doi:10.11887/j.cn.202004001　　http://journal.nudt.edu.cn

强化学习框架下移动自组织网络分步路由算法*

蒯振然,王少尉
(南京大学 电子科学与工程学院, 江苏 南京　210023)

摘　要:移动自组织网络是一种无基础设施、由移动通信节点组成的无线网络,具有高动态特性。传统的路由协议并不能适应节点移动性带来的频繁拓扑变化,简单的洪泛路由也会因开销过大降低网络的性能。针对如何在移动自组织网络中自适应地进行路由选择,提出强化学习框架下的分步路由选择算法。该算法以最小链路总往返时延为目标,基于强化学习进行路由搜寻,在筛选出符合目标需求节点集合的基础上,结合置信度选择路由。在链路变得不可靠时,数据包被广播给筛选出的邻居节点集以提升路由可靠性并降低开销。对提出的算法在分组到达率和路由开销等主要性能指标进行数值仿真分析。仿真结果表明,提出的分步路由算法相比于基于强化学习的智能鲁棒路由,在降低开销的同时,保持着相当的吞吐率。

关键词:移动自组织网络;强化学习;路由算法
中图分类号:TN92　　**文献标志码**:A　　**开放科学(资源服务)标识码(OSID)**:
文章编号:1001-2486(2020)04-001-06

听语音　与作者互动　聊科研

Stepwise routing algorithm in mobile ad hoc network under reinforcement learning framework

KUAI Zhenran, *WANG Shaowei*
(School of Electronic Science and Engineering, Nanjing University, Nanjing 210023, China)

Abstract: Mobile ad hoc network is a communication network formed by mobile nodes with non-infrastructure, which has highly dynamic characteristics. Conventional routing protocols cannot adapt to the frequent topology changes brought by node mobility, and the flooding routing also causes the network performance degradation due to the excessive routing overhead. A stepwise routing algorithm based on reinforcement learning was proposed for adaptive routing in mobile ad hoc networks. This algorithm aims at total round trip time minimization and uses the reinforcement learning algorithm to select the next hop. After selecting the set of nodes that meet the requirements of the target, it combines the confidence parameters to select the route. When the link becomes unreliable, packets are broadcasted to filtered neighbor nodes to improve the reliability and reduce the routing overhead. The main property indication of the proposed algorithm, such as throughput and routing overhead, were analyzed theoretically. The simulation results show that, compared with the reinforcement learning based smart robust routing, the proposed routing algorithm reduces the overhead and maintains a competitive throughput.

Keywords: mobile ad hoc network; reinforcement learning; routing algorithm

移动自组织网络(Mobile Ad hoc NETwork, MANET)是一种无基础设施、由移动通信节点组成的无线网络,具有组网灵活、配置方便和抗毁性强等特点[1]。由于节点的通信范围有限,源节点与目的节点之间的通信经过中间节点的多跳路由完成。然而,在一些动态场景下,节点的移动性使得拓扑结构频繁变化,传统的路由协议需要不断地计算端到端的路由来保证数据的传输,缺乏对动态网络的自适应能力,而简单的洪泛机制会产生大量开销[2-3]。因此,研究针对 MANET 动态特性的具备自适应性和可靠性的路由方式,有重要的理论和应用价值。

最早将强化学习应用于 MANET 路由的工作见诸文献[4],文中提出的 Q-Routing 算法使用了强化学习的经典 Q-Learning 框架,将衡量路径优劣的权值置于每个节点维护的 Q 表中,根据 Q 表选择下一跳节点,传回的确认字符(ACKnowledge character, ACK)用来更新 Q 表以选择更好的路由。文献[5]则根据网络拓扑结构中节点的度来调整强化学习的学习速率,更高的节点度对应更高的学习速率,从而使用更少的时间来探测网络的真实状态。

* **收稿日期**:2019-12-25
基金项目:国家自然科学基金资助项目(61671233, 61801208, 61931023)
作者简介:蒯振然(1996—),男,安徽合肥人,博士研究生,E-mail:DZ1923021@smail.nju.edu.cn;
王少尉(通信作者),男,教授,博士,博士生导师,E-mail:wangsw@nju.edu.cn

