基于 Q-learning 的分布式自适应拓扑稳定性算法

黄庆东*，石斌宇，郭民鹏，袁润芝，陈 晨

(西安邮电大学通信与信息工程学院信息与通信技术国家级实验教学中心 西安 710121)

【摘要】针对移动节点对网络拓扑稳定性的影响，提出了一种预测相邻节点稳定联接的自适应分布式强化学习算法。各节点采用强化学习和学习区间自适应划分相结合的方法，利用相邻节点间的接收信号强度信息对相邻节点间的联接状态进行判定，最终预测出能够保持稳定联接的邻居节点集。通过多种条件下随机游走模型仿真，结果表明预测准确度均保持在95%左右，验证了该算法的有效性和稳定性。

关 键 词 分布式； 移动自组织网络； 强化学习； 拓扑稳定性

中图分类号 TN911.23 **文献标志码** A doi:10.12178/1001-0548.2019076

Q-Learning Based Distributed Adaptive Algorithm for Topological Stability

HUANG Qing-dong*, SHI Bin-yu, GUO Min-peng, YUAN Run-zhi, and CHEN chen

(Informations and Communications Technology of National Experimental Teaching Center, School of Communication and Information Engineering, Xi'an University of Posts and Telecommunications Xi'an 710121)

Abstract Aiming at the influence of mobile nodes on network topological stability, an adaptive distributed reinforcement learning algorithm is proposed to predict the stable connection of adjacent nodes. Each node uses the method of combining reinforcement learning with adaptive division of learning intervals, uses the received signal strength information between adjacent nodes to determine the connection state between adjacent nodes, and finally predicts the set of neighbor nodes that can maintain stable connection. The simulation results of random walk model under various conditions show that the prediction accuracy is about 95%, which verifies the effectiveness and stability of the algorithm.

Key words distributed; mobile Ad hoc networks; reinforcement learning; topological stability

移动自组织网络 (mobile Ad hoc networks, MANET) 是由移动节点组成复杂分布式系统。移动节点可以自由和动态地自组织成临时网络拓扑结构来传输每个节点收集到的信息。MANET 的特点是有限的存储资源、处理能力以及高度移动性。在网络中，移动节点可以动态地加入或离开网络，导致了频繁和难以预测的拓扑改变，加重了网络任务的复杂程度，降低了网络通信质量。由于网络拓扑结构的不断变化[1-2]，无线链路在高速移动环境中经常发生断裂，如何保持通信链路的持续性成为一个巨大挑战。因此，在临时网络拓扑结构信息交互过程中选择稳定联接链路节点进行传输对于链路联接的持续性有重要意义。

为了增强网络的性能因素，目前最有效方法是通过节点的移动特性来预测网络中链路联接的稳定性程度和网络拓扑结构。文献 [3] 提出了基于自适应神经模糊系统来预测节点的运动轨迹，根据预测得到的轨迹来选择链路节点进行传输。文献 [1] 通过收集节点的接收信号强度指示 (received signal strength indication, RSSI)，将其进行深度学习训练，预测节点的运动轨迹。文献 [4-5] 通过深度学习或机器学习方法对节点的位置进行预测或进行链路质量预测来选择最短可靠路径进行信息传输。文献 [6] 提出一种基于接收信号强度选择稳定路径的

收稿日期：2019－03－25；修回日期：2020－01－06

基金项目：国家科技重大专项 (2017ZX03001012-005)；陕西省教育厅科学研究计划 (17JK0693)；陕西省重点科技创新团队计划 (2017KCT-30-02)

作者简介：黄庆东 (1976－)，男，博士，副教授，主要从事自适应信号处理、无线传感器网络、分布式算法、机器学习等方面的研究. E-mail：huangqingdong@xupt.edu.cn

