基于 Q 学习的受灾路网抢修队调度问题建模与求解

苏兆品 [1, 2, 3] 李沫晗 [1] 张国富 [1, 2, 3] 刘 扬 [1]

摘 要 受损路网的修复是灾害应急响应中的一个重要环节, 主要研究如何规划道路抢修队的修复活动, 为灾后救援快速打通生命通道. 本文首先构建了抢修队修复和路线规划的数学模型, 然后引入马尔科夫决策过程来模拟抢修队的修复活动, 并基于 Q 学习算法求解抢修队的最优调度策略. 对比实验结果表明, 本文方法能够让抢修队从全局和长远角度实施受损路段的修复活动, 在一定程度上提高了运输效率和修复效率, 可以为政府实施应急救援和快速安全疏散灾民提供有益的参考.

关键词 灾害应急响应, 受损路网, 抢修队调度, 马尔科夫决策过程, Q 学习

引用格式 苏兆品, 李沫晗, 张国富, 刘扬. 基于 Q 学习的受灾路网抢修队调度问题建模与求解. 自动化学报, 2020, **46**(7): 1467−1478

DOI 10.16383/j.aas.c180081

Modeling and Solving the Repair Crew Scheduling for the Damaged Road Networks Based on Q-Learning

SU Zhao-Pin[1, 2, 3] LI Mo-Han[1] ZHANG Guo-Fu[1, 2, 3] LIU Yang[1]

Abstract Repairing the damaged road network is an important issue in disaster emergency response, which mainly focuses on how to schedule the repair activities of the repair crew to clear the life path as soon as possible. In this paper, we first present a mathematical model of the repairing and scheduling of the repair crew. Next, we introduce the Markov decision process to simulate the repair activities of the repair crew. Additionally, the Q-learning algorithm is proposed to search for the optimal scheduling strategy of the repair crew for the damaged road network. Finally, the comparative experimental results demonstrate that the proposed approach is able to make the repair crew repair the damaged road network from the global and long-term perspective, improves the transport and repair efficiencies to a certain extent, and provides a useful reference for our government to carry out the emergency rescue and evacuate the victims as quickly and safely as possible.

Key words Disaster emergency response, damaged road network, repair crew scheduling, Markov decision process, Q-learning

Citation Su Zhao-Pin, Li Mo-Han, Zhang Guo-Fu, Liu Yang. Modeling and solving the repair crew scheduling for the damaged road networks based on Q-learning. *Acta Automatica Sinica*, 2020, **46**(7): 1467−1478

伴随着全球气候变化以及我国经济快速发展和城市化进程不断加快, 我国的资源、环境和生态压力日益加剧. 各类自然灾害多发频发, 给我国经济和商业社会造成的损害也日益严重[1−2]. 仅在 “十二五” 期间, 年均造成 3.1 亿人次受灾, 因灾死亡失踪 1 500 余人, 紧急转移安置 900 多万人次, 倒塌房屋近 70 万间, 农作物受灾面积 2 700 多万公顷, 直接经济损失达 3 800 多亿元. 为此, 国务院在《“十三五” 国家科技创新规划》和《国家综合防灾减灾规划 (2016–2020 年)》中明确提出了 “要提高防灾减灾救灾工作规范化、现代化水平, 强化科技创新, 有效提高防灾减灾救灾科技支撑能力和水平”.

在灾害应急响应中[3−4], 及时修复受损路网、打通生命通道是开展灾后救援工作的一个重要环节[5]. 主要研究如何利用智能决策理论和计算机辅助工具规划道路抢修队的调度, 即如何修复路段、修复哪些路段可以使得路网的修复时间最小、运输效率最大, 这对应急救援的实施和灾民的快速安全疏散具有重

收稿日期 2018-02-02 录用日期 2018-05-07
Manuscript received February 2, 2018; accepted May 7, 2018
国家自然科学基金 (61573125), 安徽省重点研究与开发计划 (202004d07020011), 中央高校基本科研业务费专项资金 (PA2020GDKC0015, PA2019GDQT0008, PA2019GDPK0072) 资助
Supported by National Natural Science Foundation of China (61573125), Anhui Provincial Key Research and Development Program (202004d07020011), Fundamental Research Funds for the Central Universities (PA2020GDKC0015, PA2019GDQT0008, PA2019GDPK0072)
本文责任编委 张敏灵
Recommended by Associate Editor ZHANG Min-Ling
1. 合肥工业大学计算机与信息学院 合肥 230601 2. 工业安全与应急技术安徽省重点实验室 合肥 230601 3. 安全关键工业测控技术教育部工程研究中心 合肥 230601
1. School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230601 2. Anhui Province Key Laboratory of Industry Safety and Emergency Technology, Hefei 230601 3. Engineering Research Center of Safety Critical Industrial Measurement and Control Technology, Ministry of Education, Hefei 230601

要的现实意义. 因此, 近年来, 灾后路网的修复问题越来越受到各国政府和学者的重视和关注[6].

Chen[7] 将应急抢修的后勤保障调度构建为一个整数多商品的网络流问题, 并设计了一种基于问题分解和变量固定技术的启发式算法, 可在约束的运营时间内降低后勤保障的短期运营成本. 霍建顺[8] 针对道路抢修的不确定性, 基于模糊机会约束规划和模糊比较排序确定震后道路抢修排程. Nurre 等[9] 针对在极端破坏下的基础设施系统恢复服务的问题, 提出了一种新的启发式调度规则用来分配工作组将节点和弧建立到网络中, 以最大限度地提高网络中的累积加权流量. 但是, 这种路网修复是以增加新的点和边为代价, 成本较大.

Yan 等[10−11] 基于应急抢修和救灾物资分配的双层时空网络, 将道路抢修和物资分配构建为一个多目标、混合整数、多商品的网络流问题, 并引入蚁群优化搜索最优抢修、物资分配路线和时间表. 花丙威等[12] 基于灾后路网状态识别提出了路网脆弱性评价指标, 并构建了基于路网脆弱性和路径出行费用的双层规划模型. 邱慧[13] 将灾后公路网修复分为应急修复期和全面修复期, 在时间和设备资源的约束下, 构建了基于最大连通子图的大小和公路网加权效率的两阶段模型. Aksu 等[14−15] 研究灾后道路障碍清理规划问题, 提出了一个基于动态路径的数学模型来识别道路阻塞的临界性, 并以有限的设备资源清除路上障碍, 随后设计了一个相应的启发式方法, 以最大化整个路网的连通性和最小化路障清理时间. Kasaei 等[16−17] 基于弧路由问题来规划灾后道路的障碍清除, 设计了基于整数规划和启发式的混合算法以最小化路网重新连通的时间开销或在给定的时间约束内最大化路网重新连通后的的总收益. 不过, 上述方法均是着眼于路网本身, 构建的路网大都过于理想化, 仅考虑修复路网中的哪些路段可以实现目标的最优化, 而没有考虑这些受损路段是否可达和修复工程队的在危险环境中的路线问题, 也没有考虑受损路段的修复顺序对应急救援的影响.

李爱庆[18] 基于用多目标的受损道路抢修与紧急物资配送的混合整数多重网络规划模型求解各受损道路的抢通时间, 各个工作队的抢修路径. Duque 等[19] 从道路抢修队 (包含人员、设备和原材料等) 自身的视角出发, 针对灾后道路抢修队的调度和路线规划问题, 提出了一个较为完整的数学模型, 不仅考虑了应该修复哪些受损边, 还考虑了修复时间和修复路线的影响, 并基于动态规划 (Dynamic programming, DP) 和贪婪策略优化路网中那些需要救援的地点的可达性. Kim 等[20] 考虑额外损失和可变损失率等灾害因素下的道路抢修队调度问题, 以尽量减少因受灾点不可达而造成的总损失. 上述工作虽然考虑了抢修队的调度路径, 并能够给出道路抢修队的修复策略集, 但在其数学模型中, 应急需求点和受损路段均是以受损节点来表示, 不能明确表现路网结构, 也不利于系统演示. 此外, Duque 等[19] 提出的 DP 算法是通过每一步的贪婪策略选择局部最优的行为来逐步达到全局最优, 因此, 其算法往往容易陷入局部最优解而无法跳出.

基于上述背景, 本文在整理和分析已有工作的基础上, 首先, 构建了考虑连续受损路段的抢修队调度和路线规划问题的数学模型; 然后, 基于马尔科夫决策过程[21] 来模拟抢修队的修复活动, 并基于 Q 学习算法[22] 求解道路抢修队的最优调度策略; 最后, 通过对比实验验证了本文方法的有效性.

