基于Q学习和动态权重的改进的区域交通信号控制方法

张 辰 喻 剑 何良华
(同济大学电子与信息工程学院 上海 400047)

摘 要 Q学习在交通信号控制中具有广泛的应用。在区域交通中,基于Q学习的传统区域交通信号控制方法通过agent之间互相交流的方式获取周边路口信息,并作出最有利的决策。传统交通控制方法在大部分情况下具有良好的表现。然而,由于其对周边路口拥堵程度的回馈计算不准确,因此在周边路口堵塞程度相差较大时将出现决策失误,从而导致局部热点拥堵。针对该问题进行分析,并以传统的区域交通信号控制方法为基础,提出一种新的基于Q学习和动态权重的改进的区域交通信号控制方法,引入"路口权重"的概念,通过多目标组合法将其应用于回馈计算,且权重随路口实际交通情况动态改变,解决了易陷入局部热点拥堵的问题。应用仿真软件在3种不同的交通状况下进行模拟,结果表明,所提算法在"拥堵"的状况下较传统控制方法具有更突出的表现。

关键词 Q学习,区域控制,路口权重
中图法分类号 U491.5+1 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.8.035

Promoted Traffic Control Strategy Based on Q Learning and Dynamic Weight

ZHANG Chen YU Jian HE Liang-hua
(Institute of Electronic and Information, Tongji University, Shanghai 400047, China)

Abstract Q-Learning is widely used in traffic signal control. In traditional multi-agent traffic signal control policy, agents gain intersection information via network, and make the best control decision. It works well in most cases. But traditional policy has a weakness that the global reward is calculated by simple average. This may cause local block in some cases. This paper introduced a promoted area traffic signal control based on Q learning. "Intersection Weight" is used in the new calculation method, which varies dynamically according to the real traffic condition. Both traditional and promoted methods were used to experiment. The results show the advantage of the promoted one.

Keywords Q learning, Area traffic control, Intersection weight

1 引言

当今,汽车已经成为人们出行的主要工具。随着汽车的普及,城市路口建设的局限性问题愈发突出,如何有效管理交通成了一个非常重要的课题。近年来,随着计算机科学的发展,智能化成为交通控制的主要特征。为了解决路口交通问题,国内外许多学者进行了大量研究。其中,强化学习算法(Reinforcement Learning,RL)[1,2]被认为是一种利用自学习来解决交通控制的有效方法。Q学习[3]则是强化学习中最具代表性的算法,其在区域交通控制中具有较好表现。其中,Bazzan等人[4,5]阐述了在区域交通信号控制中互相协作的必要性。SCOOT[6]、SCATS[7]是已经比较成熟的强化学习技术,其作为中心化控制方法的代表,已经在实际生产中投入使用。然而,中心化控制对硬件要求较高,需要以极快的速度传输、接收以及处理数据。Abdulhai [8]将强化学习算法引入交通信号控制,并且证明了Q学习作为强化学习中的代表性算法可以为自适应交通信号控制提供很好的解决方案。Q学习的模型无关性、非监督学习等特性令其十分适合实际交通控制。Araghi[9]、Lu Shou-feng[10]各自提出了一种基于Q学习的单路口交通信号控制,并获得了良好的效果。Prabuchandran[11]与Kar[12]证明了在区域多路口中应用强化学习算法的可行性,为Q学习扩展到多路口区域控制打下了理论基础。Abdoos[13]将单路口Q学习控制方法扩展到区域中,每个节点以本地路口信息为输入,控制本地路口,却忽略了路口之间的影响。Weiring[14,15]对多节点Q学习交通控制模型进行了研究,并得出其可以有效控制信号的结论。Xu Lun-hui[16]对使用Markov过程进行交通信号控制的方法进行了分析与阐述,提供了理论依据。Chanloha[17]提出了一种基于优先度的Q学习控制策略。Arel[18]将Q学习应用到区域中,通过agent互相传递reward信息,达到区域控制的目的。但是,其使用简单平均的方法计算周边路口的reward值,该计算方法容易忽略单个路口的堵塞情况,从而导致一些路口产生堵塞,并成为区域交通的瓶颈。