方法，根据一段时间内节点接收信号强度平均值将链路分为强联接和弱联接两类，设定阈值选择某一阈值内的链路进行路由传输。上述算法在研究方法上不尽相同，但都存在一定的局限性。现有的预测链路稳定性的算法中，大多都是仅考虑节点相对移动性，或仅采集节点某个时期的运动参数，而这些参数不能及时反映节点移动特性的变化，没有考虑对链路稳定性的综合影响。通常在预测节点的未来移动性时需大量的测量数据以及控制信息，这些因素会形成巨大开销造成网络拥塞，降低网络性能。在预测过程中节点移动特性是假设不变的，然而在实际的网络中这些情况都会实时变化，算法不能很好地自适应环境变化。因此，本文提出一种基于强化学习的分布式自适应拓扑稳定性方法，通过对网络中各个邻居节点接收信号强度值自适应学习，得到每个节点对未来链路稳定性和拓扑结构的判断依据，提升网络性能。

本文将接收信号强度与强化学习方法结合，每个分布式节点通过邻居节点的信号强度值进行分布式强化学习，自适应划分区间边界分级处理，形成直接决策区间和自适应强化学习区间，对不同环境下节点的联接状态进行分级判断以及实时更新学习。经过不断学习每个节点得到最优联接策略表，根据策略表中的值预测和判断下一状态的邻居节点联接情况，解决了综合因素对链路稳定性的影响。

1　理论基础及模型

1.1　链路稳定性概念

为了说明链路稳定性研究在移动自组织网络中的重要性，通过图 1 所示场景进行简要说明。从图 1 中可以观察到，移动自组织网络包含 4 个节点 A，B，C，D。节点 A 需要向 D 发送数据包，所以节点 A 广播路由请求分组并发现要发送数据包到 D 必须经过节点 B 或 C。此时节点 B 正迅速远离 A 和 D 节点，而节点 C 缓慢向 A 移动。如果节点 A 选择 B 作为转发节点，由于 B 的移动性，(A, B) 链路不稳定，很容易断开。由于 C 是缓慢向 A 节点移动，所以在传输的过程中 (A, C) 链路相比 (A, B) 将会有更长的时间保持良好稳定联接。A 选择 C 作为下一跳传输节点转发到 D，更有助于信息的可靠网络传输。通过上述场景可以看出，根据平均联接有效时长选择最稳定的路径可以避免未来链路失效，从而改善路由。

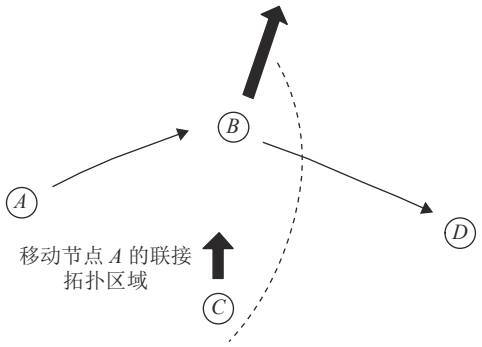

图 1　MANET 移动场景

由于每个节点具有移动性，作为最短路径的一条链路可能在联接建立之后迅速断开。中断的链路会导致路由服务质量下降。因此，在 MANET 中节点之间构建相对稳定的拓扑联接可以避免链路故障，很大程度上改善了网络通信服务质量。

1.2　强化学习基本模型

强化学习算法是一类经典的在线机器学习算法，智能体根据环境状态输入，通过与环境交互得到反馈奖赏来选择当前环境状态的最佳动作[7]。强化学习系统主要包括 5 个部分：环境、状态s、动作a、奖励r和智能体 (Agent)。强化学习以“尝试”的方式进行学习和强化，并形成好的动作策略。整个系统的框架如图 2 所示。

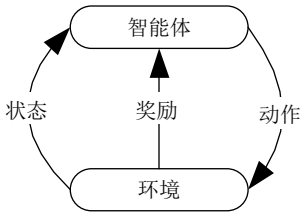

图 2　强化学习基本框架

强化学习是由仿生学习、自动控制等理论发展而来，其基本原理是：如果 Agent 的某个行为策略导致环境正的奖励 (强化信号)，则此行为策略便会加强，反之减弱。Agent 的目标是在每个离散状态学习最优策略使期望奖赏最大化。