1 问题描述

考虑如下路网 $G = \{V, E\}$, V 为 G 中的道路节点序号集合, $E = V \times V$ 为 G 中的所有路段的集合.

如图 1 所示, V 中包含一个救援物资的储备点序号和 $n \in \mathbf{N}$ 个需求节点的序号. 不失一般性, 我们用节点 "0" 表示储备点, 用 $V^* = \{1, 2, \cdots, n\}$ 表示需求节点的集合 (所有的非需求节点已经从路网中剔除), 则 $V = V^* \cup \{0\}$. $\forall i \in V^*$ 为一个急需救援的应急点, 代表某城市、村庄或社区等. 在实际的灾害应急响应中, 由于每一个应急点的受灾程度不同, 对救援需求的迫切性也不同. 因此, 我们用 $I_i \in \mathbf{R}$ 表示和区分每个应急点 i 的受灾程度.

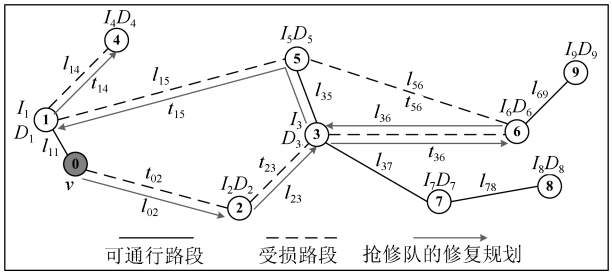


图 1 受损路网示意图

Fig. 1 Schematic diagram of the damaged road network

在应急这个特殊场景中, E 中的每条路段 $e_{ij} \in E, \forall i, j \in V$ 是双向通行的, 且每条路段 e_{ij} 有一个长度 $l_{ij} \in \mathbf{R}$. 此外, 我们用 $\xi_{ij} \in \{0, 1\}$ 来表示每条路段 e_{ij} 是否是畅通的. 如果 $\xi_{ij} = 1$, 则路段 e_{ij} 是可通行的; 否则, e_{ij} 目前为一条受损路段. 基于此, E 可以划分为如下两部分: $E_c = \{e_{ij} | e_{ij} \in E, \xi_{ij} = 1, \forall i, j \in V\}$ 为 E 中所有可通行路段的集合; $E_r = \{e_{ij} | e_{ij} \in E, \xi_{ij} = 0, \forall i, j \in V\}$ 为 E 中所有受损路段的集合. 显然, $E = E_c \cup E_r$.

如果路网 G 中任一节点 i 到任一节点 j 之间

是可通行的, 则我们用 $\mathcal{D}_{ij} \in \mathbf{R}$ 表示节点 i 到节点 j 之间的最短路径的长度, 可由经典的 Dijkstra 算法[23] 根据每条路段的长度计算得到. 特别的, $\mathcal{D}_{i0}$ 为应急点 i 到储备点 0 的最短路径的长度. 需要注意的是, 救援工作必须在一定的时限内完成, 比如黄金救援 72 小时等. 因此, 对于每个应急点 $i \in V^*$, 有一个最大可接受距离 $D_i \in \mathbf{R}$ 的约束. 基于此, 我们用 $\delta_i \in \{0,1\}$ 来表示储备点 0 到应急点 i 是否可达. 如果储备点 0 到应急点 i 是连通的, 而且其距离 $\mathcal{D}_{i0} \leq D_i$, 则 i 是可达的, $\delta_i = 1$; 如果储备点 0 到应急点 i 是不连通的, 或者即使储备点 0 到应急点 i 是连通的, 但 $\mathcal{D}_{i0} > D_i$, i 是不可达的, $\delta_i = 0$, 因为将会错过应急点 i 的最佳救援时间.

道路抢修队 (包含人员、设备和原材料等) 将从储备点 0 出发对受损路网进行一系列的修复活动, 以使储备点 0 到各个应急点 $i \in V^*$ 都是可达的. 抢修队在高速路和山路的前进速度显然是不一样的, 和已有工作一样[19], 假设抢修队在各种路段上的平均前进速度为 $v \in \mathbf{R}$, 修复任一受损边 $e_{ij} \in E_r$ 的时间开销为 $t_{ij} \in \mathbf{R}$, 则抢修队在任意一条路段 $e_{ij} \in E$ 上的时间开销 $c_{ij} \in \mathbf{R}$ 为

$$c_{ij} = \begin{cases} \frac{l_{ij}}{v}, & \xi_{ij} = 1 \\ \frac{l_{ij}}{v} + t_{ij}, & \xi_{ij} = 0 \end{cases} \tag{1}$$

注意, 如果 e_{ij} 是受损路段, 那么 e_{ij} 在被修复后应该更新 $\xi_{ij} \leftarrow 1$.

抢修队一个完整的修复和路线规划可以用向量 $\boldsymbol{H} = \langle e_{i_1 j_1}, \cdots, e_{i_k j_k} \rangle$ $(k \in \mathbf{N})$ 来表示. 例如, 图 1 中的抢修队修复和路线规划为 $\boldsymbol{H} = \langle e_{02}, e_{23}, e_{36}, e_{63}, e_{35}, e_{51}, e_{14} \rangle$. 不同的 $\boldsymbol{H}$ 构成一个修复规划集 $\Theta = \{\boldsymbol{H}_1, \cdots, \boldsymbol{H}_m\}$ $(m \in \mathbf{N})$, 满足

$$|\Theta| = \sum_{i=0}^{|E_r|} P(E_r, i)$$

其中, $P(E_r, i)$ 为从受损路段集 E_r 中挑选 i 条路段的排列数.

对于一个可能的修复规划 $\boldsymbol{H} \in \Theta$, 首先需要考虑救灾物资的运输效率, 要求每个应急点 $i \in V^*$ 到储备点 0 的距离要尽可能短, 即

$$\min f_1(\boldsymbol{H}) = \sum_{\forall i \in V^*} (\mathcal{D}_{i0} \cdot I_i) \tag{2}$$

其中, $\mathcal{D}_{i0}$ 为在修复规划 $\boldsymbol{H}$ 下的应急点 i 到储备点 0 的最短路径长度, 例如, 图 1 中, $\mathcal{D}_{40} = l_{01} + l_{14}$, $\mathcal{D}_{90} = l_{02} + l_{23} + l_{36} + l_{69}$. 受灾程度 I_i 起到了一个偏好的作用, 对于受灾较严重的应急点 i, 输送距离 $\mathcal{D}_{i0}$ 要尽可能的短. 此外, 还需要考虑路网的修复效率, 要求每个应急点 $i \in V^*$ 要尽可能快地与储备点 0 连通, 以尽快打通生命通道, 即

$$\min f_2(\boldsymbol{H}) = \sum_{\forall i \in V^*} (\mathcal{C}_{i0} \cdot I_i) \tag{3}$$

其中, $\mathcal{C}_{i0} \in \mathbf{R}$ 为在修复规划 $\boldsymbol{H}$ 下, 应急点 i 与储备点 0 连通时抢修队的累积时间开销, 可由式 (1) 计算得到. 例如, 在图 1 中, $\mathcal{C}_{20} = l_{02}/v + t_{02}$, $\mathcal{C}_{40} = (l_{02}/v + t_{02}) + (l_{23}/v + t_{23}) + (l_{36}/v + t_{36}) + (l_{36}/v) + (l_{35}/v) + (l_{15}/v + t_{15}) + (l_{14}/v + t_{14})$.

基于上述考虑, 抢修队的修复和路线规划可描述为如下的约束优化问题:

$$\begin{cases} \min f(\boldsymbol{H}) = \lambda \cdot f_1(\boldsymbol{H}) + (1-\lambda) \cdot f_2(\boldsymbol{H}) \\ \text{s.t.} \qquad \mathcal{D}_{i0} \leq D_i, \forall\, i \in V^* \end{cases} \tag{4}$$

其中, $\lambda \in (0,1)$ 为加权因子, 控制着 f_1 和 f_2 在整体应急响应目标中的偏好.

2　马尔科夫决策模型

从前面的描述可以明显的看出, 抢修队的修复和路线规划是一个典型的序贯决策, 而且具有部分随机、部分可由决策者控制的动态特征, 这与马尔可夫决策过程 (Markov decision process, MDP)[21, 24−25] 具有天然的契合性. 因此, 我们从道路抢修队的视角出发, 把抢修队看成是一个智能体 (agent)[26], 而受损路网即是这个 agent 待感知的环境, 基于马尔可夫决策型来描述 agent 的决策过程.