本文针对传统区域信号灯控制方法容易产生局部堵塞的问题展开研究,分析局部堵塞产生的根本原因,并对此提出一种新的基于Q学习的改进的区域交通信号控制方法。引入

到稿日期:2015-05-27 返修日期:2015-08-24
张 辰(1991—),男,硕士,主要研究方向为智能交通控制;喻 剑 男,博士,副研究员,主要研究方向为智能交通控制;何良华 男,博士,副研究员,主要研究方向为多媒体研究。

"动态路口权重",权重值随交通情况实时改变。基于路口权重和新的评价函数,对 Q 学习控制中的 state,action,reward 以及 Q 值更新函数进行修改。新的信号控制方法可以有效地避免局部堵塞情况的出现。

本文第 2 节对传统区域交通控制模型进行了描述;第 3 节详细描述了经过改进的基于 Q 学习的算法;第 4 节给出了实验的设计和结果,并将实验结果与传统控制方法进行比较,给出了结果分析;最后对方法进行了总结。

2 传统控制方法

2.1 基于 Q 学习的控制

Q 学习是 RL 算法中具有代表性的一种学习算法。该算法的依据为 Markov 决策过程[19,20]。相对其它 RL 算法,Q 学习有 3 个很重要的优势。第一,模型无关性。建立区域路口模型非常复杂。Q 学习算法不需要建立详细可靠的模型,由此也避免了因模型的不完善而带来的各种问题。第二,Q 学习是非监督学习[21]。监督学习[21]需要大量的案例作为训练集,通过反复训练获得最优模型,因此模型的优劣在很大程度上取决于训练集的选取。如果训练集没有完善的案例覆盖,那么所训练出来的模型依旧是有缺陷的。第三,Q 学习是一种真正意义上的自适应学习。其可以通过在线学习与真实交通环境进行互动来达到最优化的目的。

在区域路口交通信号控制中,Q 学习具有良好的应用。以 Arel[18] 中的传统区域控制方法为例,其采用单节点控制单路口的方案。如图 1 所示,每个控制节点配有环境探测器,且通过网路与周边路口的控制器连接。环境探测器实时监测本交叉口的交通状况信息,并反馈给本地控制器。周边路口的控制器将自身的信息通过网路传输给本地控制器。

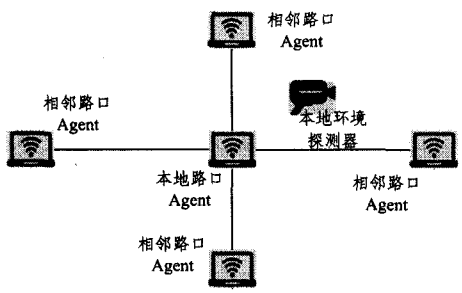

图 1 控制节点架构示意图

Q 学习控制流程如图 2 所示。

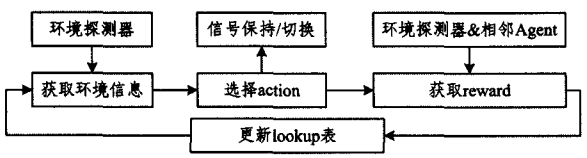

图 2 一轮 Q 学习迭代

Step1 agent 通过环境探测器获取当前路口的环境信息。

Step2 根据环境信息查找 lookup 表,找到最有利于当前环境的 action,并通知信号灯保持/改变当前相位。

Step3 一定时间后,从环境探测器获得本地路口的 local reward,从相邻路口处获得相邻路口的 global reward,并进行数据处理。

Step4 根据 Q 值更新函数,更新 lookup 表中的相应项。至此完成一轮迭代。

2.2 传统控制方法存在的问题

传统的区域控制方法以简单平均的方法处理 global reward 值。在大部分情况下,该方法具有良好的表现。但是,在实际计算时,若其中的 3 个路口具有很低的 reward 值(越高的 reward 值代表越差的堵塞情况),而另一个路口具有较高的 reward 值,则它们的 reward 平均值有可能仍显示为"不堵塞"(高 reward 值被低 reward 值抵消)。因此,实际存在的某个相邻路口的堵塞因平均值而被忽略,该堵塞将持续维持甚至增长,最终出现区域交通瓶颈。