强化学习中 Q-learning 算法由于其较好的算法性能，被广泛研究和使用。其状态集S由集合$\{s_1, s_2, \cdots, s_i, \cdots\}$组成，动作集$A$由$\{a_1, a_2, \cdots, a_j, \cdots\}$组成。不同状态动作对$(s_i, a_j)$对应 Q 值矩阵$i$行$j$列的元素，状态动作集对应的 Q 值可表示为 Q 值矩阵。分布式强化学习时，每个节点独立训练学习，并保持一个 Q 值矩阵不断学习更新。定义评估函数值$Q_t(s_i, a_j)$为 Agent 在t时刻状态s_i下选取动作

a_j计算获得的 Q 值，其中$s_i \in S$，$a_j \in A$，并且在下一状态选取最优动作的折扣奖励累积值。在 Q-learning 算法不断的学习过程中，每个网络节点的 Agent 通过递归的方式不断更新该节点 Q 值，以获得最大的长期累积奖励，最终可以得到预期目标下此节点的最佳学习策略。各个节点的 Q 值更新函数为[7]：

$$Q_{t+1}(s_i,a_j)=(1-\alpha)Q_t(s_i,a_j)+\alpha(R_{s_i\to s'_i}^{a_j}+\gamma\max_{a'_j}Q_t(s'_i,a'_j)) \tag{1}$$

式中，α为学习率，$0<\alpha<1$；γ为奖励折扣因子，$0<\gamma<1$；a_j为当前动作，a'_j为策略在s'_i状态上对应的最大 Q 值动作；s_i为当前状态；s'_i为s_i执行动作a_j后转移到的状态；$R_{s_i\to s'_i}^{a_j}$为在状态s_i下执行动作a_j后转移到状态s'_i得到的奖励值；$\max\limits_{a'_j}Q_t(s'_i,a'_j)$表示状态$s'_i$下所有状态动作对中最大 Q 值，代表当前策略取得的新状态最好预期值对当前策略 Q 值计算的影响。

强化学习应用到 MANET 中，多数情况下是解决动态情况下找寻最短路径的问题和解决 QoS 问题[8-10]。本文在强化学习的基础上结合移动自组织网络中节点之间信息交互时携带的 RSSI 值，提出了自适应拓扑稳定性算法寻找稳定链路联接。

2 基于 Q-learning 的分布式自适应拓扑稳定性算法

基于 Q-learning 的分布式自适应拓扑稳定性算法是由强化学习 Q-learning 算法与自适应区间更新算法两种方法结合产生一种预测周围移动邻居节点拓扑稳定联接的算法。该方法利用强化学习思想建立模型，通过实时处理当前节点接收到的邻居节点 RSSI 值进行强化学习，并对此邻居节点的链路联接状态进行预测，每个节点都维护一张状态 Q 值矩阵表以及一个自适应学习区间，根据 RSSI 值来分区间判断当前链路质量，算法的结构框图如图 3 所示。

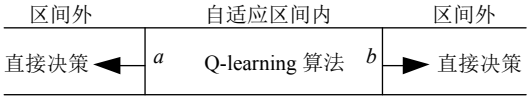

图 3 自适应区间算法结构图

图 3 中，如果当前节点接收到某个邻居节点 RSSI 值处于自适应区间$[a,b]$内，则执行 Q-learning 算法进行联接状态预测；若处于自适应区间外，则执行联接状态直接决策。自适应区间$[a,b]$的边界依据直接决策失误情况进行上、下边界的区间扩展调节。通过两种方法的结合可以提高判决效率，提升算法判决精度，从而使预测模型更加高效、快速适应环境的变化做出准确状态判断。