如图 1 所示, agent 从储备点 0 出发, 沿着某一方向前进, 并决定是否修复受损路段 e_{ij}, agent 的动作空间 $\mathcal{A}$ 即为受损路段集合 E_r, 即 $\mathcal{A} = E_r$. 显然, $\mathcal{A}$ 是有限的. agent 的状态由三元组 $S = \langle \mathcal{V}, \mathcal{D}, \mathcal{E} \rangle$ 组成. 其中, $\mathcal{V} = \{i | \delta_i = 1, \forall i \in V^*\}$ 为已经可达的应急点列表; $\mathcal{D} = \{\langle i, \mathcal{D}_{i0} \rangle | \delta_i = 1, \forall i \in V^*\}$ 为已经可达的每个应急点 i 到储备点 0 的最短路径长度列表; $\mathcal{E} = \{e_{ij} | \xi_{ij} = 1, e_{ij} \in E_r\}$ 为已经修复的受损路段列表. 当 agent 决定并修复受损路段 e_{ij} 后, $\xi_{ij} \leftarrow 1$, agent 将确定 j 是否可达并更新 $\langle \mathcal{V}, \mathcal{D}, \mathcal{E} \rangle$, 然后将 e_{ij} 从动作空间 $\mathcal{A}$ 中剔除继续前进. 例如, 在图 1 中, agent 的初始状态为 $\langle \{1\}, \{\langle 1, l_{01} \rangle\}, \emptyset \rangle$, 动作空间为 $\{e_{02}, e_{23}, e_{36}, e_{56}, e_{15}, e_{14}\}$, 在执行动作 e_{02} 后状态为 $\langle \{1,2\}, \{\langle 1, l_{01} \rangle, \langle 2, l_{02} \rangle\}, \{e_{02}\} \rangle$, 动作空间为 $\{e_{23}, e_{36}, e_{56}, e_{15}, e_{14}\}$. 因为 $\mathcal{V}$、$\mathcal{D}$ 和 $\mathcal{E}$ 均是有限的, 所以 agent 的状态空间也是有限的.

agent 从状态 $s \in S$(不妨设此时 agent 位于应急点 $k \in V^*$) 决定执行动作 $e_{ij} \in \mathcal{A}$ 并转移到下一个状态 $s' \in S$. 注意, k 和 i 或 j 之间的最短路径可能不止一条路段, 比如图 1 中 agent 在应急点 6

选择动作 e_{15} 后到达应急点 1. 由 s 到 s', 可能发生如下变化: 1) $\mathcal{V}$ 中有新的应急点加入, 比如图 1 中 agent 从应急点 3 执行动作 e_{36} 后, 应急点 6 和 9 均可达; 2) $\mathcal{V}$ 没有变化, 但 $\mathcal{D}$ 发生变化, 即没有新的应急点可达, 但在已达的应急点中, 到储备点 0 的最短路径长度变短, 如图 1 中 agent 在应急点 6 选择动作 e_{15} 到达应急点 1 后没有新的应急点可达, 但应急点 5 到储备点 0 的距离变短.

对于第一种情形, 因为有新的可达应急点加入 $\mathcal{V}$, agent 执行动作 e_{ij} 在运输效率上获得的回报为

$$\mathcal{R}_1 = \sum_{i \in \mathcal{V}^{s'} - \mathcal{V}^s} \frac{I_i}{\mathcal{D}_{i0}^{s'}}$$

此外, agent 执行动作 e_{ij} 的时间开销为

$$c_{e_{ij}} = \frac{\min\{\mathcal{D}_{ki}, \mathcal{D}_{kj}\}}{v} + \frac{l_{ij}}{v} + t_{ij}$$

而 agent 从储备点 0 执行一系列动作到达状态 s' 的累积时间开销为

$$\mathcal{C}^{s'} = \sum_{e_{ij} \in \mathcal{E}^{s'}} c_{e_{ij}}$$

据此, agent 执行动作 e_{ij} 在修复效率上获得的回报为

$$\mathcal{R}_2 = \frac{\sum\limits_{i \in \mathcal{V}^{s'} - \mathcal{V}^s} I_i}{\mathcal{C}^{s'}}$$

对于第二种情形, 因为没有新的可达应急点加入 $\mathcal{V}$, 只有 $\mathcal{D}$ 中刷新了更短的 $\mathcal{D}_{i0}$. 需要注意的是, 设置这种情况下的 agent 回报, 不能直接使用减小的距离或增加的开销. 这是因为 s' 状态下 agent 所在的应急点 k 的可达性不是由当前执行 e_{ij} 这一个动作造成的, 而是之前一系列动作的结果. 因此, 我们设计 agent 执行动作 e_{ij} 在运输效率上获得的回报为

$$\mathcal{R}_1' = \sum_{i \in \mathcal{V}^s \wedge \mathcal{D}_{i0}^{s'} < \mathcal{D}_{i0}^s} \left(\frac{I_i}{\mathcal{D}_{i0}^{s'}} - \frac{I_i}{\mathcal{D}_{i0}^s} \right)$$

在修复效率上获得的回报为

$$\mathcal{R}_2' = \left(\frac{1}{\mathcal{C}^{s'}} - \frac{1}{\mathcal{C}^s} \right) \cdot \sum_{i \in \mathcal{V}^s \wedge \mathcal{D}_{i0}^{s'} < \mathcal{D}_{i0}^s} I_i$$

基于上述考虑, 并结合目标函数 (4), 我们设计 agent 执行动作 e_{ij} 后从状态 s 转移到状态 s' 的回报函数为:

$$\begin{aligned} \mathcal{R}(e_{ij}, s, s') = {} & \lambda \cdot (\mathcal{R}_1 + \mathcal{R}_1') + \\ & (1 - \lambda) \cdot (\mathcal{R}_2 + \mathcal{R}_2') \end{aligned} \tag{5}$$

其中, $\lambda \in (0, 1)$ 控制着在运输效率和修复效率上的偏好. 如果修复一个路段带来的应急点可达性改善很小而耗费的时间开销过大, 则有可能 $\mathcal{R}(e_{ij}, s, s') < 0$, 这意味着 agent 因为过度修复而受到了惩罚, 这样设计的目的是为了防止总的修复时间开销过大, 满足不了应急救援的需求. 此外, agent 在任意状态可以选择任何受损且未修复的路段进行动作, 在实际情况中, 可能被选中的这一受损路段的两个节点在当前状态下并不是连通的, 这时 agent 选择执行此动作的时间开销设置为 ∞, 促使 agent 接收到的回报趋于 0.

agent 的前进目标是使所有的应急点与储备点 0 连通并可达, 因此, 当 agent 的当前状态中 $\mathcal{V} = V^*$ 时, agent 会立即终止前进 (即 agent 达到终止状态), 而没必要修复所有的受损路段. 此外, 依据 MDP, agent 的策略模型中还有一个折扣因子 $\gamma \in [0, 1]$, 表示未来回报与当前回报之间的差异. 理想的情况下, agent 修复一个受损路段的效果应该放在整个路网中观测才最合理. 但是, 由 agent 状态空间可以看出, agent 只能获取路网的部分信息, 这就给 agent 的决策造成了困境. agent 可能要执行一系列动作之后才能看到明显的修复效果, 或者 agent 可能需要牺牲一部分当前回报来获取更好的长期回报, 这时通过折扣因子 γ, 在 agent 的当前回报中考虑未来的回报可以有效解决上述决策困境. agent 决策的目的就是在达到终止状态时能够最大化一系列动作 $\mathcal{E}$ 的累积回报

$$\mathcal{R}(|\mathcal{E}|) = \sum_{x=1}^{|\mathcal{E}|} [\gamma^x \cdot \mathcal{R}(e_{ij}, s, s')] \tag{6}$$

3 基于 Q 学习的最优调度策略求解

在本文的模型中, 只保留真实路网中的储备点节点、受灾区和潜在灾区的一些节点, 其他节点可以通过求最短路径方法剔除掉, 因此, 路网的规模经过处理后会大大减小. 然而, 路网信息是部分可观察的, agent 从储备点 0 出发时只能获取路网环境的部分信息 (即受损路段信息). 网络中有哪些路段是可通行的, 以及每个应急点之间的距离相对关系均是未知的. 对于这些未知信息, agent 只能在修复的过程中通过探索逐步获取. agent 的状态信息中, agent 维护一个当前执行过的动作的集合, 对于受损路网来说, 只要当前修复的受损路段已知 (无论受损路段的具体执行顺序如何), 受损路网未来的状态将只取决于当前的状态下, 接下来 agent 执行何种动作有关, 而与之前的状态的无关, 因此, agent 的决策满足马尔科夫性. 这就意味着 agent 在决策时有必要综合考虑以前的观测和当前的状态信息, 而

Q 学习 (Q-learning) 算法中的 agent 不需知道整体的环境, 仅需知道当前状态下可以选择哪些动作, 已成功应用于求解一些复杂的离散规划问题[22, 27−28]. 基于上述考虑, 本文引入 Q 学习算法求解抢修队的最优调度策略.