考虑如下场景:主路口 X_0 的 4 个周边路口为 X_1,X_2,X_3,X_4。在某次信号灯切换中,若 X_0 保持当前相位,将收到 1,2,1,8 的 reward 值。若 X_0 切换相位,将收到 3,3,4,4 的 reward 值。根据"reward 值越大,交通情况越糟",将得出保持相位更有利的判断。但是,如此将令 X_4 陷入及其糟糕的交通情况,且堵塞会持续恶化,并以 X_4 为中心向四周扩散,从而导致局部堵塞。

实际上,若假设车流具有随机性,则在周边 4 个路口通行能力相差过大时,容易出现以上情况。如,X_4 的通行能力远远不如 X_1,X_2,X_3,从而相同的交通流分流对后三者仅造成轻微的堵塞,对 X_4 却造成致命的伤害。

3 改进的控制方法

3.1 评价函数

评价函数用以评价路口交通情况水平。交通协调模型的评价指标一般为平均延误、总行程时间、流量、平均速度或平均等待时间等性能指标中的一项或几项,控制模型根据交通控制目标构造评价函数,评价函数用以评价路口交通状况。本方法选用综合评价指标函数,使用权重的方法组合多个评价目标。评价函数值越高,则路口交通情况越差。

选用的目标如下。

1)平均延误(Average Delay Time)

定义:当一辆车从进入交叉口区域一直到离开,由于交叉口不畅通导致的额外行驶时间。延误时间通常由停车时间和缓慢行驶所浪费的时间组成。平均延误计算公式如下:

$$\bar{D}=\frac{1}{n}\sum_{i=1}^{n}d_i \tag{1}$$

其中,$\bar{D}$表示平均延误,n 表示车辆总数,d_i 表示每辆车的延误时间。

2)平均速度(Average Speed)

定义:车辆在经过该路口区域时的平均速度。平均速度计算公式如下:

$$\bar{S}=\frac{1}{n}\sum_{i=1}^{n}s_i \tag{2}$$

其中,$\bar{S}$表示平均速度,n 表示车辆总数,s_i 表示每辆车的延误时间。

3)排队长度(Queue Length)

定义:红灯方向队列上的平均排队长度。排队长度计算公式如下:

$$\bar{L}=\frac{1}{2}(L_1+L_2) \tag{3}$$

其中，$\bar{L}$表示平均排队长度，L_1，L_2 分别为两个方向上的红灯排队长度。

评价函数如下：

$$\bar{E}=\alpha\bar{D}+\beta\bar{S}+\gamma\bar{L} \tag{4}$$

其中，α，β，γ 分别为相应的权重。

3.2 路口权重

传统信号控制方法容易产生局部热点拥堵的原因在于在计算周边 reward 信息时简单地使用了平均值$(r_1+r_2+r_3+r_4)/4$，从而会出现高 reward 值被低 reward 值抵消，无法发现潜在的高拥堵危险。

为了解决该问题，引入“路口权重”的概念。在计算周边 reward 值时，为每一个路口赋予权重，权重随实际交通情况动态改变。权重值越高，代表路口越容易陷入堵塞。在进行计算时，权重将令脆弱路口得到更高的关注度，从而决策时更偏向保护脆弱路口。

权重与交叉口车流量系数正相关，车流量系数的计算公式为：

$$y=\frac{q}{s} \tag{5}$$

其中，q 表示经过的车流量，s 表示饱和流量。整个交叉口的 Y 值是各车道的 y 值之和，$Y=\sum y$。

在实际情况中，车流量系数在一定程度上反映了当前交叉口的堵塞情况，也反映了当前交叉口在未来所能额外承受的车流压力。将其与 reward 值结合，可以大幅提高堵塞路口的受关注程度，及时发现因简单平均计算方法而隐藏的堵塞风险，避免堵塞局势的进一步恶化。

改进后的周边 reward 信息为：

$$(y_1r_1+y_2r_2+y_3r_3+y_4r_4)/4 \tag{6}$$

3.3 改进的 Q 学习算法

3.3.1 改进的 lookup 表

在区域交通控制中，相邻路口之间由于共用交通流的原因，势必互相产生影响。单路口的相邻路口越多，其所受到的影响也越大。因此，控制节点在调度时必须要考虑到周边路口的情况。为了解决这一问题，需要修改 lookup 表，令其同时包含本地路口情况和周边路口情况。改进后的 lookup 表如表 1 所列。