2.1 自适应区间更新算法结构

自适应区间更新算法服务于 Q-learning 算法，为其提供更适合的强化学习区间范围。本文假设节点发射功率为 0 dBm，考虑环境等因素影响，节点间稳定联接临界强度值为−77 dBm。初始化区间$[a,b]$中上界a与下界b的值都等于−77 dBm，这样形成 3 个区间$[0,a)$、$[a,b]$、$(b,-\infty)$。区间$[a,b]$为 Q-learning 算法学习区间，进行强化学习决策；区间外$[0,a)$、$(b,-\infty)$进行状态的直接决策。随着算法执行，区间$[a,b]$的值不断更新，进行区间扩展。自适应区间更新及决策算法流程如下：

1) 设定初始的阈值$d_{\text{win}}=-77$ dBm，节点根据当前采集到某邻居节点的 RSSI，当大于阈值判定为稳定联接状态s_1，小于阈值判定为非稳定联接状态s_2。状态变量s表示节点与邻居节点的联接状态，表示为：

$$s=\begin{cases}s_1 & d_{\text{win}}<\text{RSSI}\\ s_2 & d_{\text{win}}\geqslant\text{RSSI}\end{cases} \tag{2}$$

2) 当前节点根据其邻居节点的 RSSI，按照式 (2) 进行状态判定，作为下一时刻节点与此邻居节点联接状态的预测$\hat{s}$；假设下一时刻信号强度为 RSSI′，又根据式 (2) 判定下一时刻实际联接状态为s'，若$s'\neq\hat{s}$，则根据情况调整区间$[a,b]$，初始状态$a=b=-77$ dBm。按照流程 1) 判定出错时，若$a<\text{RSSI}$，则调整$a=\text{RSSI}$；若$b>\text{RSSI}$，则调整$b=\text{RSSI}$。直接决策调整边界公式表示为：

$$[a,b]\Leftarrow\begin{cases}\text{上界}a=\text{RSSI}:s'\neq\hat{s}\text{RSSI}>a\\ \text{下界}b=\text{RSSI}:s'\neq\hat{s}\text{RSSI}<b\end{cases} \tag{3}$$

3) 直接决策：按照式 (2) 进行状态直接决策，在直接决策区间$[0,a)$内，直接判决为s_1状态；在直接决策区间$(b,-\infty)$内，直接判决为s_2状态。

4) 节点根据每一邻居节点前后时刻接收信号强度值，按照式 (2) 进行决策区间边界调整；按照流程 3) 进行直接决策区间的状态判定；而对于直接决策区间外的自适应区间$[a,b]$，按照 Q-learning 算法进行强化学习和状态决策，并对 Q 值矩阵进行持续更新。

5) 不同时刻，节点按照流程 2)～流程 4) 邻居节点接收信号强度进行边界循环更新和状态决策。

该算法可以异步分布式执行，网络中各个节点独立按照上述算法进行自主学习决策。每个节点对其各邻居节点进行联接状态稳定关系判定，最终由稳定联接状态的邻居节点构成此节点的稳定邻居集。由相互稳定联接的节点形成移动无线自组织网络的稳态拓扑。

2.2　Q-learning 算法结构

基于 Q-learning 的分布式自适应拓扑稳定性算法中，每一个移动节点可以视为一个 Agent，这样整个网络的动态变化都可认为是一个分布式多 Agent 协作系统。对于每个 Agent，假设其环境状态集为S，动作集为A，奖赏函数为$R^{a_j}_{s_i\to s'_i}$，动作选择策略为$\pi(s_i,a_j)$。根据 Q-learning 算法基本结构描述如下：

1) 状态集S：由离散的状态构成。状态定义为：

$$S=\{s_1,s_2\} \tag{4}$$

式中，s_1状态为根据当前接收到某邻居节点 RSSI，节点与某邻居节点处于稳定联接状态；s_2状态为根据当前接收到某邻居节点 RSSI，与某邻居节点处于非稳定联接状态。