在强化学习算法收敛到最优路由的过程中会伴随着吞吐率的损失,在高度动态的场景下,这种损失会使信息的传递效率降低,所以上述方法并不能直接应用于对信息完整性要求较高的情况。文献[6]提出从节点的广播消息获得邻居节点的 Q 值,通过这种方式减少探索网络状态所需的时间,降低算法在学习过程中的性能损失。文献[7]提出通过随机轮询邻节点的自适应 Q-routing,用于防止数据包回传,该方法提高了路由在高负载条件下的稳定性。文献[8]提出了基于强化学习的智能鲁棒路由(Smart Robust Routing, SRR)算法,对每个节点,使用置信度来衡量邻居节点的可靠性。当网络状态相对稳定时,使用结合置信度的 Q-routing 方法进行路由,而节点移动导致网络状态不稳定时,则以一定概率使用广播的方式确保将信息传到目的节点,这种方式加快了在网络状态不稳定时路由的收敛速率,并保证了系统的吞吐率。然而,广播机制增加了路由开销。并且在 Q 值和置信度一直动态变化的情况下,按照统一度量选择下一跳节点,会有陷入路由环路的情况,使重复包检测机制进行丢包处理。

本文研究了 MANET 中的路由问题,对 SRR 算法的局限性做出改进,提出了一种基于强化学习的分步路由算法,通过结合强化学习路由 Q-Routing 和利用 Q 值筛选符合路由目标节点的方式,使节点更倾向于选择提升 MANET 网络性能的路由,保障数据包到达率。在筛选出的节点基础上,结合置信度实现在网络条件较差时只向部分节点广播,在提升路由可靠性的同时,降低了网络的路由开销。

1 强化学习路由框架

在强化学习任务中,状态空间为 X,智能体处在某一个状态 $x\in X$,通过从动作空间 A 选择动作 a,与环境进行交互,得到奖赏 r,使智能体学习并得到提升。在使用确定性策略的情况下,目标是选择最优的策略 $\pi: X\rightarrow A$ 来最大化 γ 折扣累积奖赏 $R(x)=E[\sum_{t=0}^{+\infty}\gamma^t r_{t+1}\mid x_0=x,\pi]$。

Q-Learning 算法是强化学习算法中基于值函数的算法。其中智能体维护并更新一个 $|X|\times|A|$ 的 Q 表,表中每一项为 $Q(x_i,a_i)=E[\sum_{t=0}^{+\infty}\gamma^t r_{t+1}\mid x_0=x_i,a_0=a_i,\pi]$。对应于在状态 x_i 下,采取动作 a_i 会产生的期望累积奖赏。在时刻 t,状态为 x_t 时,最优策略 π^* 就可以选择表中 Q 值最大的动作 $a_t=\arg\max_a Q(x_t,a)$。在转移到下一个状态 x_{t+1},并观察到奖赏 r_{t+1} 之后,$Q(x_t,a_t)$ 进行更新的过程表示为:

$$Q_{t+1}(x_t,a_t)=(1-\alpha_t)Q_t(x_t,a_t)+\alpha_t[r_{t+1}+\gamma\max_{a'}Q_t(x_{t+1},a')] \quad (1)$$

式中,$\alpha_t\in(0,1]$ 是学习速率。式(1)代表以 α_t 的速率用估计的累积奖赏对当前的累积奖赏进行更新[9]。在与环境交互的过程中,随着 Q 表的更新,Q 值会逐渐接近于真实的累积奖赏,智能体在各种状态下会趋向于选择最优的动作。

1.1 *Q*-Routing

本节将介绍文献[4]如何将 MANET 路由问题归约到 Q-learning 算法的框架下,并具体说明在强化学习框架下所采用的奖赏与更新方式。对每个节点,可以将其看作智能体,每个节点只有在数据包到达节点时,才需要转发,所以只存在一种需要做动作的状态。节点需转发数据包至邻居节点,最终目的地是目的节点。

假设发送端可以检测发出数据包和接收到 ACK 之间的往返时延(Round Trip Time, RTT),可以将其作为节点选择下一跳节点的奖赏。以 Q-Learning 算法为框架,让每个节点 m 维护一个元素数目为其邻居节点数量的 Q 表。表中的每一项为 $Q_m(d,n)$,代表选择邻居节点 n 作为下一跳节点的情况下,在到达目的节点 d 前的总 RTT。该数值可以一定程度上体现节点的端到端时延性能,可以将最小化总 RTT 作为 Q-Routing 的目标。