如图 2 所示, 基于 Q 学习的抢修队调度算法基本思路为: 基于式 (5) 设计 Q 学习中的即时奖励矩阵 $\mathcal{R}$, 用于表示从状态 s 到下一个状态 s' 的动作奖励值, 然后由即时奖励矩阵 $\mathcal{R}$ 计算得出指导 agent 行动的 $\mathcal{Q}$ 矩阵; agent 的状态为 $S = \langle \mathcal{V}, \mathcal{D}, \mathcal{E} \rangle$, 当 $\mathcal{V} = V^*$ 时达到终止状态; 在评估即时奖励时, 我们基于 Dijkstra 算法[23] 计算两个节点之间的最短路径长度. 基于上述思想, 抢修队调度算法的具体流程如下:

步骤 1. 初始化路网结构, 包括每个应急点的受灾程度 I_i、最大可接受距离 D_i, 每个路段的长度 l_{ij}, 受损路段的修复时间 t_{ij}, 抢修队的行驶速度 v.

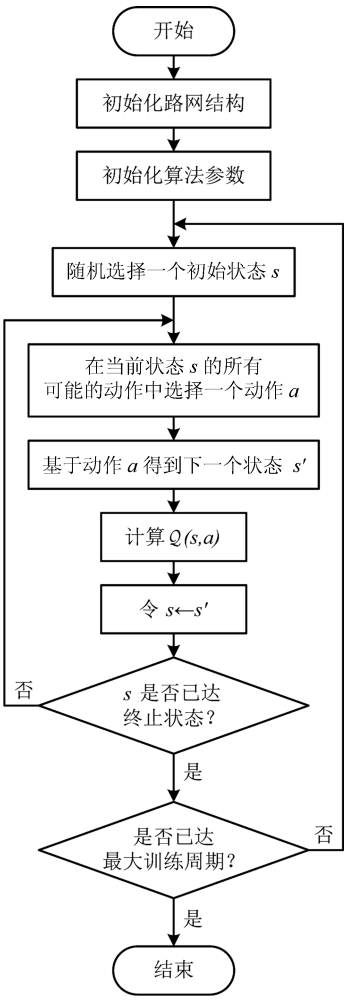


图 2　Q 学习算法流程图

Fig. 2　Flowchart of the Q-learning

步骤 2. 初始化算法参数: 最大训练周期数 (从初始状态到终止状态为一个训练周期), 折扣因子 γ, 即时奖励矩阵 $\mathcal{R}$, $\mathcal{Q}$ 矩阵.

步骤 3. 随机选择一个初始状态 s.

步骤 4. 在当前状态 s 的所有可能的动作中依据 ε-greedy 策略[29] 选择一个动作 a, 即有 $\varepsilon \in (0, 0.1]$ 的 (微小) 概率随机选取动作, 有 $1-\varepsilon$ 的 (较大) 概率从 $\mathcal{Q}$ 矩阵中选择当前最优的动作.

步骤 5. 利用选定的动作 a, 得到下一个状态 s'.

步骤 6. 对于下一个状态 s', 基于其所有可能的动作, 获得最大的 $\mathcal{Q}$ 值:

$$\begin{aligned} \mathcal{Q}(s,a) = &(1-\alpha)\cdot \mathcal{Q}(s,a) + \\ &\alpha \cdot (\mathcal{R} + \gamma \cdot \max_{a'} \mathcal{Q}(s',a')) \end{aligned} \tag{7}$$

其中, $\alpha \in (0,1]$ 表示旧的 $\mathcal{Q}$ 值被更新的程度; $\mathcal{R}$ 是 agent 在状态 s 执行动作 a 接收的立即回报, 由式 (5) 计算; $\gamma \in [0,1]$ 是折扣因子.

步骤 7. 设置下一个状态作为当前状态: $s \leftarrow s'$. 如果 s 未达到终止状态 (即 $\mathcal{V} \neq V^*$), 转步骤 4.

步骤 8. 如果算法未达到最大训练周期数, 转步骤 3, 否则根据 $\mathcal{Q}$ 矩阵输出 agent 的最优动作序列.

在上述算法中, 学习率 α 使用的是固定值, 这样智能体最近访问的状态–动作 $\mathcal{Q}(s,a)$ 能够在 $\mathcal{Q}$ 值更新中占有更大的权重, 而且, 状态 s 保留了所有已达节点的节点号和最短路径长度等历史信息.

4　实验结果与分析

正如前述, 本文是在 Duque 等[19] 提出的数学模型基础上, 对其所提模型和算法的改进和拓展, 因此, 为了验证本文所提 Q-learning 算法的有效性, 我们首先给出受损路网的实验环境和参数设置, 然后将 Q-learning 算法与 Duque 等[19] 所提的 DP 算法在不同的路网环境中进行对比分析. 此外, 与本文模型不同的是, Duque 等[19] 的模型是采用受损节点来表示受损边, 因此, 为了对比的公平性, 我们在按照本文的模型随机生成受损路网测试实例后, 严格按照 Duque 等[19] 的模型对测试实例进行转换, 以适应其所提的 DP 算法.

4.1　实验环境与参数设置

本文设计了 4 种规模的受损路网, 表 1 给出了受损路网和 Q-learning 算法的参数设置. 其中, $|V|$ 为节点数; $|E_r|/|E|$ 为在路网中受损边占全部边的比例, 我们根据这个比列来随机挑选路段设为受损路段; $D_i/\mathcal{D}_{i0}$ 为各应急点 i 到储备点 0 的最大可接受距离 D_i 相对于其最短路径长度 $\mathcal{D}_{i0}$ 的比例, 我们依据这个比例和 Dijkstra 算法算出的 $\mathcal{D}_{i0}$ 来设置每个应急点的 D_i. 此外, 对于 Q-learning 算法来

说, 算法参数取值不同可能会导致算法性能的波动, 表 1 中的 Q-learning 算法参数为我们结合已有工作[22, 27−28] 并通过大量测试所获得.

对于每一个规模 $|V| \in \{26, 31, 36, 41\}$, 我们设计了如表 2 所示的 9 组分别在 $|E_r|/|E|$、$D_i/\mathcal{D}_{i0}$ 和训练周期数上不同的参数, 对应每个 $|V|$ 下 9 个不同的测试实例, 构成 36 个测试实例. 这样设计的目的在于, 对于相同的 $|V|$, 随着 $|E_r|/|E|$ 的值的增加, 测试实例中受损边的数目也会增加, 意味着路网的受损情况变得更加复杂. 与此同时, agent 的状态空间也会增加. 因此, 需要调整 Q-learning 算法中的训练周期数以确保算法能够收敛. 此外, 相同的 $|V|$ 和相同的 $|E_r|/|E|$ 条件下, 随着 $D_i/\mathcal{D}_{i0}$ 的值的增加, 应急点的最大可接受距离会相应增加, 意味着受损路网放松了对距离的约束.

此外, 我们从算法运行时间、路网修复效果和抢修队规划方案三个方面衡量两种算法的效果. 每个测试样本均是根据表 1 的参数随机生成, 并在 Intel Core i5 CPU 3.2 GHz、内存 8 GB、操作系统 Windows 10 的个人计算机上独立运行 10 次 (其中统计分析实验运行 30 次).

4.2 算法运行时间

图 3 给出了两种方法的平均运行时间 (秒). 可以看出, 在路网规模和路段受损率都比较大的时候, Q-learning 算法明显比 DP 算法更耗时. 具体来说, 随着 $|E_r|/|E|$ 的值的增加, 路网中的受损路段越来越多, 为了 Q-learning 算法能够收敛, 训练周期数也相应增加, 导致算法的运行时间明显增加. 这是因为, Q-learning 算法在每一个次训练周期内, 会反复利用 Dijkstra 算法来估计 agent 的状态和计算 agent 的立即回报, 而 Dijkstra 算法的时间复杂度为 $O(|V|^2)$. 此外, 在 DP 算法中, 一旦一个路段被修复, 后续就不会再考虑这个路段. 因此, 受损路网的规模会不断减小, 而且 DP 算法一旦发现所有尚未修复的路段不能使新的应急点连通, 算法就会终止, 而 Q-learning 算法在每一个训练周期都是从头开始的. 不过, 即使是在 $|V| = 41$ 时, Q-learning 算法的最大平均运行时间也不到 17 秒, 从工程角度来说, 这是一个可接受的时间. 在后面, 我们可以考虑基于堆这种数据结构对 Dijkstra 算法进行优化, 提高 Q-learning 算法的运行效率.

