基于 Qlearning 的单点信号配时方案选择算法

朱海峰[1,2]
(1.中国电子科技南湖研究院，浙江 杭州 310012;2.浙江海康智联科技有限公司，浙江 杭州 310012)

摘要：为提高单点控制交叉口时段内信号配时的准确性，采用强化学习方法构建时段内信号配时优化模型。该模型以时段内原始固定信号配时方案为基准，向其上下区域探索建立状态空间及动作空间，同时以时段内交通状态为依据，设置常规及异常状态开关，用于区分学习常规及异常状态下 Q 值表，并在回报函数上进行特别设置，以快速响应交通的短期突变及长期缓慢变化，减少因交通环境变化导致 Q 值表不能及时适应交通状况的现象。最后通过仿真对该算法的有效性进行验证，结果表明采用该算法能生成合理的配时方案，且可将交叉口车辆总延误降低 24%。

关键词：城市交通；交叉口；信号配时；Qlearning
中图分类号：U491.5 **文献标志码：**A **文章编号：**1671－2668(2022)01－0044－04

当前，交叉口信号控制多采用多时段固定配时方案，随着交通环境的长期或短期变化经常会无法适应交通需求，引起不必要的延误甚至部分时段拥堵。因此，对时段内的配时方案进行实时优化非常必要。常见的实时优化方法未根据反馈进行学习，且计算过程复杂或变化过于灵活不安全，不利于实施和流程化操作，不能完全满足动态交通信号配时的需要。该文兼顾信号配时优化的稳定性和灵活性，以交叉口时段内原始固定配时方案为基准，向其上下两个安全区域进行搜索和选择，实现控制的稳定性，同时对时段内相对长久缓慢或异常变化作出及时响应，体现控制的灵活性，从而稳定、灵活地改善路口交通运行状况。

1 基本原理

如图 1 所示，将交叉口看作智能体，通过对时段内交通环境状态的判别，区分常规与异常状态，选择并执行相应状态下配时方案动作行为，作用于当前交通环境，分析交叉口状态，并依据状态给出相应的奖励或惩罚反馈。该奖励或惩罚用于强化环境状态与最佳方案选择之间的映射关系，反复执行这个映射过程，学习模型即可获得时段内常规与异常环境状态下选择最佳方案的能力。

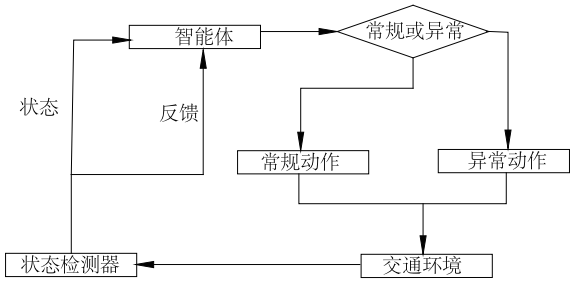

图 1 单点信号配时方案选择原理示意图

1.1 状态空间定义

为描述常规与异常状态，将状态空间定义为 $S=(C,F)$，其中 C 代表状态集，F 代表状态开关。

为使算法能快速收敛并迅速响应交通环境变化，对状态集进行简化设计。以某一时段内运行的固定配时方案为基准，向其上下 2 个方向各拓展 2 套方案。状态集 C 共设置 5 套方案，即 $C=(P_1,P_2,P_3,P_4,P_5)$，其中 P_3 为原始基准方案，P_1 为向下拓展方案 1，P_2 为向下拓展方案 2，P_4 为向上拓展方案 4，P_5 为向上拓展方案 5。以城市主干道上三相位十字交叉口为例进行状态集 5 套方案的设置展示：P_3 方案各相位时长分别设置为 54、34、44 s，周期为 132 s；P_1 各相位时长分别设置为 30、24、32 s，周期为 86 s；P_2 各相位时长分别设置为 43、29、40 s，周期为 112 s；P_4 各相位时长分别设置为 56、35、46 s，周期为 137 s；P_5 各相位时长分别设置为 59、36、47 s，周期为 142 s。