由式(1)可知,在转发数据包之后,节点 Q 表的更新还需要下一个状态的最大估计奖赏。在 MANET 网络的多跳路由中,由于节点转发的次数是有限的,所以考虑直接将数据包在发送之后传输到目的节点的总 RTT 来作为节点的累积奖赏,即 $Q_m(d,n)=E[\sum_{i=n}^{d}T^i_{\mathrm{RTT},t}\mid m_t=m,n_t=n]$。

那么在下一跳节点 n 接收包之后,需要通过 ACK 的方式传回该节点最优的估计奖赏,以利于上一跳节点更新 Q 值,即 $Q_n^*=\min_l Q_n(d,l)$。该值表示从节点 n 传输到目的 d 的最小链路总 RTT。节点 m 的 Q 表对应条目更新过程为:

$$Q_{m,t+1}(d,n)=(1-\alpha_t)Q_{m,t}(d,n)+\alpha_t(T_{\mathrm{RTT},t}+Q^*_{n,t}) \quad (2)$$

式中,等号右边第一项代表 Q 值的原始值,第二项可以代表 Q 值的更新部分,$T_{\mathrm{RTT},t}$ 是第 t 次传输返回的 RTT,$\alpha_t\in(0,1]$ 是第 t 次更新的学习速率,代表了 Q 值更新信息所占的权重。随着估计

值接近真实值,α_t 通常随着更新次数减小,为了使 Q 值收敛,将把 α_t 的取值与置信度的取值相联系。当更新次数增加到 j 次时,经过迭代,对于节点 m,对应邻节点为 n 的 Q 值为:

$$Q_{m,j}(d,n) = \prod_{t=1}^{j}(1-\alpha_t)Q_{m,0}(d,n) + \sum_{p=0}^{j-1}\alpha_p\prod_{t=p+1}^{j-1}(1-\alpha_t)(T_{\mathrm{RTT},p}+Q^*_{n,p}) \tag{3}$$

式中,$Q_{m,0}(d,n)$ 是节点 m 对应于邻节点 n 的初始 Q 值。从等号右边第二项可以看出,Q 值更新信息距离当前时刻越近,赋予的权重越高,对 Q 值的估计也因此越容易跟踪网络的变化。

1.2　置信度

网络中每个节点同样维护一个置信度表,每一项为 $C_m(d,n)$,代表节点 m 选择邻居节点 n 作为下一跳节点的情况下,能到达目的节点 d 的可信度,其中 $C_m(d,n)\in[0,1]$。

在路由算法中,可以结合 Q 表和置信度表做出决策,选出可信度高,并且总 RTT 小的下一跳节点。在节点 m 确定下一跳节点 n,转发数据包之后,置信度表也需要进行更新,与 Q 值更新一样,下一跳节点 n 通过 ACK 的方式传回需要更新的信息 $C^*_n=C_n(d,k^*)$。其中,k^* 表示下一跳节点 n 选定的节点,即节点 n 确定了所选的下一跳节点 k^* 后,再将 C^*_n 通过 ACK 反馈给上一跳节点 m。若节点 m 并未成功转发数据包至 n,则需要对置信度表中的对应项 $C_m(d,n)$ 更新以降低其可信度,这是为了保证只有确认接收数据后才会增加相应的置信度。具体可以将置信度的更新分为两步。

当数据包被转发时,更新过程[5]为:

$$C_{m,t+1}(d,n)=(1-\eta)C_{m,t}(d,n) \tag{4}$$

节点 m 接收节点 n 的 ACK 后,更新过程为:

$$C_{m,t+1}(d,n)=C_{m,t}(d,n)+\eta C^*_{n,t} \tag{5}$$

其中,η 是置信度更新的学习速率,代表置信度更新信息所占权重。然而,ACK 包存在延迟的情况,即在节点 m 发射端 TX 接收到接收端 RX 相应的 ACK 包之前,又转发了 k 个数据包,如图 1 所示。因为不同时刻的 ACK 影响是不同的,例如 $k=4$ 时,前两个数据包转发失败,后两个数据包转发成功意味着通信链路条件变好。因此仅使用式(4)、式(5)更新并不能反映类似的变化。于是将数据包转发成功时更新过程改为:

$$C_{m,t+1}(d,n)=C_{m,t}(d,n)+\eta(1-\eta)^kC^*_{n,t} \tag{6}$$

式中,$(1-\eta)^k$ 是对接收到的置信度更新信息根据延迟的结果进行缩放处理。文献[8]证明了式(5)与式(6)在更新次数足够多的情况下是趋向于相等的,因此成功转发数据包时,采用式(6)对置信度进行更新。

