degree}}(s_t),360-f_{\text{degree}}(s_t)),$$

式中,getdegree 表示得到两点之间方位角的函数;origin 表示路径的起点;destination 表示路径的终点。

(4)距离比例 distance_ratio。distance_ratio 表示当前所在位置与终点的距离以及起点与终点的距离之间的比例关系,比例越小则越接近终点,所有距离均用 haversine 公式[18]进行计算:

$$f_{\text{distance\_ratio}}(s_t)=\frac{\text{haversine}(\text{location}(t),\text{destination})}{\text{haversine}(\text{origin},\text{destination})},$$

式中,haversine (A,B)表示点 A 与点 B 之间的 haversine 距离。

(5)绕路 loop。loop 表示寻找某起点终点对的路径过程中是否重复访问某一位置。

$$f_{\text{loop}}(s_t)=\begin{cases}1,\text{if location}(t)\in\text{runway\_history}\\0,\text{otherwise}\end{cases},$$

式中,runway_history 表示电动汽车走过的节点位置的集合,若当前位置已经在 runway_history 中,则说明电动汽车已经访问过该节点,即出现了绕路。若出现绕路往往说明某步的决策不是最优,因此并不希望路径中出现绕路的情况。绕路情况如图 2 所示。

给定起点、终点对(A,H),电动汽车由 E 到达 D 后,在 D 点若根据角度选择下一个点,极有可能选择点 E,此时便出现了绕路,虚线表示可能出现的绕路情况。

在上述特征中,电量对逆强化学习算法得到的奖励产生正向影响,充电时间、角度、距离比例、绕路均对奖励产生负向影响。

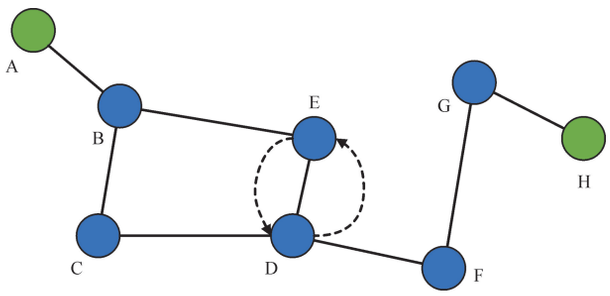

图 2　绕路情况

Fig. 2　Situation of loop

在对特征进行选择时,路径因素以及充电因素都需要被考虑以适用于 EVTP 问题,从而得到一条兼顾行驶和充电的最优路径。

2.4　Dueling DQN

除了对特征的选择外,对动作的选择在逆强化学习中也十分重要。逆强化学习算法与传统强化学习算法的区别仅在于奖励函数,虽然其算法流程与传统强化学习相似,但是在得到奖励函数之后仍需要正向训练更新 Q 值来学习策略。可用 $Q(s,a)$ 近似累计奖励,即

$$Q(s,a)=E[r_t+\gamma\cdot r_{t+1}+\gamma^2\cdot r_{t+2}+\cdots\mid s_t=s,a_t=a,\mu]。$$

根据上述设定,需要求得最优动作值 Q^*,最优动作值函数基于贝尔曼方程[19],表示如下:

$$Q^*(s,a)=E_{s'}[r+\gamma\cdot\max_{a'}Q^*(s',a')\mid s,a]。$$

为得到最优解,引入 Dueling DQN 算法,该算法是 DQN 算法的改进形式,采用非线性表示形式逼近动作值函数。与 DQN 方法类似[20,21],Dueling DQN 算法[22]用到了两个神经网络,分别表示为 $Q(s,a;\theta)$ 和 $\hat{Q}(s,a;\theta')$,前者被称为评估网络,后者被称为目标网络。通过这两个网络得到损失函数,进而对神经网络的参数进行更新,第 i 轮的损失函数如下:

$$L_i(\theta_i)=E_{s,a,r,s'}[(r+\gamma\cdot\hat{Q}(s,\mathrm{argmax}_{a'}Q(s',a';\theta_i);\theta'_i)-Q(s,a;\theta_i))^2]。$$

在实际问题中,有时采取的动作对下一个状态的变化情况没有太大的影响,因此使用 Dueling DQN 算法可以提升学习效果,加速收敛。该算法将每个动作的 Q 值拆分成价值函数 V 以及每个动作的优势函数 A,因此 Q 值可改写 $\hat{Q}$:

$$Q(s,a;\theta_i)=Q(s,a;\theta_i,\alpha,\beta)=V(s;\theta_i,\alpha)+\left(A(s,a;\theta_i,\beta)-\frac{1}{|A|}\sum_{a'}A(s,a';\theta_i,\beta)\right),$$

式中,α 和 β 分别为价值函数和优势函数的参数。

在普通 DQN 算法中,若想更新某个动作的 Q 值会直接更新 Q 网络,使得该动作的 Q 值改变。然而在 Dueling DQN 算法中,由于所有动作的优势函数 A 值的和为 0,不便于对优势函数进行更新,因此 Q 网络将优先对价值函数进行更新,即相当于对所有动作的 Q 值均进行更新,从而提升学习效果。Dueling DQN 的网络结构如图 3 所示。在得到所有动作的 Q 值后,ε-greedy 可被应用于动作的选择中。