2) 动作集A：每个 Agent 可以采取的动作分为预判稳定联接状态和预判非稳定联接状态两个类型。动作集定义为：

$$A=\{a_1,a_2\} \tag{5}$$

式中，a_1为预判稳定状态；a_2为预判非稳定状态。

3) 奖励函数$R^{a_j}_{s_i\to s'_i}$：$s_i\to s'_i$表示前后时刻的实际状态转移。强化学习过程中，奖励函数是 Agent 在状态s_i下采取行动a_j预判状态，参照实际转移状态s'_i后的奖惩值，它表明在特定状态下采取动作决策的好坏程度。在算法设计的过程中，设定奖赏函数值如表 1 所示，由 8 种情况组成。由于非稳定联接关系误判为稳定联接关系，以及稳定联接关系误判为非稳定联接关系，上面两种情况造成的网络影响相对恶劣，所以加重了奖惩数值，分别予以$R^{a_1}_{s_1\to s_2}=-5$和$R^{a_1}_{s_2\to s_2}=-5$的惩罚。

表 1　奖励函数值表

$R^{a_j}_{s_i\to s'_i}$	a_1	a_2
$s_1\to s_1$	+1	−1
$s_1\to s_2$	−5	+1
$s_2\to s_1$	+1	−1
$s_2\to s_2$	−5	+1

根据表 1 分析，可以得到奖赏函数定义式：

$$R^{a_j}_{s_i\to s'_i}=\begin{cases}-5: s_1\xrightarrow{a_1}s_2; s_2\xrightarrow{a_1}s_2\\ +1: s_1\xrightarrow{a_1}s_1; s_1\xrightarrow{a_2}s_2;\\ \quad s_2\xrightarrow{a_1}s_1; s_2\xrightarrow{a_2}s_2\\ -1: s_1\xrightarrow{a_2}s_1; s_2\xrightarrow{a_2}s_1\end{cases} \tag{6}$$

式中，$s_i\xrightarrow{a_j}s'_i$表示状态s_i采取动作a_j，发生状态转移$s_i\to s'_i$。不同状态执行动作后，根据执行动作的状态预判与转移的实际状态的异同设置奖励值。综合上述状态、动作、奖励值的结构描述得到本文 Q-learning 算法的状态转移图如 4 所示。

4) 动作选择策略$\pi(s_i,a_j)$：Q-learning 算法的策略选择决定了 Agent 怎样去平衡探索和开发之间的问题。Agent 通过探索可以持续学习发现更优的策略；通过开发选择转向期望状态最佳动作。本文算法选择ε−贪心策略来确定最优动作，每次选择 Q 值最大的动作。即：

$$\pi(s_i,a_j)=\arg\max_{a_j}Q(s_i,a_j) \tag{7}$$

5) 更新 Q 值函数：综合动作、奖励值的设计，根据式 (1) 的方法进行函数的更新。

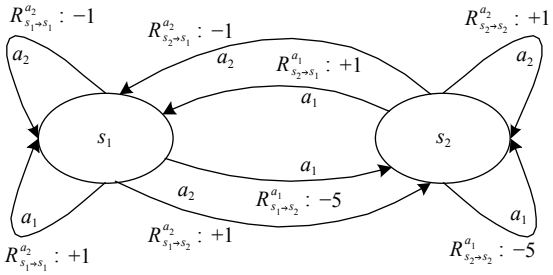

图 4　本文 Q-learning 算法状态转移图

3　仿真与结果分析

为了验证算法有效性和稳定性，通过 Python 仿真环境设计了 3 组实验来研究本文所提出算法的性能。为了能够更加真实地建立 MANET 中节点运动的随机性以及各个节点之间速度以及运动方向的随机性模型，在仿真场景的设计中采用了 MANET 中经典的运动模型——随机游走移动模型[11] (random walk mobility model, RWM) 来验证本文算法性能。