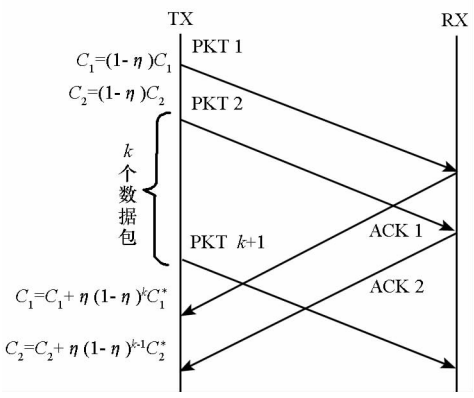


图 1　置信度更新过程

Fig. 1　Process of updating confidence factors

置信度值除了用来选择更加可靠的节点,还用来在接收到 ACK 而未更新置信度之前,调整 Q-Routing 学习速率 α_t[10]。

$$\alpha_t=\max\left(C^*_{n,t},1-\frac{C_{m,t}(d,n)}{1-\eta}\right) \tag{7}$$

目的是在邻居节点置信度更高,或当前节点转发数据包前的置信度更低时,对邻居节点返回的 Q 值在更新时赋予更高的权重。

2　分步路由选择算法

为了保证路由可靠性的同时,降低路由开销,提出了分步路由选择算法,使用 Q-Routing 和置信度结合来进行路由,目标是最小化总 RTT。源节点 s 发送和中间节点转发数据包时,先结合 Q 表和置信度表选出下一跳节点,根据该节点的置信度来选择是广播还是直接转发数据包至该节点。

在选择下一跳节点时,结合 Q-Routing 算法和置信度来进行选择,本文考虑的选择度量为 $S_{CQ}=Q_m(d,n)[1-C_m(d,n)]$,目标是选择 Q 值较小、可信度较高的下一跳节点。然而,节点的移动性导致节点的 Q 值和置信度会变化,仅基于该度量进行下一跳节点的选择,会增大选择不符合路由目标节点的概率,甚至陷入路由环路,从而增加路由开销。为了提升节点选择的容错率,先筛选出符合路由目标的节点。设节点 m 的邻居节点数量为 N,设节点 Q 表中的 Q 值按数值大小排序,即 $Q_m(d,a_1)<Q_m(d,a_2)<\cdots<Q_m(d,a_N)$。

第一步选择 Q 表中比例为 λ、Q 值最小的邻居节点子集 $\mathcal{N}_c$,大小为$\lceil\lambda N\rceil$,该集合表示为:

$$\mathcal{N}_c = \{a_1, a_2, \cdots, a_{\lceil \lambda N \rceil}\} \tag{8}$$

第二步从 $\mathcal{N}_c$ 中选择度量 S_{CQ} 最小的节点。

$$n^* = \arg\min_{n \in \mathcal{N}_c} Q_m(d,n)[1 - C_m(d,n)] \tag{9}$$

先基于 Q 值进行选择也是为了减小在置信度未收敛时，对路由开销带来的负面影响。因为在确定了下一跳节点 n^* 之后，SRR 算法根据置信度确定是否需要进行广播来保证路由可靠性。而 SRR 算法以广播的方式进行传输在增加了路由开销的条件下可以保证数据包到达率的性能，所以为了保留可靠性，并降低路由开销，提出的分步路由算法将数据包广播给在节点选择阶段筛选出的最符合路由目标的邻居节点子集 $\mathcal{N}_c$，而是否进行广播的概率则根据置信度计算，可以表示为：

$$p_U = \varepsilon + (1 - \varepsilon)[1 - C_m(d,n^*)] \tag{10}$$

由式(10)看出：置信度接近 0 时，节点进行广播的概率更大；置信度接近 1 时，广播的概率应当减小并趋于 ε。在节点发送数据包或拓扑改变的初期，置信度值不够高，就会有更高的概率选择广播的方式传输，Q 值和置信度值也会更快地更新，而且在转发初期或网络拓扑改变初期，Q 表未收敛时，广播的方式也使路由的可靠性得到了保障。