4.3 路网修复效果

图 4 给出了 Q-learning 算法和 DP 算法在每个测试实例下目标函数 (2) 和 (3) 上的值. 注意, 如果图中没有出现某些测试实例的序号, 则表示在这些测试实例下, 算法没有找到可行解.

表 1 受损路网和 Q-learning 算法的基本参数设置

Table 1 Basic parameter settings of the damaged road network and the Q-learning algorithm

$\|V\|$	$\frac{\|E_r\|}{\|E\|}$	$\frac{D_i}{\mathcal{D}_{i0}}$	I_i	l_{ij}	t_{ij}	v	λ	训练周期数	ε	α	γ
$\{26, 31, 36, 41\}$	$\{0.1, 0.25, 0.5\}$	$\{1.05, 1.25, 1.5\}$	$[1, 10]$	$[1, 10]$	$[1, 10]$	1	0.9	$[100, 10\,000]$	0.1	0.4	0.2

表 2 36 个不同测试实例的参数设置

Table 2 Parameter settings of the 36 different test instances

测试实例		1	2	3	4	5	6	7	8	9
$\|V\| = 26$	$\frac{\|E_r\|}{\|E\|}$	0.1	0.1	0.1	0.25	0.25	0.25	0.5	0.5	0.5
	$\frac{D_i}{\mathcal{D}_{i0}}$	1.05	1.25	1.5	1.05	1.25	1.5	1.05	1.25	1.5
	训练周期数	100	100	100	500	500	500	1 000	1 000	1 000
$\|V\| = 31$	$\frac{\|E_r\|}{\|E\|}$	0.1	0.1	0.1	0.25	0.25	0.25	0.5	0.5	0.5
	$\frac{D_i}{\mathcal{D}_{i0}}$	1.05	1.25	1.5	1.05	1.25	1.5	1.05	1.25	1.5
	训练周期数	500	500	500	1 000	1 000	1 000	3 000	3 000	3 000
$\|V\| = 36$	$\frac{\|E_r\|}{\|E\|}$	0.1	0.1	0.1	0.25	0.25	0.25	0.5	0.5	0.5
	$\frac{D_i}{\mathcal{D}_{i0}}$	1.05	1.25	1.5	1.05	1.25	1.5	1.05	1.25	1.5
	训练周期数	300	300	300	900	900	900	5 000	5 000	5 000
$\|V\| = 41$	$\frac{\|E_r\|}{\|E\|}$	0.1	0.1	0.1	0.25	0.25	0.25	0.5	0.5	0.5
	$\frac{D_i}{\mathcal{D}_{i0}}$	1.05	1.25	1.5	1.05	1.25	1.5	1.05	1.25	1.5
	训练周期数	300	300	300	3 000	3 000	3 000	10 000	10 000	10 000

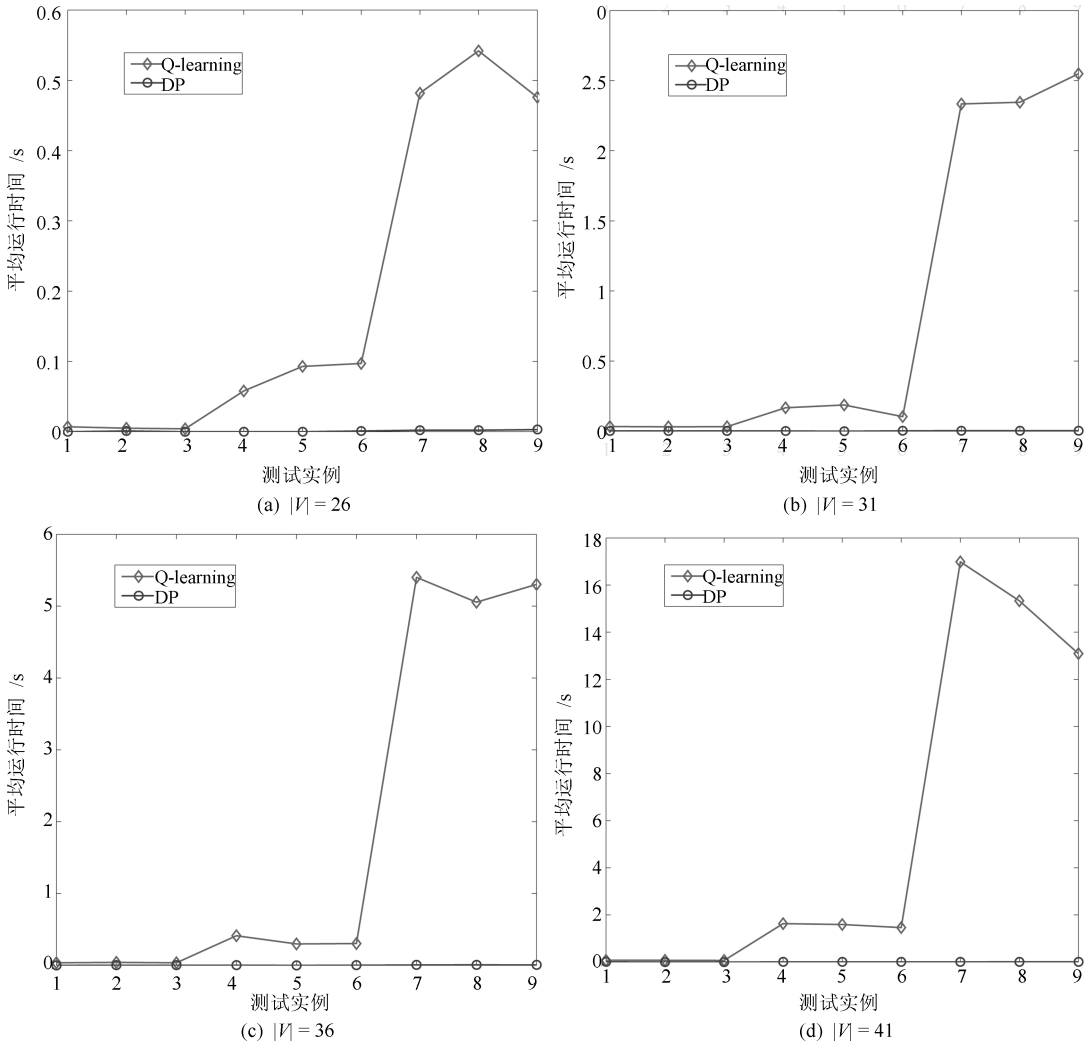


图 3　两种算法的平均运行时间 (秒)

Fig. 3　The average running time of the two algorithms (s)

表 3　36 个不同测试实例的加权目标函数值 (均值 ± 标准差)

Table 3　Weighted objective function values (mean and standard deviation) of the 36 different test instances

测试实例	$\|V\| = 26$		$\|V\| = 31$		$\|V\| = 36$		$\|V\| = 41$	
	Q-learning	DP	Q-learning	DP	Q-learning	DP	Q-learning	DP
1	2 176.40 ± 4.21	2 160.05	**5 388.19 ± 0.91**	–	**3 655.55 ± 34.83**	3 739.2	**6 096.11 ± 63.38**	6 832.49
2	2 201.18 ± 1.7	2 192.71	**5 438.54 ± 53.68**	–	**3 540.41 ± 0.76**	3 964.6	**6 005.27 ± 56.46**	–
3	2 201.18 ± 1.7	2 192.71	**5 427.4 ± 41.31**	–	**3 625.27 ± 36.28**	3 960.7	**6 276.5 ± 126.58**	–
4	**3 265.48 ± 4.71**	3 924.92	**2 800.84 ± 63.33**	2 880.5	**4 919.75 ± 33.79**	5 657.62	**4 844.56 ± 288.19**	6 011.95
5	**3 318.17 ± 43.28**	3 839.56	**2 697.35 ± 43.7**	2 851.61	**4 756.02 ± 94.22**	5 765.13	**4 586.82 ± 238.45**	6 233.67
6	**3 367.82 ± 75.62**	3 705.91	2 784.31 ± 82.92	2 578.67	**4 663.86 ± 71.74**	6 777.95	**4 309.49 ± 129.79**	6 355.16
7	**3 222.2 ± 179.3**	–	**4 405.91 ± 39.26**	–	**3 566.12 ± 62.31**	–	**6 189.04 ± 125.91**	–
8	**2 813.94 ± 15.41**	–	**3 824.53 ± 145**	–	**3 350.95 ± 76.4**	–	**5 891.23 ± 184.08**	–
9	**2 930.07 ± 69.46**	–	**3 919.55 ± 201.8**	–	**3 004.85 ± 135.9**	–	**5 356.4 ± 177.42**	–