实验设定在 150×150 m^2 的网络区域内生成移动节点，每个移动节点选择随机的方向运动、随机的运动时间、随机的停顿时间，实验中设定节点数目为 15 个且每个节点之间的运动互不影响。表 2

所示为仿真实验的系统参数。

表 2 实验参数设置

参数	值
移动模型	RWM
仿真区域/m²	150×150
节点数目/个	15
节点随机移动速度/m·s⁻¹	[0, 10]
节点随机停顿时间区间/s	[0, 10]
节点随机移动角度区间	[0, 2π]
节点最大通信距离/m	70
仿真时间/s	1 000
采集数据间隔/s	1
临界联接信号强度/dBm	−77

根据上述的仿真参数设定，将本文算法应用到 RWM 移动模型中进行算法的有效性测试。仿真中设定 RSSI 的测量模型为自由空间传播模型[12]，计算公式如下：

$$\text{Loss} = 32.44 + 20\lg d + 20\lg f \tag{8}$$

式中，Loss 是传播损耗，单位为 dB，与传输路径有关；d是距离，单位为 km，f是工作频率，单位为 MHz。假设各个节点发射信号为窄带信号，工作频率为 2 400 MHz，并且发射功率为 0 dBm 时，可以得到RSSI = − Loss，根据节点的最大通信距离 $d = 0.07$时计算得到 RSSI 值为−77 dBm。考虑电磁波在空气中的损耗，设定了可以稳定联接的临界值为−77 dBm。

在算法开始执行前，设定初始的学习迭代次数为 200 轮、通过学习 200 轮之后得到策略表以及强化学习区间，对测试数据进行 100 轮预测来计算准确率，将 100 轮预测的联接状态结果与节点在实际移动过程中各个节点联接状态进行统计平均，计算出每个节点在 100 轮预测过程中的准确率。

图 5 为仿真环境都相同的情况下，分别设定不同学习率α为 0.1、0.5、0.7 的准确率值对比图。

根据图 5 中不同学习率α对准确率的影响曲线分析可知，当学习率α的取值为 0.1 时所有节点的准确率值均维持在 95% 左右，并且各个节点之间的预测准确率变化值相差不大，整个曲线变化比较平缓；而在学习率α取值为 0.5 或 0.7 时准确率比 0.1 时均有所下降，并且各个节点的预测准确率相差变大，曲线的变化程度较明显。出现该现象是由于在执行本文算法进行预测的过程中，节点主要根据邻居节点过去运动经验来判断下一传输时刻联接的状态程度，如果学习率α增大将增大 Agent 的探索过程则对节点的运动经验的取值变小，从而导致节点的预测错误的几率增加。但是在不同学习率α的影响下准确率维持在 0.8～0.95，从而证明算法的稳定性。因此，在接下来的实验过程中均选取学习率α为 0.1 作为本文算法中的参数。

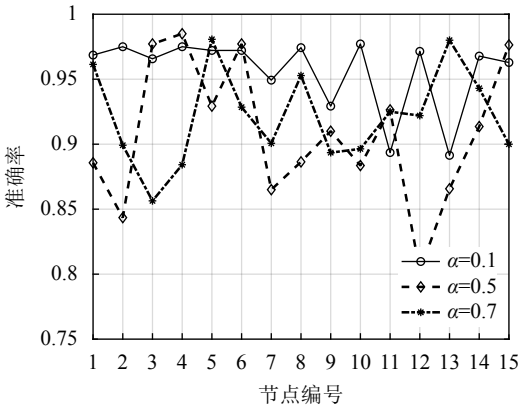

图 5 不同α对准确率影响

为了证明算法的有效性，通过在 RWM 模型中分别应用本文提出的基于 Q-learning 的分布式自适应拓扑稳定性算法与通过强化学习算法直接得到策略表来判断稳定联接次数比较。实验设定两次仿真环境均相同的情况下，分别统计测试数据 100 轮中每个节点预测联接状态的准确次数率。