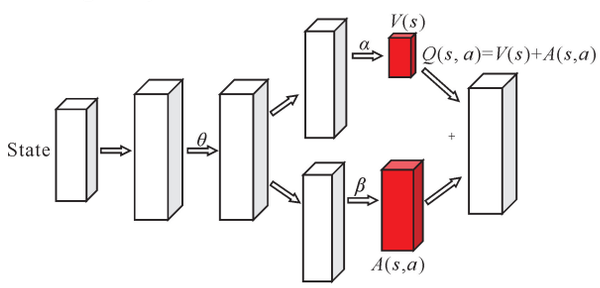

图 3　Dueling DQN 的网络结构

Fig. 3　Network architecture of Dueling DQN

2.5　逆强化学习

在马尔可夫决策过程中,除状态和动作外,奖励也尤为重要。无论是传统强化学习还是逆强化学习,最基本的假设都是最大化累积的奖励,因此奖励的设计非常关键。传统强化学习的即时奖励通常凭借经验人为设定,然而像电动汽车出行规划等复杂问题,仅凭经验设置奖励往往难以精准量化,因此,逆强化学习算法应运而生。本文将逆强化学习方法应用到电动汽车出行规划中以得到最优的奖励函数。

逆强化学习的思想与模仿学习有一定的相似之处,均需要借助专家示例库对问题进行优化。行为克隆是最早的模仿学习形式,与行为克隆简单学习一个状态到动作的映射不同,逆强化学习算法中奖励函数的学习方法原理如下:假设专家策略本身是根据某种奖励学到的最优策略,那么专家示例即可取得最优奖励值。根据此假设,设置参数化的奖励函数,并且寻找参数使得专家示例的奖励优于其他任意数据,这样即可求解出奖励函数。然而,上述“其他任意数据”并不具备可操作性,因此常采用迭代过程:根据当下学到的奖励函数学习策略并生成轨迹数据,该轨迹数据替代“其他任意数据”,对比专家示例和生成的轨迹数据,来学习下一轮奖励函数。逆强化学习的流程如图 4 所示。

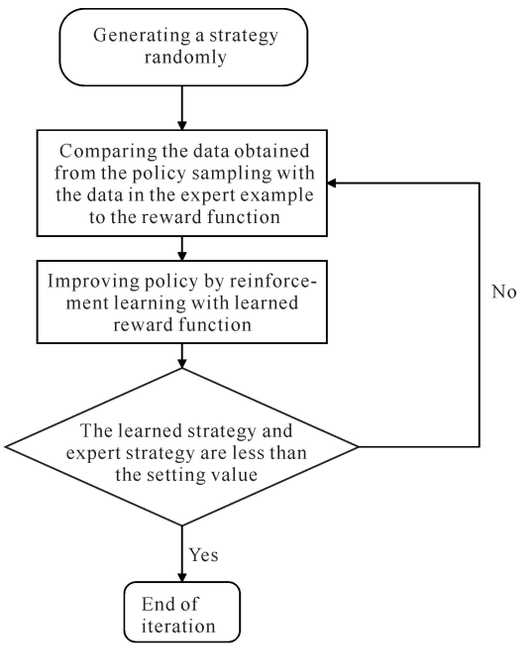

图 4　逆强化学习流程图

Fig. 4　Flow chart of inverse reinforcement learning

奖励被定义为线性结构,即特征向量的加权和,表示如下:

$$r(s_t)=\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{f}(s_t),$$

式中,$\boldsymbol{\theta}=[\theta_1,\theta_2,\cdots,\theta_K]$为 K 维权值向量,K 为特征向量维度;$\boldsymbol{f}(s_t)=[f_1(s_t),f_2(s_t),\cdots,f_K(s_t)]$为根据状态定义的与奖励相关的特征向量。因此,一条轨迹 ζ 的奖励可以表示为

$$R(\zeta)=\sum_t r(s_t)=\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{f}_\zeta=\boldsymbol{\theta}^{\mathrm{T}}\sum_{s_t\in\zeta}\boldsymbol{f}(s_t),$$

给定 Dijkstra 算法得到的专家示例集 $D=\{\zeta_1,\zeta_2,\zeta_3,\cdots,\zeta_M\}$,其中 M 为专家示例轨迹的数目,问题转化为找到参数 $\boldsymbol{\theta}$ 使得专家示例的奖励最优。因此,目标函数如下:

$$\max J=\max(R-\widetilde{R}),$$

式中,R 表示专家示例集获得的奖励,$\widetilde{R}$ 表示根据上一步学到的奖励函数学习策略并生成的轨迹数据获得的奖励。

由上述奖励函数的表示方式,目标函数转化为

$$\max_{\boldsymbol{\theta}}J(\boldsymbol{\theta})=\max\left(\frac{1}{M}\boldsymbol{\theta}^{\mathrm{T}}\sum_{\zeta_i\in D}^{M}\boldsymbol{f}_{\zeta_i}-\frac{1}{N}\boldsymbol{\theta}^{\mathrm{T}}\cdot\sum_{\tilde{\zeta}_j}^{N}\boldsymbol{f}_{\tilde{\zeta}_j}\right)=\max\boldsymbol{\theta}^{\mathrm{T}}\left(\frac{1}{M}\sum_{\zeta_i\in D}^{M}\boldsymbol{f}_{\zeta_i}-\frac{1}{N}\sum_{\tilde{\zeta}_j\in\widetilde{D}}^{N}\boldsymbol{f}_{\tilde{\zeta}_j}\right),$$

式中,N 表示生成的轨迹数据集 $\widetilde{D}=\{\tilde{\zeta}_1,\tilde{\zeta}_2,\tilde{\zeta}_3,\cdots,\tilde{\zeta}_N\}$中轨迹的数目,$\widetilde{D}$ 中的轨迹与 D 中的轨迹具有相同的初始状态以及起点终点对分布。要得到最优的参数,本文使用了基于梯度的方法,将 $J(\boldsymbol{\theta})$对 $\boldsymbol{\theta}$ 求偏导:

$$\nabla_{\boldsymbol{\theta}}J(\boldsymbol{\theta})=\frac{1}{M}\sum_{\zeta_i\in D}^{M}\boldsymbol{f}_{\zeta_i}-\frac{1}{N}\sum_{\tilde{\zeta}_j}^{N}\boldsymbol{f}_{\tilde{\zeta}_j}。$$