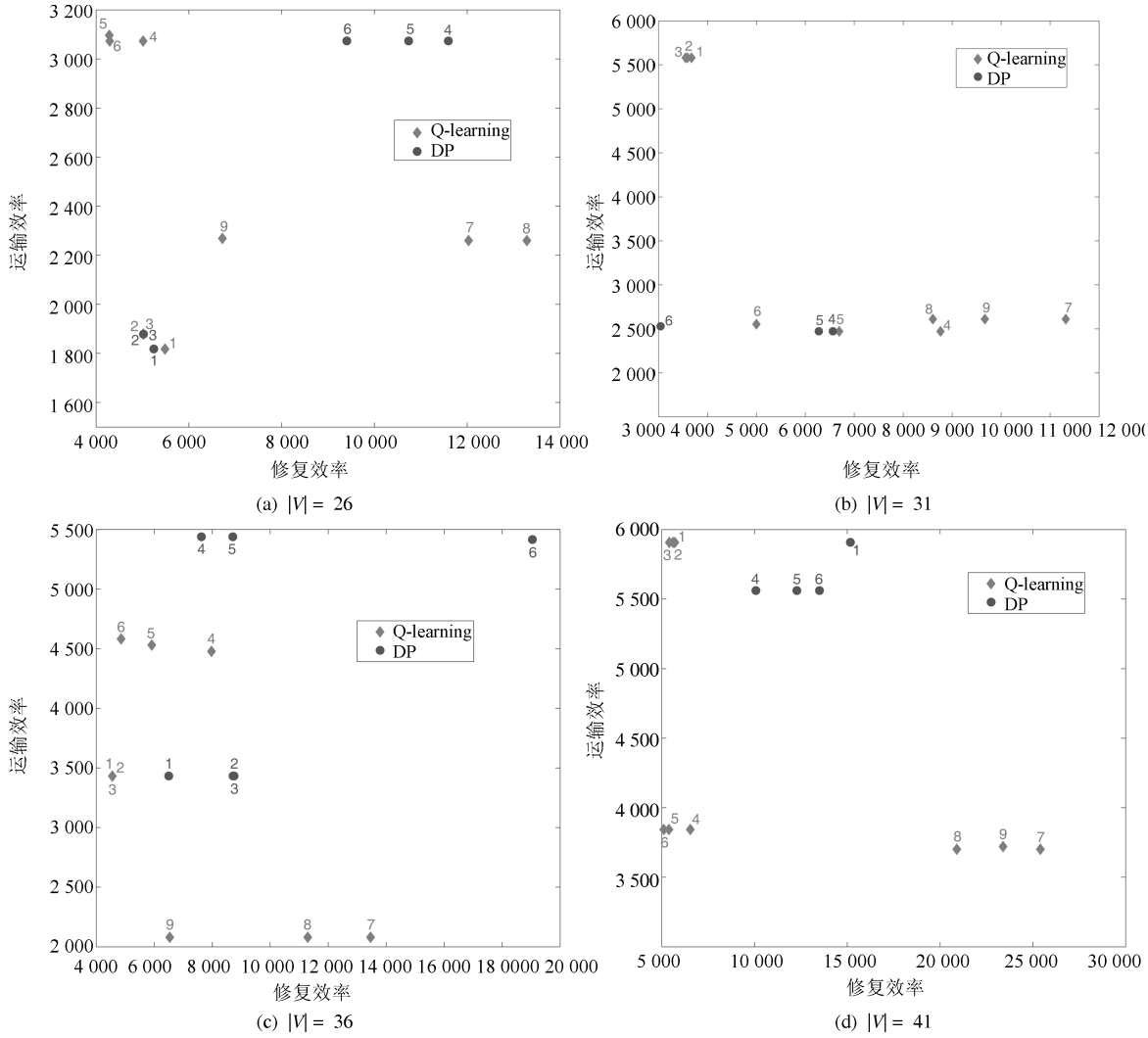


图 4 两种算法在每个测试实例下的目标函数值

Fig. 4 The objective function values of the two algorithms for each test instance

从图 4 可以看出, 当 $|V| = 26$ 时, Q-learning 算法分别在测试实例 4、6、7、8 和 9 上的解可以支配 DP 算法, 而 DP 算法只有在测试实例 1 上占优; 当 $|V| = 31$ 时, Q-learning 算法分别在测试实例 1、2、3、7、8 和 9 上的解可以支配 DP 算法, 而 DP 算法只有在测试实例 4、5 和 6 上占优; 当 $|V| = 36$ 时, Q-learning 算法除了测试实例 4 以外在其他 8 个测试实例上的解都可以支配 DP 算法, 而 DP 算法在测试实例 4 上的解不能支配 Q-learning 算法; 当 $|V| = 41$ 时, Q-learning 算法在所有 9 个测试实例上的解都可以支配 DP 算法. 此外, Q-learning 算法在所有 36 个测试实例上都能找到可行解, 而 DP 算法在 17 个测试实例上没有找到可行解. 因此, 本文的 Q-learning 算法鲁棒性更强, 对参数的敏感性较低, 所给的抢修队调度方案在运输效率和修复效率上要明显优于 DP 算法.

不过, 需要指出的是, 在 $|V| = 31$ 时, DP 算法有在测试实例 4、5 和 6 上的修复效率要优于 Q-learning 算法, 这是因为 Q-learning 算法比 DP 算法修复了更多的受损路段, 导致 f_2 的值增加. 如图 5 所示, 在灾情较严重的情况下 (即路段受损率较大、最大可接受距离较小), Q-learning 算法的受损路段修复率随着路网规模的增加要明显优于 DP 算法. 在 $|V| = 36$ 和 $|V| = 41$ 时, DP 算法的受损路段修复率几乎只有 Q-learning 算法的一半. Q-learning 通过修复更多的受损路段, 可以为每个应急点找到更短的运输路径, 有助于恢复受损路网交通系统的运输能力.

此外, 与确定性的 DP 算法不一样的是, Q-learning 具有随机不确定性特征. 因此, 我们将 Q-learning 独立运行 30 次, 从而为每个测试实例获取 30 个数据样本, 并进行统计分析, 用黑体标出较

优值. 表 3 给出了加权目标函数 (4) 的均值和标准差. 可以看出, 在总共 36 个测试实例中, Q-learning 在 32 个实例上的测试结果要优于 DP. 而且, 随着道路节点规模的增加, Q-learning 在加权目标函数 (4) 上的优势更加明显.

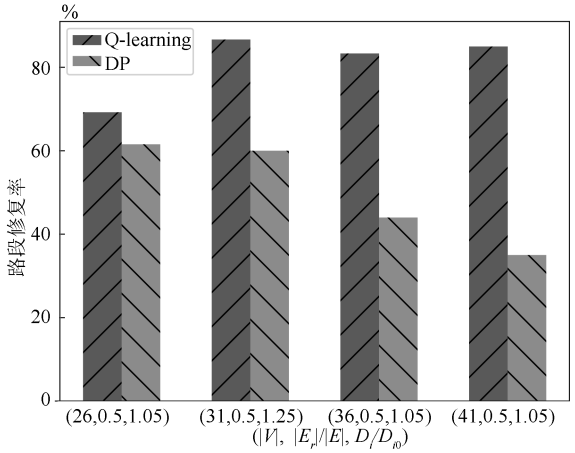


图 5　灾情严重情况下两种算法的受损路段修复率

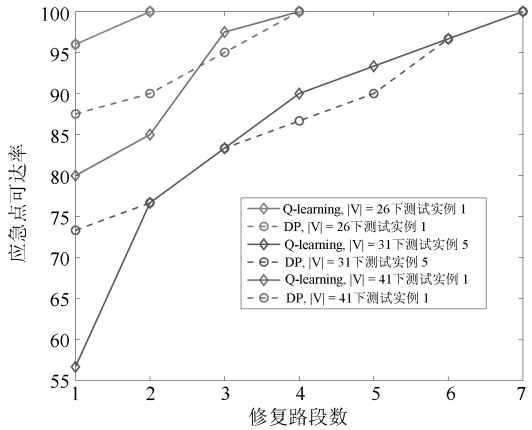
Fig. 5　Repair rate of the damaged edges of the two algorithms under a serious disaster

4.4　抢修队规划方案

在这个实验中, 我们考察在表 1 所给的路网实例下两种算法具体的修复规划的优劣. 图 6 给出了路网规模或最大可接受距离逐步增加时两种算法给出的抢修队规划方案所对应的修复路段数和应急点可达率. 其中, 图 6 (a) 专注于路段受损率较小时路网规模逐渐增加的变化, 图 6 (b) 专注于路段受损率较大时路网规模逐渐增加的变化, 图 6 (c) 专注于路网规模和路段受损率都较大时, 应急点最大可接受距离逐渐增加的变化.