表 1 改进的 lookup 表

state		action	Q value	
local	global	action	local	global
ls []	gs []	1	Lv_1	Gv_1
ls []	gs []	0	Lv_2	Gv_2

由表 1 可知，传统 lookup 表由 3 部分构成：state 部分存储环境信息，action 部分存储决策行为，Q value 部分存储 Q 值信息。

为了将 Q 学习应用在区域交通信号控制算法中，对其中的 state 和 Q value 进行修改，各自细分为 local 部分与 global 部分。local 部分存储本地交叉口信息，global 部分存储相邻交叉口信息。

state 部分的存储形式为一串向量。向量由多个变量构成，描述了交叉口的交通环境信息，如下式所示，其中 s_1，s_2，s_3 等各代表一项环境变量。

$$ls=[s_1,s_2,s_3\cdots]$$

action 部分存储决策行为。在交通控制中，行为分为“保持当前相位”和“切换当前相位”。分别以 0/1 值存储。

Q value 部分存储对应的 Q 值。其中，local Q value 存储本地路口 Q 值。global Q value 存储周边路口 Q 值。

3.3.2 改进的 Q 值更新函数

对 Q 值更新函数的修改主要涉及 reward 部分。reward 向量同时包含本地路口(r_{Local})和周边路口信息(r_{Global})。其中，本地路口 reward 可以直接由探测器获取。周边路口 reward 则为周边所有路口 reward 值的加权平均值。

改进的 Q 值更新函数如下：

$$Q_{t+1}(s_t,a_t)=(1-\alpha_t)Q_t(s_t,a_t)+\alpha_t[r_{t+1}+\gamma\max Q_t(s_{t+1},a_{t+1})] \tag{7}$$

其中：

$$Q(s_t,a_t)=[Q_{Local}(s_t,a_t),Q_{Global}(s_t,a_t)] \tag{8}$$

$$r_{t+1}=[r_{Local},r_{Global}]\alpha_t \tag{9}$$

$$r_{Global}=\sum_{i=1}^{n}\frac{h_i\cdot r_i}{n} \tag{10}$$

式(8)表示 Q 值以一对变量的形式存在。$Q_{Local}(s_t,a_t)$表示本地路口的 Q 值，$Q_{Global}(s_t,a_t)$表示周边路口的 Q 值。

式(9)表示 reward 值 r_{t+1} 以一组向量的形式存在。r_{Local} 表示本地路口的 reward 值，r_{Global} 表示周边路口的 reward 平均值。

式(10)表示 r_{Global} 的计算方法为所有周边路口的加权平均值。其中，n 表示周边路口总数，r_i 为路口 i 的经过转换的 reward 值，h_i 为路口 i 的权重。

此处 r_i 的意义为本地路口所作出的 action 对相邻路口 i 产生的影响的度量。然而，如果简单使用路口 i 的本地 reward 值是不准确的。因为一个交叉口的 reward 值是所有相邻路口车流共同作用下的结果。因此，应当尽量保留本地路口对路口 i 所造成的影响，而剔除其余部分。在改进的控制方法中，相邻路口 A_i 接收来自主路口 A 的 reward 的计算方法为：

$$R_i=\frac{q_i}{\sum q}R \tag{11}$$

其中，R_i 为相邻路口 A_i 接收到的来自主路口 A 的 reward 值。q_i 为路口 A_i 对路口 A 的车流量贡献。$\sum q$ 为进入主路口 A 的所有车流量。R 为主路口 A 的本地 reward 值。

3.3.3 改进的 action 策略

由于 lookup 表的 Q 值部分由 local 与 global 两项构成，因此需要对 action 选择算法进行修正。修正的后算法如下。

算法 1 action 选择算法(参数请参考表 1)

```
1. if(|lv_1-lv_2|≫|gv_1-gv_2|)
2.    choose min(gv_1,gv_2);
3. if(|lv_1-lv_2|≫|gv_1-gv_2|)
4.    choose min(lv_1,lv_2);
5. if(min(lv_1,lv_2)≪min(gv_1,gv_2))
6.    choose min(lv_1,lv_2);
7. if(min(lv_1,lv_2)≫min(gv_1,gv_2))
8.    choose min(gv_1,gv_2);
9. choose min(lv_1,lv_2,gv_1,gv_2);
```