奖励函数参数 $\boldsymbol{\theta}$ 的更新使用了梯度上升的方法,直到其收敛。为防止过拟合,本文引入了参数 $\boldsymbol{\theta}$ 的 L2 正则化,目标函数改写为

$$J(\boldsymbol{\theta})=\boldsymbol{\theta}^{\mathrm{T}}\left(\frac{1}{M}\sum_{\zeta_i\in D}^{M}\boldsymbol{f}_{\zeta_i}-\frac{1}{N}\sum_{\tilde{\zeta}_j\in\widetilde{D}}^{N}\boldsymbol{f}_{\tilde{\zeta}_j}\right)-\lambda\boldsymbol{\theta}^2,$$

式中,λ 为正则化参数,且满足 $\lambda>0$。

因此,梯度公式变为两轨迹集间特征期望的差异加上正则化项的梯度:

$$\nabla_{\boldsymbol{\theta}}J(\boldsymbol{\theta})=\frac{1}{M}\sum_{\zeta_i\in D}^{M}\boldsymbol{f}_{\zeta_i}-\frac{1}{N}\sum_{\tilde{\zeta}_j}^{N}\boldsymbol{f}_{\tilde{\zeta}_j}-2\lambda\boldsymbol{\theta},$$

参数 $\boldsymbol{\theta}$ 的更新公式为

$$\boldsymbol{\theta}=\boldsymbol{\theta}+\alpha\nabla_{\boldsymbol{\theta}}J(\boldsymbol{\theta}),$$

式中,α 表示学习率。

经过上述过程的迭代,即可求得最优参数 $\boldsymbol{\theta}$,从而得到最优的奖励函数。然而,上述算法如果被运用,还需要得到专家示例库,才可通过与专家策略进行对比求得奖励函数的参数。

2.6　构建 IRL 专家示例

在逆强化学习中,专家示例的选择会直接影响策略的学习,因此,如何构建专家示例十分关键。Dijkstra 算法使用了广度优先搜索解决赋权有向图或者无向图的单源最短路径的问题,在路径规划问题中得到了广泛的应用[23,24]。考虑到 Dijkstra 算法能生成最短路径的特点,本文将 Dijkstra 算法生成的路径设置为专家示例。然而,传统的 Dijkstra 算法只考虑最短路径的问题,而无法同时考虑选择充电点以及计算充电时间等充电相关动作,因此需要在传统 Dijkstra 算法上做出一定改进。在本文中,给定起点、终点对,需要先利用传统 Dijkstra 算法得到最短路径,再考虑该条最短路径有哪些点在充电桩节点的集合中,由这些点作为分割点得到子路径 Sub_path。对于每条子路径 $\mathrm{Sub\_path}_i$ 的耗电量情况进行分析,充电指示 charge_index 取值如下:

$$\mathrm{charge\_index}=\begin{cases}1,\text{if use\_power}(\mathrm{Sub\_path}_i)>\mathrm{Soc}_{i-1}-\mathrm{limit}_{\mathrm{batt}}\\0,\text{otherwise}\end{cases},$$

式中,use_power($\mathrm{Sub\_path}_i$)表示子路径 $\mathrm{Sub\_path}_i$ 的耗电量;Soc_{i-1} 表示电动汽车在最短路径中第 $i-1$ 个充电桩节点时的剩余电量;$\mathrm{limit}_{\mathrm{batt}}$ 表示电动汽

车的最低电量,本文设定 $\mathrm{limit}_{\mathrm{batt}}=0.1$。charge_index=1 时表示需要充电,即对于某一个路径中的充电桩节点,若下一条子路径耗电量大于剩余电量与最低电量之间的差值,则在该充电桩节点需要充电。

对于充电的目标电量,本文采取部分充电的策略[25]:若下一条子路径的耗电量小于等于 0.7,则在此充电桩将电量充到 0.8;若下一条子路径的耗电量大于 0.7,则此次充电需要将电量充到 1。此外,为了更接近现实中的情况,本文也考虑了分段充电的策略,从电量 0 充到目标电量 x 的充电时间 charge_time 如下:

$$\text{charge\_time}=\begin{cases}\dfrac{x}{0.8\times0.01}, & x<0.8\\ \dfrac{x-0.8}{0.5\times0.01}+\dfrac{0.8}{0.8\times0.01}, & 0.8\leqslant x\leqslant1\end{cases},$$

式中,电动汽车的电量达到 0.8 之前的充电效率高于其电量达到 0.8 之后的充电效率,这与现实中的设定相同。为简化运算,本文将电量达到 0.8 之后所需的充电时间也用线性函数表示。关于电量的充电时间函数如图 5 所示。

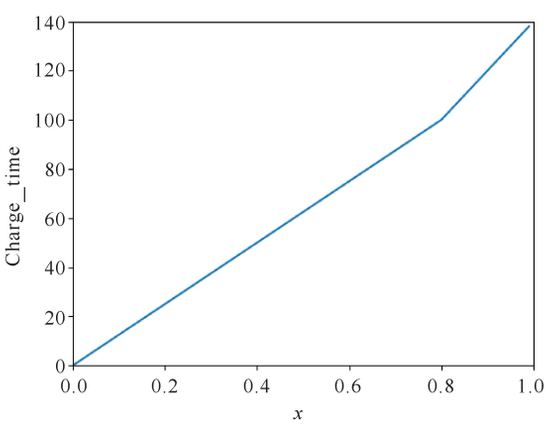

图 5 充电时间函数

Fig. 5 Function of charging time

因此,一条包含充电策略的最短路径可被求得,并将此路径作为专家示例。Dijkstra 算法在较小的路网中有较好的性能,然而当其运用在大型路网中,该算法需要花费很长的时间得到邻接矩阵,且它的运行效率与路网规模成正比。而本文基于逆强化学习的方法是在专家示例中学习其行走与充电的策略,这种策略对于不同的路网或大型路网也同样适用,而无需对模型重新进行训练。