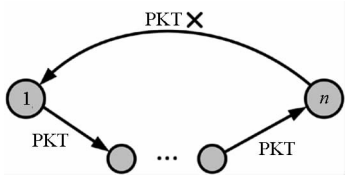
所有收到广播数据包的节点，都需要传回 ACK，包含其自身的节点序列号、下一跳节点的 Q 值和置信度等信息。

$$\begin{cases} Q_n^* = Q_m(d,n^*) \\ C_n^* = C_m(d,n^*) \end{cases} \tag{11}$$

目的是在拓扑变化时能及时更新邻居节点信息。尽管有部分邻居节点并未在筛选的集合 $\mathcal{N}_c$ 中，但是如果判断为广播数据包，仍然需要传回 ACK，使广播节点掌握邻居节点的信息。

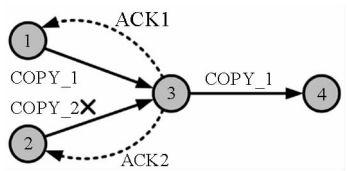
当根据广播数据包的 ACK 更新信息时，并不像 SRR 算法一样保存 Q 表和置信度表中所有节点的信息，而是选择删除表中不再是邻居节点的信息。若仍为邻居节点，则按照规则更新 Q 值和置信度，新的邻居节点将对应 Q 值设为表中其余节点 Q 值的均值，置信度则设为 1 以缩短收敛时间。若不需要进行广播，节点就根据选择的下一跳节点转发，转发后根据式(4)更新置信度。目的是在拓扑变化影响到路由时，更新节点本地信息。在转发成功，下一跳节点接收到数据包后，发送的 ACK 数据包包括了所选下一跳节点 n^* 的 Q 值和置信度以及相关节点信息。接收 ACK 的节点，则由 ACK 包含的信息根据式(7)、式(2)、式(6)更新 α_t、Q 值和置信度。

所有数据包传输的过程都采用了重复包检测机制，如图 2 所示。当出现路由环路时，接收包的节点进行丢包处理，如果只是广播的包的复制，进行丢包处理，但会传回 ACK，以提供确认该路由可用的信息来更新置信度。分步路由选择算法的步骤如算法 1 所示。



(a) 形成路由环路

(a) Traverse a routing loop



(b) 重复包到达

(b) Arrival of duplicate packets

图 2　重复数据包检测

Fig. 2　Duplicate packet detection

算法 1　分步路由算法

Alg. 1　Stepwise routing algorithm

1. 参数：$\varepsilon \in [0,1]$，$\lambda, \eta \in (0,1]$
2. 节点 m 接收数据包：
3. **if** 接收 ACK 包 **then**
4. 读取 C_n^*，Q_n^*，d，n 节点序列号等信息
5. k = ACK 对应包发送后转发包的数量
6. 根据式(7)、式(2)、式(6)更新 α_t，Q 值和置信度
7. **if** 为广播数据包 ACK **then**
8. 更新 Q 表和置信度表中的节点.
9. **else**
10. **if** 数据包形成路由环路 **then** 丢包
11. 根据式(8)选择邻居节点子集 $\mathcal{N}_c$
12. 根据式(9)计算下一跳节点 n^*
13. 根据式(11)计算 C_n^*，Q_n^*，发送 ACK
14. **if** 接收包为广播数据包 **then**
15. 标记 ACK 为广播数据包 ACK
16. **if** 包的复制已转发过 **then** 丢包
17. **if** 目的节点为 m **then** 发至传输层
18. 根据式(10)计算 p_U 决定是否广播至 $\mathcal{N}_c$
19. **if** 决定广播 **then**
20. 广播数据包至邻居节点子集 $\mathcal{N}_c$
21. 根据式(4)更新 $C_m(d,n^*)$，$n^* \in \mathcal{N}_c$
22. **else**
23. 根据式(4)更新 $C_m(d,n^*)$
24. 转发数据包至节点 n^*
25. **end if**

3　仿真结果与分析

本节给出了提出的分步路由选择算法的仿真结果。将仿真的场景设置在 1000 m × 600 m 的矩形区域,源节点和目的节点设为静止,分别位于区域 600 m 宽边的中间位置。中间节点数为 30,在矩形区域内均匀分布,节点的移动速度在[0 m/s,30 m/s]内均匀分布,方向在[0,2π]范围内均匀分布,以当前方向和速率持续的时间也在仿真时间内均匀分布,如果持续的时间结束,或者超出仿真区域范围,则重新分配速度、方向和持续时间,节点通信的范围为 300 m。仿真模拟了从源节点发送数据包到目的节点的过程。源节点以每秒 10 个1500 Byte数据包的速率发送数据,即速率为 120 kbit/s,置信度更新参数 $\eta = 0.3$。随着 Q 值和置信度更新,置信度接近于 1,此时根据 Q 值和置信度已经可以确保路由的可靠性,参数 ε 过高会增加广播概率,进而增加路由开销,因此选取 $\varepsilon = 0.05$。