对于步骤 1－4：考虑同一类型(local 或者 global)的两个

值的差(i.e.|lv_1−lv_2|,|gv_1−gv_2|)。差值越大,说明不同的action对同一类型路口造成的影响差别越大。假设local差值远小于global差值,则说明action的不同选择对本地路口的影响远大于对相邻路口的影响,此时应当选择对本地路口有利的action。

对于步骤5−8:若两者的差值比较相近,则说明action的不同选择对同一类型路口的影响相对平衡。此时应当考虑不同action所能带来的效益。若其中某一个Q值非常小,则说明其可以大大改善路口情况,因此选择对应的action。

对于步骤9:若4个数值均在一个较小范围内波动,则说明路口改善情况对action的选择的敏感度不高。此时综合考虑local和global值,并做出最有利的选择。

3.4 Q学习参数

3.4.1 state

state参数保存了对当前环境状态的描述。state值由以下参数构成:

1)本路口当前信号相位s_local;

2)相邻路口当前信号相位s_global;

3)本路口红灯方向上队列red_1,red_2;

4)绿灯方向上通行车流量green_1,green_2。

其中,对相位有如下规定:

①一共有4种相位。

②相位切换的顺序是固定的,不可以直接跳过中间的相位。

③每一个相位都有最小时间和最长时间的限制,以保证不会出现"饿死"现象。

四基本相位如图3所示。

图3 四基本相位

红灯方向的排队长度在一定程度上反映了处于红灯相位的车流的紧迫程度。排队长度越长,说明该方向上紧迫程度越高,越需要切换当前相位至通行。

绿灯方向的车流数量在一定程度上反映了处于绿灯相位的车流的通行量水平。通行量越大,说明该方向上车流量越大,越需要保持当前绿灯相位。

3.4.2 action

action表示控制节点对当前环境作出的行为选择。由于采用可变周期,因此通用的描述每相位增加/减少时间的行为并不适合。本算法采用的行为描述为"(保持/切换)当前相位"。相对前者,后者不仅简单,适合可变周期,并且避免了lookup表状态空间的爆炸问题。

3.4.3 reward

reward函数描述了在控制节点对当前环境做出选择之后环境对选择行为的回报。回报函数记录了环境是否改善,以及改善程度的信息。回报由评价函数衡量。回报越高,表示改善程度越低。

3.5 参数离散化

Q学习无法处理连续空间中的数据。为了解决这一问题,需要将连续数据离散化。在离散化的同时也需要避免状态空间的爆炸。过大的状态空间即使可以勉强工作,也势必会导致过长的运行时间。在非常注重时间因素且不容出错的实时交通控制中,过慢的响应速度是不可忍受的。因此,离散化程度必须合理,既不能令系统响应速度过于缓慢,又要保证Q学习最终结果的质量。

3.6 探索策略

在使用Q学习进行交通控制时,每一时刻Agent都会查找lookup表,找到符合当前state的、最大的Q值所对应的action,并选择此action作为行为,这是最优化的行为策略。

然而,当仍处于学习的过程中时,以上默认规则却存在一个潜在的危险,即:为了每次可以获得最大的回报,Q学习都选择最优Q值对应的行为,那么其他的行为将一直被忽略,从而它们所对应的Q值无法得到及时更新,造成误差。从某种意义上来说,这时也陷入了"局部最优"的困境。

为了解决这一问题,需要在"利用(Exploitation)"的同时进行"探索(Exploration)"。探索的意义在于,在进行行为选择时不选择最优Q值,而是抱着让所有$Q(s,a)$对都达到优化的目的,以一定概率选择其他的action,使得Q学习可以得到全面的经验知识。

关于"探索",有两种机制可以选择:ε-greedy机制和Boltzmann分布机制。这里选择后者作为探索机制。

Boltzmann分布机制如下。

对于给定的随机系数$T>1$,在状态s下,动作a被选择到的概率由式(12)决定:

$$P_r\{a_t=a|s_t=s\}=\frac{e^{Q(s,a)/T}}{\sum_{a'\in A}e^{Q(s,a')/T}} \tag{12}$$

由式(12)易知,$\sum P_r\{a_t=a|s_t=s\}=1$,并且按照轮盘赌的方法进行动作的选择。A是所有可以执行的动作的集合。其步骤如下。

Step1 在[0,1]之间产生一个随机数ran,且令$sum=0$,$i=1$。

Step2 $sum\leftarrow sum+P_r\{a_t=a|s_t=s\}$。

Step3 若$ran\leqslant sum$,则被选中的动作为$i=1$,否则转Step2,且$i\leftarrow i+1$。

通常,Boltzmann分布机制中的温度控制系数T称为探索力度参数,这是因为它的取值越大,算法越偏向于随机"探索"。在智能系统学习早期,lookup表中的元素并没有达到成熟,"利用"它进行动作决策并不能得到很好的回报,应该偏重于"探索"新知识,完善lookup表中的元素,随着学习的进行,再渐渐转向于依靠lookup表进行动作决策。据此,ε和T的取值可以采取随时间推移而递减的形式,以便更好地实现从"前期重探索"到"后期重利用"的平衡过渡。

4 实验及分析

4.1 仿真设计

本实验在不同程度的交通拥堵下进行测试。实验场景设计为一块由27个交叉口组成的区域,区域是对上海市松江区部分的模拟,如图4所示。图中标出的路口为监测路口,其中圆圈标注的为正常路口,五边形标注的为通行能力较差的脆弱路口。

为了验证本方法应用的控制效果，设计了多个具有不同车流量的场景，车流量依次递增，并且将结果与传统区域控制方法所得结果进行比较。

实验采用随机车流，针对每一种车流量场景均测试 10 轮，每轮 180 分钟。

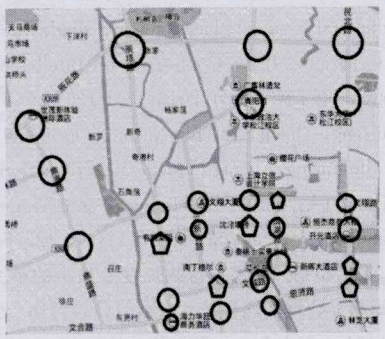

图 4　实验区域模拟图

表 2 为各场景车流数据。其中，每个方向上的车流为该方向上每个入口的平均车流量。

表 2　各场景车流数据(veh/h)

场景	东	西	南	北
1	1200	800	1000	900
2	1100	1000	800	1200
3	1400	1600	1300	1700
4	1800	2000	1900	1800
5	2200	2100	1900	1900
6	2200	2500	2500	2200
7	2300	2600	2900	2800
8	2800	3100	3300	2900
9	3500	3200	3300	3400

4.2　结果分析

实验结果如表 3 所列。表 3 记录了整个区域中所有路口的平均交通情况。表 4 记录了所有脆弱路口的平均交通情况。实验对平均延迟、平均速度和平均排队长度 3 个参数进行了记录。

表 3　区域平均数值

场景	平均延迟(s)		平均速度(km/h)		平均排队(m)	
	传统	改进	传统	改进	传统	改进
1	6.15	6.98	58	56	27	26
2	7.24	7.32	57	57	29	31
3	9.31	9.21	50	51	38	37
4	13.27	13.98	43	45	46	45
5	16.41	15.88	38	40	66	61
6	21.88	20.32	31	34	81	75
7	27.36	25.77	25	30	96	87
8	35.89	31.26	21	25	110	98
9	46.22	41.68	17	22	126	109

表 4　脆弱路口平均数值

场景	平均延迟(s)		平均速度(km/h)		平均排队(m)	
	传统	改进	传统	改进	传统	改进
1	7.02	7.09	58	58	28	27
2	7.85	8.13	57	58	31	32
3	11.68	10.85	49	51	41	39
4	17.01	14.53	39	43	56	49
5	22.37	18.26	31	36	72	67
6	30.06	24.78	23	27	88	78
7	38.12	31.22	17	23	106	91
8	47.55	39.52	14	17	134	113
9	56.23	45.35	12	14	153	130

对其中典型的 3 个场景进行分析，分别为“空闲”(场景 2)、“正常”(场景 5)和“拥堵”(场景 8)。

首先对脆弱路口平均数值进行分析。从表 4 中可以发现，在处于空闲状态时，传统区域控制方法和改进方法的实际测试效果基本一致。在正常状态时，两种方法具有较小的差异。而在拥堵情景下，改进的方法具有最好的优化效果。在重度拥堵的情景下，两种方法的控制效果反而缩小。