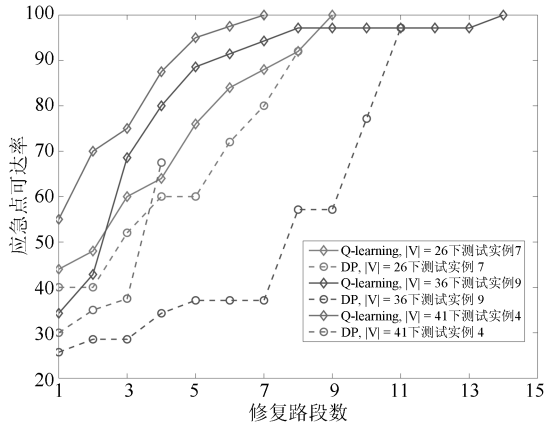
从图 6 (a) 可以看出, 在路段受损率较小时, 随着路网规模的增加, Q-learning 算法和 DP 算法均可以使所有应急点可达, 且最终修复的路段数也一样. 在路网修复初期, DP 算法可以使更多的应急点可达. 但随着修复过程的推进, Q-learning 算法很快就赶上并且超过了 DP 算法, 这是因为在 Q-learning 算法中, agent 牺牲了一部分的短期回报来获取更好的长期回报.

从图 6 (b) 可以看出, 在路段受损率较大时, 随着路网规模的增加, Q-learning 算法要明显比 DP 算法更优. 与图 6 (a) 不同的是, Q-learning 算法在路网修复整个阶段都能使更多的应急点可达, 并直至使所有应急点都可达, 且最终修复的路段数也多于 DP 算法. 而 DP 算法在整个修复过程中表现均不理想, 即使是在后期修复了 11 条路段, 应急点可达率也不到 98 %. 而且, 在 $V = 41$ 时, DP 算法在仅修复 4 条路段后就终止了. 因此, 在灾情较严重时, 与 DP 算法相比, Q-learning 算法所给的抢修队规划方案能够更快地使更多的应急点可达. 也就是说, Q-learning 算法能够更快地使路网交通系统恢复程度最大, 更有利于应急救援的实施和灾民的快



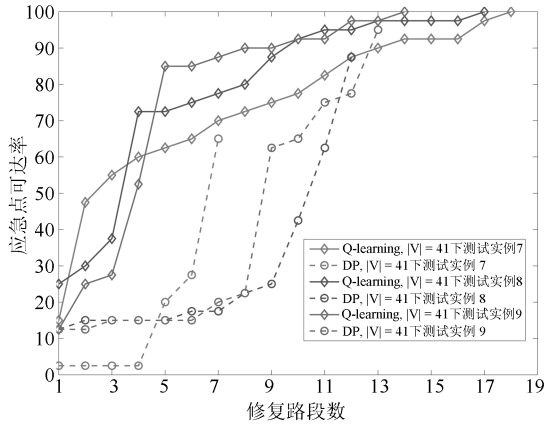
(a) 路段受损率较小，路网规模逐渐增加

(a) Increasing sizes of road network under low damage rate of road sections



(b) 路段受损率较大，路网规模逐渐增加

(b) Increasing sizes of road network under high damage rate of road sections



(c) 路网规模和路段受损率均较大、最大可授受距离逐渐增加

(c) Increasing maximum acceptable distances under high damage rate of road sections and large size of road network

图 6　两种算法的修复路段数和应急点可达率

Fig. 6　The average numbers of repaired edges and node accessibilities of the two algorithms

速安全疏散.

从图 6 (c) 可以看出, 在路网规模和路段受损率都较大时, 无论应急点最大可接受距离变大还是变小, 仍然只有 Q-learning 能够使所有应急点可达, 且修复了更多的路段. 而 DP 算法随着最大可接受距离的变小, 应急点可达率越来越小. 而且, Q-learning 在整个修复阶段都要优于 DP 算法, 在 $D_i/\mathcal{D}_{i0} = 1.05$ 时, DP 算法在仅修复 7 条路段后就终止了. 因此, 当距离约束要求发生变化时, Q-learning 更加鲁棒, 能够统筹考虑受损路网的全局, 会随着距离约束要求的变化从全局和长期收益的角度让抢修队重新适应这些约束的变化. 与之明显不同的是, 当距离约束要求变化时, DP 算法不能做出相应的调整, 尤其在约束苛刻时, DP 算法可能会找不到任何有效规划.

5 结论

本文针对灾后受损路网修复这一难点问题展开研究, 首先构建了抢修队修复和路线规划两目标优化模型, 然后基于马尔科夫决策过程设计了抢修队的决策模型, 并基于 Q-learning 算法求解抢修队的最优调度策略. 对比实验结果表明, 本文的 Q-learning 算法能够使抢修队从全局和长远角度实施受损路段的修复活动, 快速实现所有应急点可达, 而不用修复所有的受损路段. 在灾情较严重的情况下, 本文的 Q-learning 算法可以在一定程度上提高运输效率和修复效率, 为政府实施应急救援和快速安全疏散灾民提供了更多合理可靠的选择.

但是, 本文只是对灾后受损路网修复在抢修队视角下的一个初步探索, 旨在能为政府的应急响应决策提供有益的技术支撑. 本文仍有如下问题需要在以后的工作中进一步加以研究: 1) 和惯例一样[18−20], 本文是在剔除了所有的非需求节点之后建立的路网模型, 但在实际应急环境中, 剔除某些非需求节点可能会影响到整个实际路网的结构. 而且, 在应急响应初期, 地震和洪水等可能会引起余震、泥石流和山体滑坡等次生灾害, 这时原来的非需求节点可能会变成新的需求节点, 原来的可通行路段有可能变成新的受损路段. 因此, 如何在优化抢修队修复规划中考虑需求节点和受损路段的动态变化是未来一个非常有实际意义的研究课题. 2) 在有多个物资储备点的情形下, 将有多个道路抢修队从各自的储备点出发, 需要从应急全局同时规划这些抢修队的修复活动, 这可能需要研究新的数学模型、抢修队决策模型和采用新的强化学习算法. 3) 在多个物资储备点情形下, 当状态空间很大时, Q 学习算法需要较长的训练才能够收敛到较好的解, 这时直接利用列表表示 $\mathcal{Q}$ 函数已经不可能, 需要借助一些函数近似的方法, 如值梯度法、策略梯度法、Actor-Critic 法[30] 和深度强化学习[31] 等, 其次, 对于之前观测历史信息的利用, 如资格轨迹 (Eligibility traces) 问题、智能体的探索策略 (如 Soft max, UCB, Gradient Bandit)[30] 等, 这也是未来值得深入研究的方向.

References

1 Su Zhao-Pin, Zhang Ting, Zhang Guo-Fu, et al. Evaluation of emergency disposal schemes based on cloud model and fuzzy aggregation. *Pattern Recognition and Artificial Intelligence*, 2014, **27**(11): 1047−1055
(苏兆品, 张婷, 张国富, 等. 基于云模型和模糊聚合的应急方案评估. 模式识别与人工智能, 2014, **27**(11): 1047−1055)

2 Zhang Guo-Fu, Wang Yong-Qi, Su Zhao-Pin, et al. Modeling and solving multi-objective allocation-scheduling of emergency relief supplies. *Control and Decision*, 2017, **32**(1): 86−92
(张国富, 王永奇, 苏兆品, 等. 应急救援物资多目标分配与调度问题建模与求解. 控制与决策, 2017, **32**(1): 86−92)

3 Su Z, Zhang G, Liu Y, et al. Multiple emergency resource allocation for concurrent incidents in natural disasters. *International Journal of Disaster Risk Reduction*, 2016, **17**: 199−212

4 Su Zhao-Pin, Zhang Guo-Fu, Jiang Jian-Guo, et al. Multi-objective approach to emergency resource allocation using none-dominated sorting Based differential evolution. *Acta Automatica Sinica*, 2017, **43**(2): 188−207
(苏兆品, 张国富, 蒋建国, 等. 基于非支配排序差异演化的应急资源多目标分配算法. 自动化学报, 2017, **43**(2): 188−207)

5 Averbakh I. Emergency path restoration problems. *Discrete Optimization*, 2012, **9**(1): 58−64

6 Çelik M. Network restoration and recovery in humanitarian operations: Framework, literature review, and research directions. *Surveys in Operations Research and Management Science*, 2016, **21**(2): 47−61