图 3 比较了在不同 λ 情况下,300 s 内平均路由开销与数据包到达速率的比值。该比值用来体现不同 λ 对路由开销和数据包到达率的整体影响。路由开销则使用网络中所有链路的数据速率总和来衡量。可以看出,当 λ 值较高,接近于 1 时,广播时目标为所有邻节点,这大大增加了路由开销;当 λ 值较低时,路由初期无法尽快更新 Q 值与置信度信息,降低路由可靠性。通过比较,确定 $\lambda = 0.4$,此时,为了保障数据包到达速率而付出的路由开销要远小于对所有节点进行广播的方式。

将提出的分布路由算法(StepWise Routing, SWR)与 SSR 算法和开放最短路径优先(Open Shortest Path First, OSPF)路由[11]相比较,来分析不同算法对路由开销与数据包到达速率的影响。仿真中,以目的节点的数据速率来衡量数据包到达目的节点的成功率,并且以网络中所有链路的数据速率来衡量系统的路由开销。

图 4 比较了在 300 s 内不同路由方法在目的节点的吞吐率;图 5 则比较了对应的路由开销。从图 5 中可以看出,OSPF 路由拥有的路由开销更低,但是对目的节点来说,数据包到达率不够稳定,因为节点的移动会改变拓扑结构,使得以 OSPF 的方式路由的成功率降低。基于强化学习的 SRR 算法相对于 OSPF 算法提升了路由的稳定性,数据成功传输率平均增加了 70%,平均丢包率仅有 1.6%,相对地,也增加了 1 倍以上的路由开销。