3 实验验证

3.1 数据集

选用两个路网分别对模型进行训练,两个路网均是北京市路网中的一部分。路网Ⅰ包含 76 个节点以及 105 条边,纬度为 39.893°-39.898° N,经度为 116.390°-116.405° E。路网Ⅱ包含 75 956 个节点以及 101 052 条边,纬度为 39.4°-40.4° N,经度为 115.8°-117.0° E。路网的数据集由两部分构成:点集与边集,其中点集中包含各点的经纬度信息,边集中包含每条路径起点、终点的经纬度以及起点、终点之间的距离,由点集与边集构成的路网如图 6 所示。由于充电桩节点的数据集未公开,且关于充电桩的布局不是本文的研究重点,因此随机选择点集中的点作为充电桩,这对实验结果并不影响。

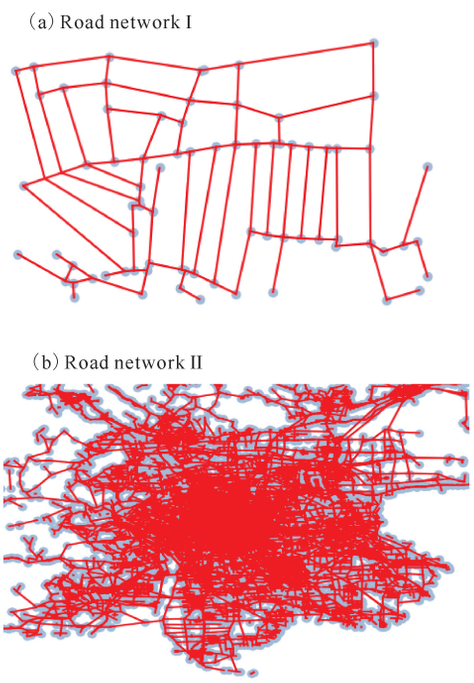

图 6 北京市部分路网

Fig. 6 Part of road network in Beijing

3.2 基准方法

通过将所提出的方法与以下基准方法进行比较,开展性能验证。

(1)Dijkstra & 完全充电策略(Dijkstra & Full charge)。Dijkstra 算法在小型路网中能够高效地得到最短路径,但未考虑充电策略。如 2.6 节所示,此基准方法对 Dijkstra 进行改进并采用完全充电策略,即每次在充电桩节点充电时都将电量充满,以验证部分充电策略的有效性。

(2)RL & Dueling DQN。与本文相同,此方法结合 Dueling DQN 算法对 Q 值进行近似。然而,强化学习中的奖励凭借人为经验进行设置,用此基准方法可验证逆强化学习算法获得的奖励函数的优越性。

(3)RL & DQN。与方法(2)相同,此方法的奖励同样人为给定,且与方法(2)设置相同的奖励,并通过经典的 DQN 算法对 Q 值进行近似,此方法用来验证 Dueling DQN 算法的有效性。

3.3 评价标准

本文从有效性和高效性两方面对模型进行评价。在有效性评估方面,对于电动汽车的出行规划,行驶时间(T_{path})以及充电时间(T_{ch})两个基本评价指标尤为重要,直接影响电动汽车的出行效率。此外,充电频率(Frq_{ch})高会导致充电服务时间长,因此充电频率也是一个重要的评价指标,本文用路径中包含的充电次数对该指标进行衡量。除上述 3 个评价指标外,由于用户在行驶过程中希望尽可能避免绕路,因此绕路次数也值得关注,这一指标用 loop 值衡量。因此本文对于有效性的评价指标包括:行驶时间、充电时间、充电频率、绕路次数。本文用 Gap 来量化所提出的方法与基准方法之间的差异,前两个指标的 Gap 表示为

$$Gap(T_{path}) = \frac{ExistF(T_{path}) - PropF(T_{path})}{PropF(T_{path})} \times 100\%,$$

$$Gap(T_{ch}) = \frac{ExistF(T_{ch}) - PropF(T_{ch})}{PropF(T_{ch})} \times 100\%,$$

式中,ExistF 表示基准方法的性能,PropF 表示本文提出的方法的性能;$Gap_{T_{path}}$ 和 $Gap_{T_{ch}}$ 分别表示两种方法之间行驶时间以及充电时间的差异。由于后面两个指标中 PropF 的值可能为 0,因而直接用两种方法的差值表示 Gap,类似地可表示为

$$Gap(Frq_{ch}) = ExistF(Frq_{ch}) - PropF(Frq_{ch}),$$

$$Gap(loop) = ExistF(loop) - PropF(loop)。$$

Gap 的值越大,说明本文提出的方法性能相对基准方法越好。

对于高效性的评估,使用运行模型所需的时间作为评价指标在不同的模型之间进行比较。模型运行时间越短,说明模型越高效。

3.4 实验结果

选取 8 对不同方向的起点、终点对测试各个方法的性能,其中 $\overline{T}_{path}$、$\overline{T}_{ch}$、$\overline{Frq_{ch}}$、$\overline{loop}$ 分别表示 8 条路径的平均行驶时间、平均充电时间、平均充电频率以及平均绕路次数,利用这些指标对模型的有效性进行验证。路网Ⅰ、Ⅱ中 4 个评估指标下所有模型的结果如表 1 所示。