7 Chen S Y. Optimal scheduling of logistical support for an emergency roadway repair work schedule. *Engineering Optimization*, 2012, **44**(9): 1035−1055

8 Huo Jian-Shun. A Study on Optimal Scheduling of Emergency Roadway Repair after Earchquake [Master dissertation]. Southwest Jiaotong University, 2007
(霍建顺. 震后应急期道路抢修优化排程研究 [硕士学位论文]. 西南交通大学, 2007)

9 Nurre S G, Cavdaroglu B, Mitchell J E, et al. Restoring infrastructure systems: an integrated network design and scheduling (INDS) problem. *European Journal of Operational Research*, 2012, **223**(3): 794−806

10 Yan S, Shih Y L. Optimal scheduling of emergency roadway repair and subsequent relief distribution. *Computers & Operations Research*, 2009, **36**(6): 2049−2065

11 Yan S, Shih Y L. An ant colony system-based hybrid algorithm for an emergency roadway repair time-space network flow problem. *Transportmetrica*, 2012, **8**(5): 361−386

12 Hua Bing-Wei, Wei Lin, Wang Fang, et al. Based on vulnerability to optimization the post-disaster road network restores. *Highway Engineering*, 2013, **38**(3): 18−21
(花丙威, 魏琳, 王芳, 等. 基于脆弱性的灾后路网修复优化. 公路工程, 2013, **38**(3): 18−21)

13 Qiu Hui. Study on the Restoration Sequence of Highway Network after Disaster [Master dissertation]. Chang′an University, 2016
(邱慧. 灾后公路网修复序列研究 [硕士学位论文]. 长安大学, 2016)

14 Aksu D T, Özdamar L. A mathematical model for post-disaster road restoration: enabling accessibility and evacuation. *Transportation Research Part E: Logistics and Transportation Review*, 2014, **61**(1): 56−67

15 Özdamar L, Aksu D T, Ergüneş B. Coordinating debris cleanup operations in post disaster road networks. *Socio-Economic Planning Sciences*, 2014, **48**(4): 249−262

16 Kasaei M, Salman F S. Arc routing problems to restore connectivity of a road network. *Transportation Research Part E: Logistics and Transportation Review*, 2016, **95**: 177−206

17 Akbari V, Salman F S. Multi-vehicle synchronized arc routing problem to restore post-disaster network connectivity. *European Journal of Operational Research*, 2017, **257**(2): 625−640

18 Li Ai-Qing. Scheduling Research of Emergency Roadway Repair and Relief Material Distribution after An Earchquake [Master dissertation]. Southwest Jiaotong University, 2007
(李爱庆. 震后紧急道路抢修与救灾物资配送调度研究 [硕士学位论文]. 西南交通大学, 2007)

19 Duque P A M, Dolinskaya I S, Sörensen K. Network repair crew scheduling and routing for emergency relief distribution problem. *European Journal of Operational Research*, 2016, **248**(1): 272−285

20 Kim S, Park Y, Lee K, et al. Repair crew scheduling considering variable disaster aspects. In: Proceedings of IFIP International Conference on Advances in Production Management Systems. Hamburg, Germany: Springer, 2017. 57−63

21 Delgado K V, Barros L N D, Dias D B, et al. Real-time dynamic programming for Markov decision processes with imprecise probabilities. *Artificial Intelligence*, 2016, **230**: 192−223

22 Su Z, Jiang J, Liang C, et al. Path selection in disaster response management based on Q-learning. *International Journal of Automation and Computing*, 2011, **8**(1): 100−106

23 Dijkstra E W. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1959, **1**(1): 269−271

24 Yang Chao-Lin, Shen Hou-Cai, Gao Chun-Yan. Joint control of component production and inventory allocation in an assemble-to-order system with lost sales. *Acta Automatica Sinica*, 2011, **37**(2): 234−240
(杨超林, 沈厚才, 高春燕. 按单装配系统中组件生产和库存分配控制策略研究. 自动化学报, 2011, **37**(2): 234−240)

25 Li Zhi, Tan De-Qing. Optimal control of ATO system with individual components and product demands based on Markov decision process. *Acta Automatica Sinica*, 2016, **42**(5): 782−791
(李稚, 谭德庆. 基于马尔科夫决策过程的 ATO 系统独立组件与产品双需求最优决策研究. 自动化学报, 2016, **42**(5): 782−791)

26 Mo Li-Po, Yu Yong-Guang. Finite-time rotating encirclement control of multi-agent systems. *Acta Automatica Sinica*, 2017, **43**(9): 1665−1672
(莫立坡, 于永光. 多智能体系统的有限时间旋转环绕控制. 自动化学报, 2017, **43**(9): 1665−1672)

27 Tang Hao, Pei Rong, Zhou Lei, Tan Qi. Coordinate control of multiple CSPS system based on state aggregation method. *Acta Automatica Sinica*, 2014, **40**(5): 901−908
(唐昊, 裴荣, 周雷, 谭琦. 基于状态聚类的多站点 CSPS 系统的协同控制方法. 自动化学报, 2014, **40**(5): 901−908)

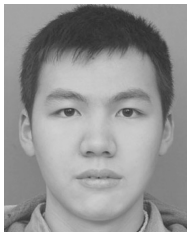
28 Xu Mao-Xin, Zhang Xiao-Shun, Yu Tao. Transfer bees optimizer and its application on reactive power optimization. *Acta Automatica Sinica*, 2017, **43**(1): 83−93
(徐茂鑫, 张孝顺, 余涛. 基迁移蜂群优化算法及其在无功优化中的应用. 自动化学报, 2017, **43**(1): 83−93)

29 Gomes E R, Kowalczyk R. Dynamic analysis of multiagent Q-learning with ε-greedy exploration. In: Proceedings of 26th International Conference on Machine Learning. Montreal, Canada: Omnipress, 2009. 369−376

30 Panait L, Luke S. Cooperative multi-agent learning: the state of the art. *Autonomous Agents and Multi-Agent Systems*, 2005, **11**(3): 387−434

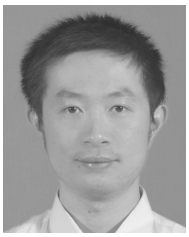
31 Gu S, Lillicrap T, Sutskever I, et al. Continuous deep Q-learning with model-based acceleration. In: Proceedings of 33rd International Conference on Machine Learning. New York, NY, USA: ACM Press, 2016. 2829−2838

苏兆品 合肥工业大学计算机与信息学院副教授. IEEE 会员. 2008 年获得合肥工业大学计算机科学与技术专业博士学位. 主要研究方向为演化计算, 灾害应急决策, 多媒体安全.
E-mail: szp@hfut.edu.cn
(**SU Zhao-Pin** Associate professor at the School of Computer Science and Information Engineering, Hefei University of Technology. She is a member of IEEE. She received her Ph. D. degree in computer science and technology from Hefei University of Technology in 2008. Her research interest covers evolutionary computation, disaster emergency decision-making, and multimedia security.)



李沫晗 合肥工业大学计算机与信息学院硕士研究生. 2014 年获得合肥工业大学光信息科学与技术专业学士学位. 主要研究方向为灾害应急决策和强化学习. E-mail: limohan@mail.hfut.edu.cn
(**LI Mo-Han** Master student at the School of Computer Science and Information Engineering, Hefei University of Technology. He received his bachelor degree in optical information science and technology from Hefei University of Technology in 2014. His research interest covers disaster emergency decision-making and reinforcement learning.)

张国富 合肥工业大学计算机与信息学院教授. 中国自动化学会、IEEE 会员. 2008 年获得合肥工业大学计算机科学与技术专业博士学位. 主要研究方向为计算智能, 多 agent 系统, 基于搜索的软件工程. 本文通信作者.
E-mail: zgf@hfut.edu.cn
(**ZHANG Guo-Fu** Professor at the School of Computer Science and Information Engineering, Hefei University of Technology. He is a member of CAA and IEEE. He received his Ph. D. degree in computer science and technology from Hefei University of Technology in 2008. His research interest covers computational intelligence, multi-agent systems, and search-based software engineering. Corresponding author of this paper.)



刘 扬 合肥工业大学计算机与信息学院博士研究生. 2005 年获得合肥工业大学通信工程专业学士学位, 2007 年获得合肥工业大学信号与信息处理专业硕士学位. 主要研究方向为灾害应急决策和演化计算. E-mail: lyy673@163.com
(**LIU Yang** Ph. D. candidate at the School of Computer Science and Information Engineering, Hefei University of Technology. He received his bachelor degree in communication engineering and master degree in signal and information processing from Hefei University of Technology in 2005 and 2007, respectively. His research interest covers disaster emergency decision-making and evolutionary computation.)