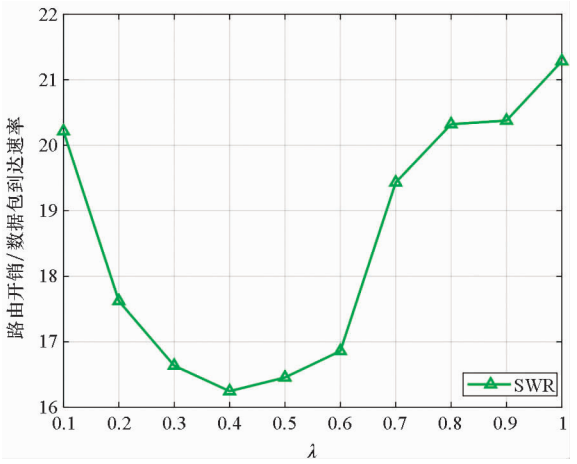


图 3　路由开销与数据包到达速率比值

Fig. 3　Ratio of routing overhead to packet delivery rate

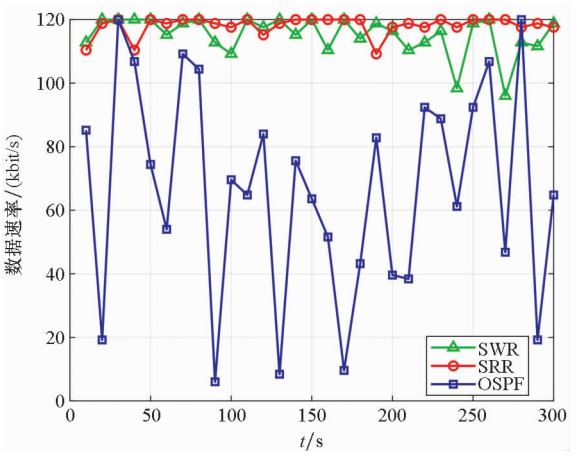


图 4　数据包到达速率仿真结果

Fig. 4　Results of packet delivery rate

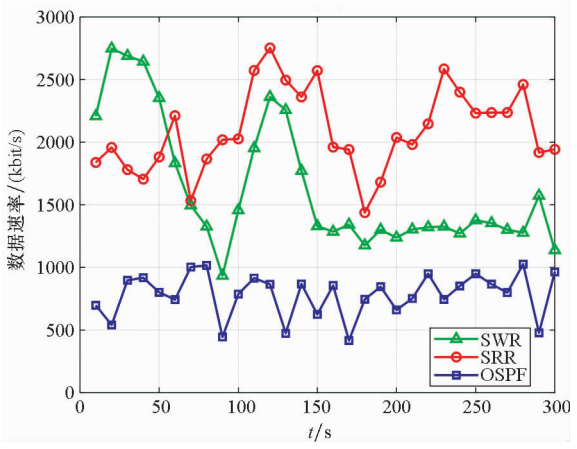


图 5　路由开销仿真结果

Fig. 5　Results of routing overhead

所提出的 SWR 算法在数据成功传输率上接近 SRR 算法,平均丢包率为 4%,而平均路由开销在 0 ~ 40 s 时相对于 SRR 算法稍高,原因是路由开始时置信度尚未收敛,SWR 广播的节点数量较

少,广播次数多,开销较大。40 s 之后相对于 SRR 降低 30%,平均开销降低了 23%。因为在路由时,SWR 算法基于 Q 值筛选出节点集合的方式再选择,所以避免了在直接结合置信度的选择中选择较差的节点;同时,即便选择广播,也利用了路由时筛选出的节点信息,从而减少无益于到达目的节点的路由选择。

4　结论

本文研究了 MANET 中的路由问题,基于强化学习提出了分步路由算法。通过结合强化学习路由 Q-Routing 和利用 Q 值筛选符合路由目标节点的方式,使节点更倾向于选择提升 MANET 网络性能的路由,保障数据包到达率。在筛选出的节点基础上,结合置信度实现在网络条件较差时只向部分节点广播,在提升路由可靠性的同时,降低了网络的路由开销。仿真结果表明:和传统 OSPF 路由相比,以少量的路由开销为代价,数据传输成功率提升了 70%;和基于强化学习的 SRR 算法相比,数据传输成功率相差仅 2.4% 的情况下,路由开销降低了 23%。

参考文献(References)

[1] Sarkar S K, Basavaraju T G, Puttamadappa C. Ad hoc mobile wireless networks: principles, protocols, and applications[M]. US: CRC Press, 2008.

[2] Sun Y, Wang T Y, Wang S W. Location optimization and user association for unmanned aerial vehicles assisted mobile networks[J]. IEEE Transactions on Vehicular Technology, 2019, 68(10): 10056 – 10065.

[3] 杜青松, 朱江, 张尔扬. 战术 MANET 中基于多态转移策略的蚁群优化 QoS 路由算法[J]. 国防科技大学学报, 2012, 34(1): 107 – 114.
DU Qingsong, ZHU Jiang, ZHANG Eryang. A novel ant-colony optimized QoS routing algorithm based on multiple transferring strategies for tactical MANETs[J]. Journal of National University of Defense Technology, 2012, 34(1): 107 – 114. (in Chinese)

[4] Boyan J A, Littman M L. Packet routing in dynamically changing networks: a reinforcement learning approach[J]. Advances in Neural Information Processing Systems, 1993: 671 – 678.

[5] Desai R, Patil B P. Enhanced confidence based Q routing for an ad hoc network[J]. American Journal of Educational Science, 2015, 1(3): 60 – 68.

[6] Haraty R A, Traboulsi B. MANET with the Q-Routing protocol[C]//Proceedings of ICN the Eleventh International Conference on Networks, 2012: 187 – 192.

[7] Kavalerov M, Shilova Y, Likhacheva Y. Adaptive Q-routing with random echo and route memory[C]//Proceedings of the 20th Conference of Open Innovations Association, 2017: 138 – 145.

[8] Johnston M, Danilov C, Larson K. A reinforcement learning approach to adaptive redundancy for routing in tactical networks[C]//Proceedings of MILCOM IEEE Military Communications Conference (MILCOM), 2018: 267 – 272.

[9] Wei Q L, Lewis F L, Sun Q Y, et al. Discrete-time deterministic Q-learning: a novel convergence analysis[J]. IEEE Transactions on Systems, Man, and Cybernetics, 2017, 47(5): 1224 – 1237.

[10] Evendar E, Mansour Y. Learning rates for Q-learning[J]. Journal of Machine Learning Research, 2004, 5: 1 – 25.

[11] Spagnolo P A, Henderson T R. Comparison of proposed OSPF manet extensions[C]//Proceedings of Military Communications Conference, 2006: 1925 – 1936.