由表 1 可知,由于 Dijkstra 算法一定能得到最短路径且不存在绕路情况,因此 Dijkstra & 完全充电策略在行驶时间和绕路方面在所有方法中表现最佳,而本文方法在行驶时间方面与 Dijkstra 算法的差距很小,且同样不存在绕路情况。在充电性能方面,由于 Dijkstra & 完全充电策略不考虑电量的可用性,相对本文方法表现较差,说明部分充电策略在提高充电效率上具有有效性。此外,Dijkstra 算法的充电次数也多于本文方法,说明本文方法在决策时考虑增加路径代价以节约充电代价,从而使得整个旅程的代价(行驶时间与充电时间相加)最小,这一点可以通过将前两个指标相加来证明。Dijkstra & 完全充电策略的 $\overline{T}_{path} + \overline{T}_{ch}$ 大于本文方法,就整个旅程的代价而言,本文方法优于 Dijkstra & 完全充电策略。对于 RL & Dueling DQN 以及 RL & DQN 两种强化学习方法,其行驶时间以及充电两方面的性能表现均不佳,且部分路径存在绕路情况,原因可能是奖励凭借人为经验设定,在复杂问题中难以考虑全面。此外,使用 Dueling DQN 的方法表现比使用 DQN 的方法好,说明 Dueling DQN 算法的有效性。综上可知,相较于基准方法,结合逆强化学习算法与 Dueling DQN 的模型,在对电动汽车进行出行规划时,可以得到使整个旅程代价最小的出行方案,证明了本文模型的有效性。

通过 Gap 详细对比不同方法生成的每条路径,将起点、终点对相同的路径归为一组,路网Ⅰ中的结果对比如图 7 所示,路网Ⅱ中的结果对比如图 8 所示。本文模型在绝大多数路径中表现最为优秀。

为验证模型的高效性,对路网Ⅰ、Ⅱ模型的运行时间进行对比,如图 9 所示。在高效性验证中,Dijkstra 算法在较小的路网中运行时间略微短于强化学习以及逆强化学习模型。然而 Dijkstra 算法的运行时间与路网规模强相关,因此在大型路网中,Dijkstra 算法的运行时间远远长于强化学习模型与逆强化学习模型。这也证明了本文模型在大型路网中的高效性。

表 1　路网Ⅰ及路网Ⅱ中本文方法与基准方法的效果比较

Table 1　Comparison among the proposed method and the baseline methods in road network Ⅰ and road network Ⅱ

路网 Road network	方法 Methods	$\overline{T}_{path}$	$\overline{T}_{ch}$	$\overline{Frq}_{ch}$	$\overline{loop}$
Ⅰ	Dijkstra & Full charge	**773.921**	235.348	1.875	**0**
	RL & Dueling DQN	843.072	248.438	2.500	0.750
	RL & DQN	858.768	259.819	2.750	1.125
	Proposed method	778.904	**212.078**	**1.750**	**0**
Ⅱ	Dijkstra & Full charge	**1 987.090**	608.630	5.750	**0**
	RL & Dueling DQN	2 036.660	627.320	6.750	2.375
	RL & DQN	2 063.460	639.570	7.375	3.250
	Proposed method	1 993.900	**592.360**	**5.375**	**0**

Note: The best results of each evaluation criterion in the test set are shown in bold