根据图 6 所示，本文提出的基于 Q-learning 的分布式自适应拓扑稳定性算法的准确率比单独使用 Q 学习算法的准确率整体提高了 30% 左右，故本文算法在预测的准确率方面明显优于单独使用 Q 学习算法，其原因是各个 Agent 通过自适应的强化学习区间的不断更新将每次的学习变化范围扩大，自适应区间外直接判断联接状态，自适应区间内随着不断的强化学习经验的积累做出更加精确地预测，提升算法的性能。两种算法的比较也说明本文算法的有效性。

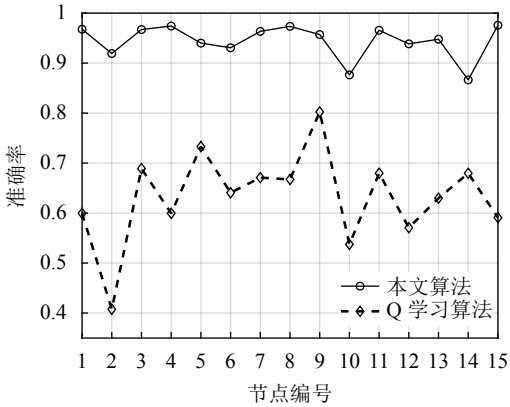

图 6 本文算法与 Q 学习算法准确率比较

图 7 为通过随机的抽取某一轮预测过程中单个节点预测得到的网络拓扑联接关系，与图 8 的节点在实际运动过程中的真实联接关系进行比较。实验仿真环境与上述两个实验相同，仿真中实际联接稳定的阈值设定为$d_{\text{win}} = -77$ dBm，根据设定阈值判断稳定联接邻居节集。

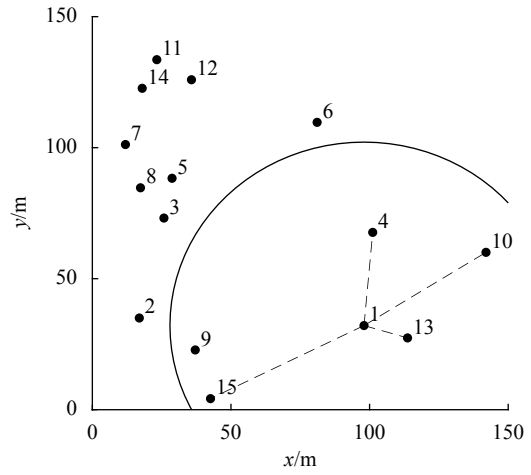

图 7　预测拓扑联接

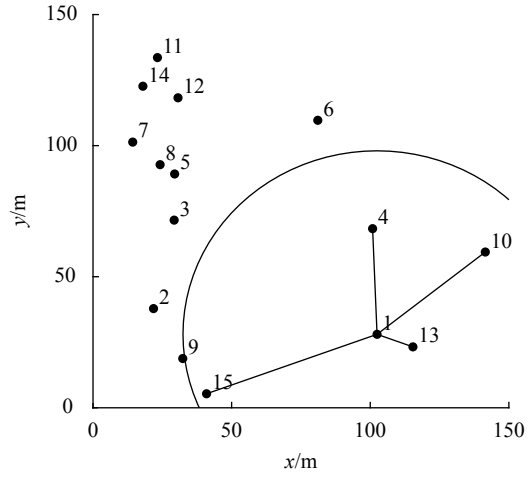

图 8　真实拓扑联接

根据图 7 中处于 1 号节点通信范围内的节点集合为{4,9,10,13,15}，在预测稳定拓扑联接过程中，生成的联接关系集合为{4,10,13,15}，预测出 9 号节点不能在下一传输时刻稳定联接。通过预测拓扑联接关系与图 8 真实拓扑联接关系比较表明，预测结果与真实联接关系相一致。强化学习的过程中每个 Agent 都会对其他节点的运动特性有累积性的学习，不会因为节点处于通信范围内判断为稳定联接链路，Agent 会根据节点当前的运动状态以及策略表中学习得到的经验来有效避免在短时间内可能会快速断开的链路联接，所以 9 号节点在预测过程中被判断非稳定联接状态。