在空闲和正常情景下，经过的车流量对通行能力较小的路口而言不会造成严重的拥堵，因此各路口的堵塞情况较轻。从算法角度分析，由于在较好的交通环境，返回的 reward 值处于较低的数值，因此不会出现因简单平均计算方法造成的误区，并且“路口权重”在低堵塞的情况下不具有明显的区分度。

而在拥堵的模拟情景下，改进的方法具有最好的优化效果(见图 5)。在较高的车流量下，脆弱路口将首先出现拥堵情况，但周边正常路口尚未到达通行量的临界值。此时，若应用传统的控制方法，脆弱路口的高 reward 值将被正常路口的低 reward 值抵消。由于错误的信息反馈，主交叉口将无视脆弱路口的拥堵，从而令其交通状况持续恶化，并首先成为区域中的瓶颈。而在改进的控制方法中，较高的权值动态提升了脆弱路口的受关注度，加强了对脆弱路口的保护，使其不至于过早陷入“局部堵塞”。

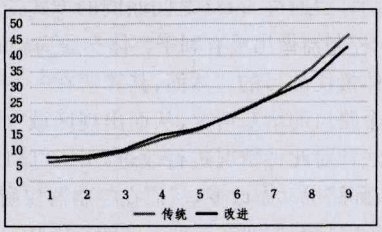

图 5　9 轮不同拥堵程度下的平均延迟差异

图 6 记录了在拥堵情景下，区域平均延迟指标的变化。随着拥堵情景进一步升级，区域进入重度拥堵的状态。此时改进方法的优化效果变差。从算法的角度分析，在传统的控制方法中，随着脆弱路口的堵塞开始扩散，正常路口的车流量也已达到临界值，“平均值陷阱”逐渐消失。而在改进的控制方法中，过差的交通环境导致路口堵塞情况差距缩小，从而“路口权重”的作用也随之减小。因此，在重度拥堵的情景下，由于外界施加的过大的车流压力，改进的控制方法的优化效果出现变差的趋势。

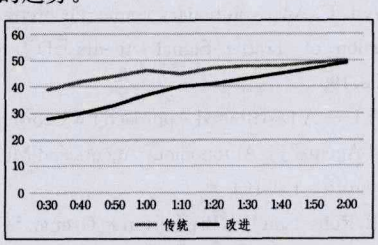

图 6　拥堵情景下区域平均时延变化

其次对整体区域平均数值进行分析，如表 3 所列。整体区域中，两种方法的结果差异较小。从算法的角度分析，在传统的控制方法中，由于其忽视脆弱路口，容易形成以之为中心的“局部堵塞”，而堵塞的效应随着向周边正常路口扩散而逐渐减轻。如图 7 所示，六角星为产生堵塞的脆弱路口，而三角

形为陷入“平均值陷阱”的路口。相反，在改进的方法中，控制器根据路口权重合理选择策略，其不仅依据路口的平均改善情况，更要防止脆弱路口出现严重堵塞的情况。因而在有些情况下，宁愿牺牲路口平均利益，也要保证脆弱路口的交通情况，不令其成为整个区域的瓶颈。

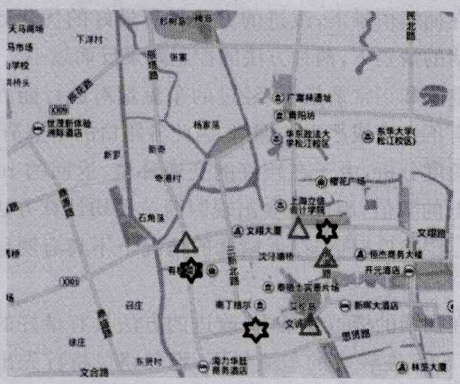

图 7　传统控制方法实验结果

综上，改进的控制方法首先保证对脆弱路口的保护，避免区域出现“局部堵塞”的现象，其次才追求整体区域的优化效果。因此，改进的控制方法在拥堵的情况中具有较好的优化效果。