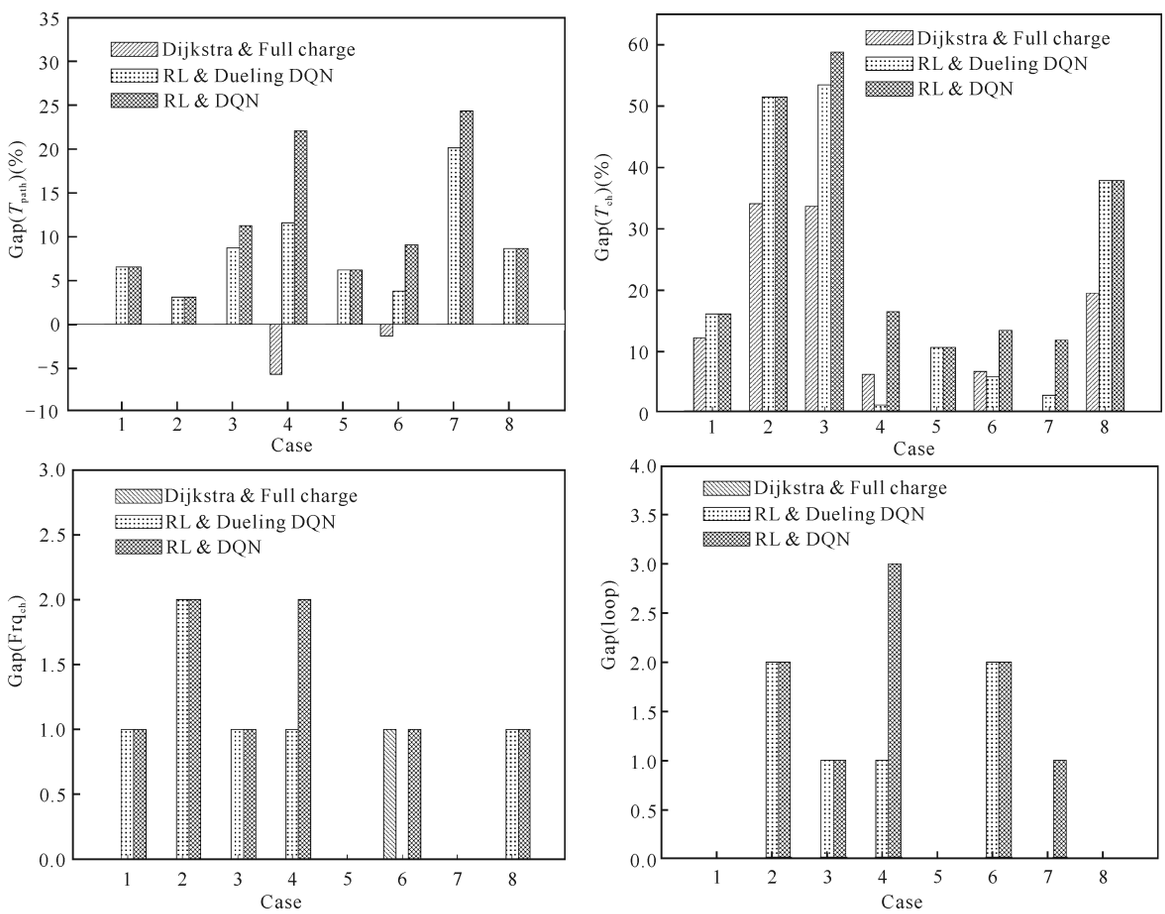

图 7　路网Ⅰ中不同方法生成的路径结果对比

Fig. 7　Comparison of path results generated by different methods in road network Ⅰ

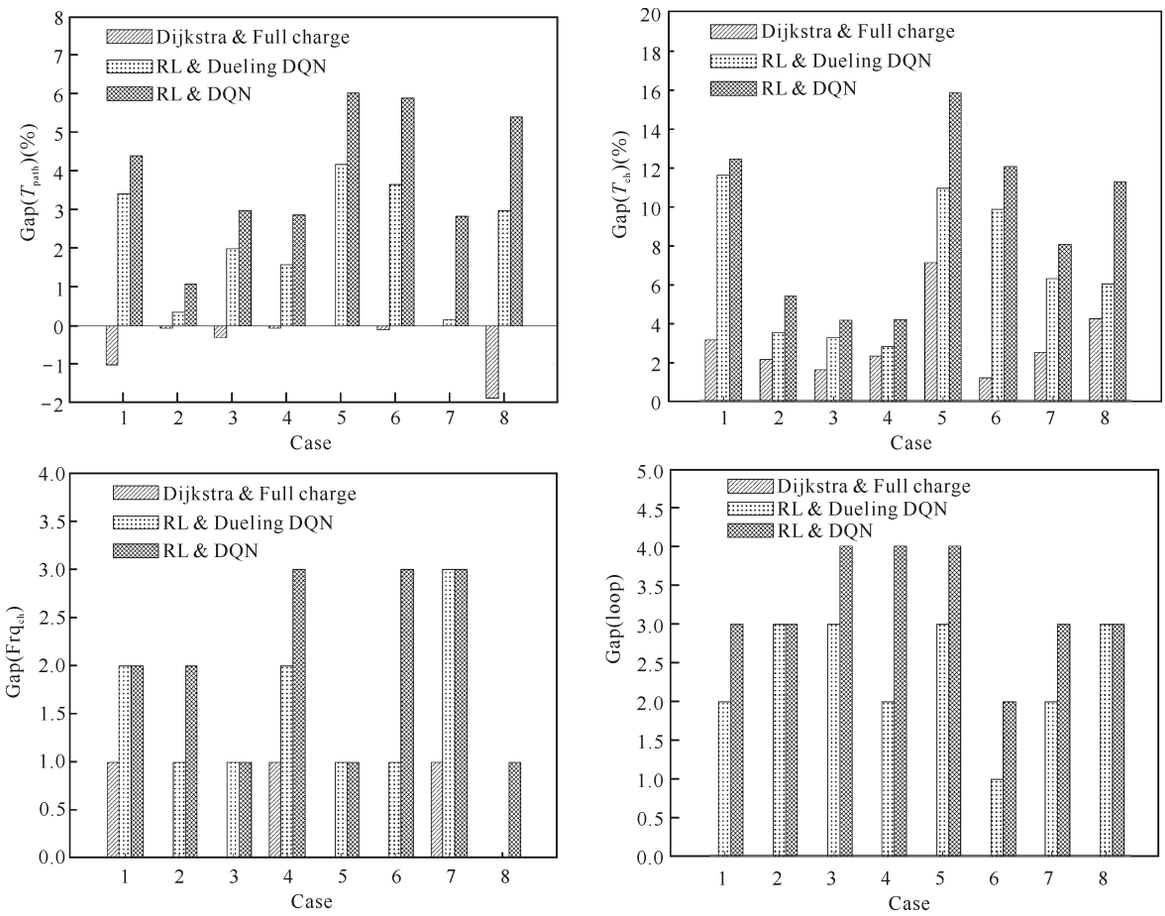

图 8　路网Ⅱ中不同方法生成的路径结果对比

Fig. 8　Comparison of path results generated by different methods in road network Ⅱ