4　结 束 语

本文通过研究 MANET 中移动节点对网络拓扑影响，提出了基于强化学习的分布式自适应算法。算法中每个节点通过对其他节点运动特性学习得到下一传输时刻稳定联接的邻居集合，通过稳定联接集合预测移动节点之间网络拓扑的稳定联接关系，可以更好地适应网络拓扑变化。MANET 中稳定的拓扑联接关系很大程度上改善了路由选择，同时也提高了网络通信服务质量。实验结果表明，基于 Q-learning 的分布式自适应拓扑稳定性算法高效稳定且准确度高，能够有效地实现网络拓扑联接的稳定性选择。

参　考　文　献

[1] YAYEH Y, LIN H, BERIE G, et al. Mobility prediction in mobile ad-hoc network using deep learning[C]//The 2018 IEEE International Conference on Applied System Invention. Japan: IEEE, 2018: 1203-1206.

[2] MAYADUNNA H, SILVA S L D, WEDAGE I, et al. Improving trusted routing by identifying malicious nodes in a MANET using reinforcement learning[C]//The 2017 Seventeenth International Conference on Advances in ICT for Emerging Regions. Colombo, Srilanka: IEEE, 2017: 1-8.

[3] ELLEUCH M, KAANICHE H, AYADI M. Exploiting neuro-fuzzy system for mobility prediction in wireless ad-hoc networks[C]//International Work-Conference on Artificial Neural Networks. Palma de Mallorca, Spain: Springer, 2015: 536-548.

[4] KAANICHE H, KAMOUN F. Mobility prediction in wireless ad hoc networks using neural networks[J]. Computer Science, 2010, 8(1): 95-97.

[5] LIU L, CHENG Y, CAI L, et al. Deep learning based optimization in wireless network[C]//The 2017 IEEE International Conference on Communications. Paris, France: IEEE, 2017: 1-6.

[6] 夏辉, 贾智平, 张志勇, 等. 移动 Ad Hoc 网络中基于链路稳定性预测的组播路由协议[J]. 计算机学报, 2013, 36(5): 926-936.
XIA Hui, JIA Zhi-ping, ZHANG Zhi-yong, et al. A link stability prediction-based multicast routing protocol in mobile Ad Hoc networks[J]. Chinese Journal of Computers, 2013, 36(5): 926-936.

[7] 梁志伟, 朱松豪. 基于强化学习的类人机器人步行参数训练算法[J]. 计算机工程, 2012, 38(8): 13-15.
LIANG Zhi-wei, ZHU Song-hao. Walking parameters training algorithm of humanoid robot based on reinforcement learning[J]. Computer Engineering, 2012, 38(8): 13-15.

[8] MAMMERI Z. Reinforcement learning based routing in networks: review and classification of approaches[J]. IEEE Access, 2019, 7: 55916-55950.

[9] KAVALEROV M, LIKHACHEVA Y, SHILOVA Y. A reinforcement learning approach to network routing based on adaptive learning rates and route memory[C]// Southeastcon.[S.l.]: IEEE, 2017: 1-6.

[10] ALHARBI A, ALDHALAAN A, ALRODHAAN M. A mobile ad hoc network q-routing algorithm: Self-aware approach[J]. International Journal of Computer Applications, 2015, 127(7): 1-6.

[11] CAMP T, BOLENG J, DAVIES V. A survey of mobility models for ad hoc network research[J]. Wireless Communications and Mobile Computing, 2002, 2(5): 483-502.

[12] 熊皓. 无线电波传播 [M]. 北京: 电子工业出版社, 2000.
XIONG Hao. Radio propagation[M]. Beijing: Publishing house of electronics industry, 2000.

编 辑 税 红