结束语　Q 学习算法在区域交通信号控制中具有良好表现。传统控制算法通过 agent 之间联网的方式传递相邻路口信息，有利于控制器做出最有利于整体交通局势的决策，从而达到区域整体最优的目的。然而，其算法在一些特殊场景下容易忽略单个路口的堵塞情况，从而出现区域“局部堵塞”的问题。本算法针对此种情况进行改进，引入“路口权重”的概念，权重随实际路况动态改变，从而在控制器权衡时得到更多的关注。最后将控制方案在 3 种不同的交通状况下进行仿真模拟，并与传统 Q 学习控制方案进行比较，实验证明本方案具有更好的控制效果。

参 考 文 献

[1] Sutton R S, Barto A G. Introduction to reinforcement learning [M]. MIT Press, 1998

[2] Sutton S. Introduction: The challenge of reinforcement learning [M]. Reinforcement Learning. Springer US, 1992: 1-3

[3] Watkins C J C H, Dayan P. Q-learning[J]. Machine Learning, 1992, 8(3/4): 279-292

[4] Bazzan A L C. An Evolutionary Game Theoretic Approach for Coordination of Traffic Signal Agents [D]. University of Karlsruhe, 1997

[5] Bazzan A L C. A Distributed Approach for Coordination of Traffic Signal Agents[J]. Autonomous Agents and Multi-Agent Systems, 2005(10): 131-164

[6] Hunt P B, Robertson D I, Bretherton R D, et al. SCOOT-a traffic responsive method of coordinatingsignals[D]. United Kingdom, 1981

[7] Sims A G, Dobinson K W. The Sydney Coordinated Adaptive Traffic (SCAT) system philosophy and benefits [J]. IEEE Trans. Veh. Technol., 1980(29): 130-137

[8] Abdulhai B, Pringle R, Karakoulas G J. Reinforcement learning for true adaptive traffic signal control[J]. Journal of Transportation Engineering, 2003, 129(3): 278-285

[9] Araghi S, Khosravi A, Johnstone M, et al. Q-learning method for controlling traffic signal phase time in a single intersection[C]// 16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013). 2013: 1261-1265

[10] Lu Shou-feng, Zhang Shu, Liu Xi-min. On-line Q Learning Model for Minimizing Average Queue Length Difference of Single Intersection[J]. Journal of Highway and Transportation Research and Development, 2014, 31(11): 116-122(in Chinese)
卢守峰，张术，刘喜敏. 平均排队长度差最小的单交叉口在线 Q 学习模型[J]. 公路交通科技，2014，31(11)：116-122

[11] Prabuchandran K J, Kumar H, Bhatnagar A N, et al. Decentralized learning for traffic signal control[C]// 2015 7th International Conference on Communication Systems and Networks. 2015: 1-6

[12] Kar S, Moura J M F, Poor H V. Distributed reinforcement learning in multi-agent networks[C]// 2013 5th IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP). 2013: 296-299

[13] Abdoos M, Mozayani N, Bazzan A L C. Traffic light control in non-stationary environments based on multi agent Q-learning [C]// 2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC). IEEE, 2011: 1580-1585

[14] Wiering M. Multi-Agent Reinforcement Learning for Traffic Light Control. Machine Learning[C]// Proceedings of the Seventeenth International Conference (ICML'2000). 2000: 1151-1158

[15] Wiering M, et al. Intelligent Traffic Light Control[R]. Technical Report UU-CS-2004-029, University Utrecht, 2004

[16] Xu Lun-hui, Xia Xin-hai, Luo Qiang. The study of reinforcement learning for traffic self-adaptive control under multiagent Markov game environment[J]. Mathematical Problems in Enginee-ring, 2013, 2013(6): 1-10

[17] Chanloha P, Chinrungrueng J, Usaha W, et al. Cell Transmission Model-Based Multiagent Q-Learning for Network-Scale Signal Control With Transit Priority[J]. Computer Journal, 2014, 57 (3): 451-468

[18] Arel I, Liu C, Urbanik T, et al. Reinforcement learning-based multi-agent system for network traffic signal control[J]. Intelligent Transport Systems, IET, 2010, 4(2): 128-135

[19] Puterman M L. Markov decision processes: discrete stochastic dynamic programming[M]. John Wiley & Sons, 2009

[20] Papadimitriou C H, Tsitsiklis J N. The complexity of Markov decision processes [J]. Mathematics of Operations Research, 1987, 12(3): 441-450

[21] Rasmussen C E, Williams K I. Gaussian processes for machine learning[M]. The MIT Press, 2006