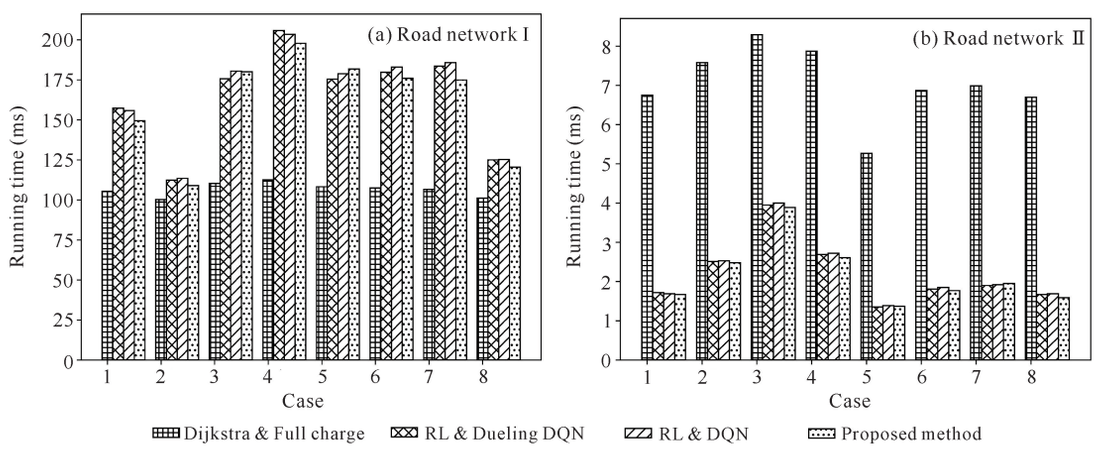

图 9　模型运行时间对比

Fig. 9　Comparison of model running time

该模型的收敛性如图 10 所示,在训练一定轮数后每一轮获得的奖励值趋于稳定且神经网络的参数也逐渐收敛。

此外,训练出来的模型也能较好地直接应用在其他路网上而无需重新训练,而 Dijkstra 算法不具备此特点。处理不同的路网时,Dijkstra 算法需要重新进行数据预处理而不能将原有模型直接应用在新路网中,即不具备迁移性。本文在两个未经训练的新路网中运行模型用以验证迁移性,在路网Ⅲ和路网Ⅳ中随机设置起点、终点对运行模型,观察能否得到可达路径,利用可达路径的比例评估不同模型的迁移性。在两个路网中各模型的可达路径比例如表 2 所示,本文

模型在两个路网中均表现最佳。

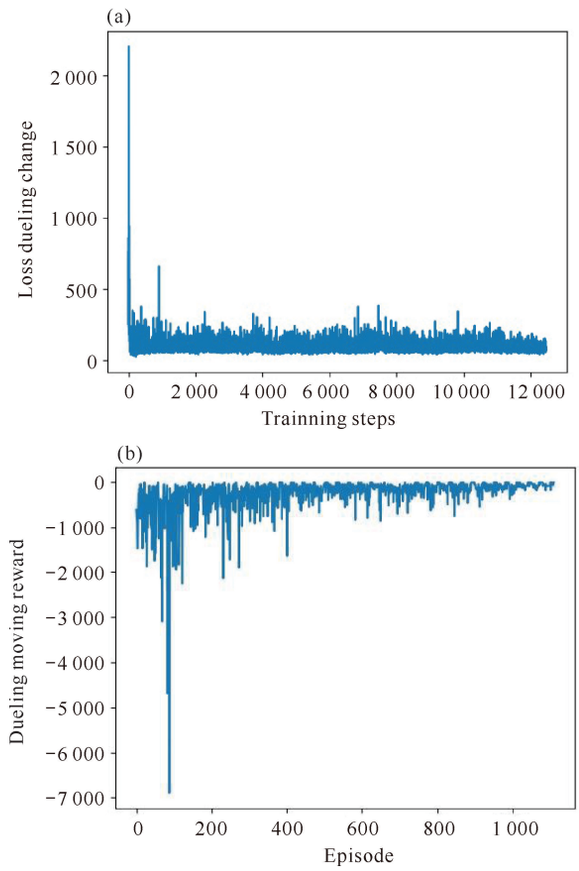

图 10　模型收敛性

Fig. 10　Convergence of model

表 2　模型迁移性对比

Table 2　Comparison of model migration

路网 Road network	Model migration (%)		
	RL & Dueling DQN	RL & DQN	Proposed method
Ⅲ	74	68	**86**
Ⅳ	70	62	**80**

Note: The best results of each evaluation criterion in the test set are shown in bold

4　结论

本文将考虑充电行为的 Dijkstra 算法生成的轨迹作为专家示例,使用 Dueling DQN 算法对 Q 值近似,提出了一种基于逆强化学习的电动汽车出行规划方法。该方法能够有效引导电动汽车用户选择一条可达目的地,并且行驶时间短、充电时间短、充电频率少的路线行走。通过引入分段充电以及部分充电策略,使研究场景更接近现实情况,从而有效地提高了充电效率;并利用北京市真实路网中的部分路网图对模型进行评估,结果表明,本文提出的方法优于其他基准方法。此外,本文方法具有很好的迁移性,可以有效地应用在其他未经训练的路网中。在后续的研究中,基于现有方法策略以及实验结果,本文模型可引入非线性奖励函数形式,进一步挖掘特征向量之间的联系以得到更适用的奖励函数,从而提升模型的性能。

参考文献

[1] 赵云峰,杨武双,李榕杰,等. 我国纯电动汽车发展趋势分析[J]. 汽车工程师,2020(7):14-17.

[2] 孙叶,刘锴. 里程焦虑对纯电动汽车使用意愿的影响[J]. 武汉理工大学学报(交通科学与工程版),2017,41(1):87-91.

[3] CUCHÝ M,JAKOB M. Electric vehicle travel planning with lazy evaluation of recharging times [C]//2019 IEEE International Conference on Systems, Man and Cybernetics (SMC). Bari,Italy:IEEE,2019:3168-3173.

[4] 郭戈,张振琳. 电动车辆路径优化研究与进展[J]. 控制与决策,2018,33(10):1729-1739.

[5] LU J,CHEN Y N,HAO J K,et al. The time-dependent electric vehicle routing problem: Model and solution [J]. Expert Systems with Applications, 2020, 161: 113593. DOI:10.1016/j.eswa.2020.113593.

[6] LEBEAU P,DE CAUWER C,VAN MIERLO J,et al. Conventional,hybrid,or electric vehicles:Which technology for an urban distribution centre? [J]. The Scientific World Journal,2015,2015:302867. DOI:10.1155/2015/302867.

[7] FELIPE Á,ORTUÑO M T,RIGHINI G,et al. A heuristic approach for the green vehicle routing problem with multiple technologies and partial recharges [J]. Transportation Research Part E: Logistics and Transportation Review,2014,71:111-128.

[8] CANDRA A,BUDIMAN M A,HARTANTO K. Dijkstra's and A-Star in finding the shortest path:A Tutorial [C]//2020 International Conference on Data Science, Artificial Intelligence, and Business Analytics (DATABIA). Medan,Indonesia:IEEE,2020:28-32.

[9] KOBAYASHI Y,KIYAMA N,AOSHIMA H,et al. A route search method for electric vehicles in consideration of range and locations of charging stations [C]//2011 IEEE Intelligent Vehicles Symposium (Ⅳ). Baden-Baden,Germany:IEEE,2011:920-925.

[10] WANG Y,JIANG J M,MU T T. Context-aware and energy-driven route optimization for fully electric vehicles via crowdsourcing [J]. IEEE Transactions on In-

telligent Transportation Systems, 2013, 14(3): 1331-1345.

[11] YANG C, TANG J R, SHEN Q. Impact of electric vehicle battery parameters on the large-scale electric vehicle charging loads in power distribution network [C]//2020 16th International Conference on Control, Automation, Robotics and Vision (ICARCV). Shenzhen, China: IEEE, 2020: 56-60.

[12] KESKIN M, AKHAVAN-TABATABAEI R, ÇATAY B. Electric vehicle routing problem with time windows and stochastic waiting times at recharging stations [C]//2019 Winter Simulation Conference (WSC). National Harbor, MD, USA: IEEE, 2019: 1649-1659.

[13] HANAWAL M K, LIU H, ZHU H H, et al. Learning policies for Markov decision processes from data [J]. IEEE Transactions on Automatic Control, 2019, 64(6): 2298-2309.

[14] HE M X, LU D J, TIAN J, et al. Collaborative reinforcement learning based route planning for cloud content delivery networks [J]. IEEE Access, 2021, 9: 30868-30880.

[15] CHREN W A. One-hot residue coding for high-speed non-uniform pseudo-random test pattern generation [C]//Proceedings of ISCAS'95-International Symposium on Circuits and Systems. Seattle, WA, USA: IEEE, 1995, 1: 401-404.

[16] ABBEEL P, NG A Y. Apprenticeship learning via inverse reinforcement learning [C]//Proceedings of the 21st International Conference on Machine Learning. Banff, Canada: ACM, 2004.

[17] QIU H M, LIU F. A state representation dueling network for deep reinforcement learning [C]//2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI). Baltimore, MD, USA: IEEE, 2020: 669-674.

[18] WINARNO E, HADIKURNIAWATI W, ROSSO R N. Location based service for presence system using haversine method [C]//2017 International Conference on Innovative and Creative Information Technology (ICITech). Salatiga, Indonesia: IEEE, 2017: 1-4. DOI: 10.1109/INNOCIT.2017.8319153.

[19] ABOUSLEIMAN R, RAWASHDEH O. A Bellman-Ford approach to energy efficient routing of electric vehicles [C]//2015 IEEE Transportation Electrification Conference and Expo (ITEC). Dearborn, MI, USA: IEEE, 2015: 1-4.

[20] YI L M. Lane change of vehicles based on DQN [C]//2020 5th International Conference on Information Science, Computer Technology and Transportation (ISCTT). Shenyang, China: IEEE, 2020: 593-597.

[21] SHARMA J, ANDERSEN P A, GRANMO O C, et al. Deep Q-learning with Q-matrix transfer learning for novel fire evacuation environment [J]. IEEE Transactions on Systems, Man and Cybernetics: Systems, 2021, 51(12): 7363-7381.

[22] SONG S N, FANG Z Y, ZHANG Z Y, et al. Semi-online computational offloading by Dueling Deep-Q network for user behavior prediction [J]. IEEE Access, 2020, 8: 118192-118204.

[23] FAN D K, SHI P. Improvement of Dijkstra's algorithm and its application in route planning [C]//2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery. Yantai, China: IEEE, 2010, 4: 1901-1904.

[24] KESSWANI N. Performance evaluation of shortest path routing algorithms in real road networks [C]//Proceedings of the International Conference on Data Engineering and Communication Technology. Singapore: Springer, 2017: 77-83.

[25] DESAULNIERS G, ERRICO F, IRNICH S, et al. Exact algorithms for electric vehicle-routing problems with time windows [J]. Operations Research, 2016, 64(6): 1388-1405.

Research on Electric Vehicle Travel Planning Based on Inverse Reinforcement Learning

LI Fanyu,ZHANG Ying,HUA Yunpeng,LI Muyang,CHEN Yuanchang
(School of Control and Computer Engineering,North China Electric Power University,Beijing,102206,China)

Abstract: With the popularization of electric vehicles,the research on travel planning of electric vehicles is particularly important. Travel planning,which is different from path planning,needs to consider both path and charging problems. This article proposes a travel planning method for Electric Vehicles Travel Planning (EVTP) based on Inverse Reinforcement Learning (IRL),which can effectively plan an accessible path for electric vehicle users with a short driving path and a short charging time. The Dijkstra algorithm was improved to obtain the shortest path considering the charging behavior,which was input into the inverse reinforcement learning algorithm as an expert example. The inverse reinforcement learning algorithm was used to obtain both walking and charging rewards. In learning strategy,Dueling DQN algorithm was used to update Q-value efficiently and improve learning performance. Partial charging strategies and segmented charging strategies were adopted to improve the charging efficiency and make the research closer to the real situation. The working performance and results of the model were analyzed in detail and compared with the benchmark method. The results show that the travel planning method of electric vehicles based on inverse reinforcement learning has better performance in both driving time and charging time. Meanwhile,our method has very good performance in portability.

Key words: inverse reinforcement learning; electric vehicle; travel planning; Dueling DQN; partial charging strategies

责任编辑:唐淑芬