为区分交通状态异常与否，设置针对常规与异常状态的开关量 F，表达式如下：

$$F=\begin{cases}1, \dfrac{y_{\text{now}}-y_{\text{last}}}{y_{\text{last}}}>e\\ 0, \dfrac{y_{\text{now}}-y_{\text{last}}}{y_{\text{last}}}\leqslant e\end{cases} \tag{1}$$

式中：y 为交叉口的关键流量比，以三相位交叉口为例，$y=(q_1+q_2+q_3)/s$；q_1、q_2、q_3 分别为相位 1、2、

3 中关键车流的流量;s 为车道的饱和流量,这里假设各车道的饱和流量相同;y_{now} 为当前关键流量比;y_{last} 为数据分析得到的该时段公允关键流量比,与 y_{last} 同比增大超过 e 可判定为异常状态,e 可根据实际交叉口情况进行设置。

1.2　动作空间定义

在单点信号配时优化中,一个完整的动作空间包括交叉口在一个时间步内所有可能的动作,也就是所有可能的信号配时方案。考虑到动作空间太大会影响算法的收敛速度,将动作简化为 5 套方案的选择。动作空间定义为 $A=(P_1,P_2,P_3,P_4,P_5)$,其中 P_1、P_2、P_3、P_4、P_5 与状态空间中的 P_1、P_2、P_3、P_4、P_5 相同。每套方案中各相位时长已确定(实际应用中,相位可根据各相位关键车流的流量比进行分配调节)。

为简化算法,将异常状态和常规状态下动作空间设置成一样,动作空间需同时覆盖常规及异常状态下配时方案空间(实际应用中,可根据常规状态和异常状态分别设置动作空间)。

1.3　回报函数

回报函数可选取延误时间、停车次数、排队长度等指标值计算得到,指标值可通过仿真软件直接获得。这里选取交叉口车辆平均延误作为评价指标。

首先通过聚类算法分析得到交叉口该时段内不同类别延误变化范围的上限值 d。如图 2 所示,类别为 0 上的"★"代表正常延误类别的聚类中心点,×代表 80%分位上正常延误值上限;为 1 上的"★"代表异常延误类别的聚类中心点,×代表 80%分位上异常延误值上限。正常延误上限 d 为 44 s,异常延误上限 d 为 66 s。

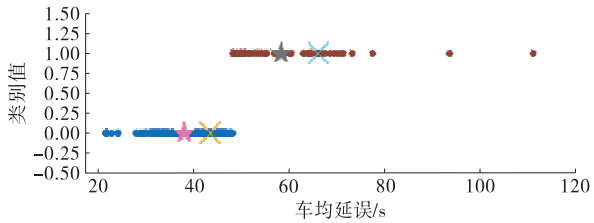

图 2　交叉口车均延误聚类结果

奖惩函数为:

$$r_t(s,a)=\begin{cases}-1, d_{t0}<d_{tk} \text{ 或 } d_{tk}>d\\ 1, d_{t0}\geqslant d_{tk}\end{cases} \tag{2}$$

式中:d_{t0} 为动作执行前的延误;d_{tk} 为动作执行后的延误。

为防止因交通自身的波动性导致延误突变及奖惩函数产生振荡,设置连续相同动作标志 b。若连续 2 次相同的动作,则 $b=2$;若连续 3 次相同的动作,则 $b=3$;以此类推,每增加 1 次连续相同动作,b 值加 1;连续动作被中断,则 $b=1$。

针对不同的 b、$r_t(s,a)$、$dif(dif=d_{tk}-d_{t0})$ 调整反馈 $r_t(s,a)$,规则如下:

$$r_t(s,a)=\begin{cases}0; r_t(s,a)=-1, dif<10, b=2\\ -1; r_t(s,a)=-1, dif\geqslant 10, b=2\end{cases} \tag{3}$$

如式(3)所示,$b=2$ 且 $r_t(s,a)=-1$ 时,说明被选中的方案已连续 2 次被选中,该算法动作选择策略采取贪婪算法,据此可知被选中的方案曾经是一套相对优秀的方案,或许是由于交通的波动性导致延误升高。延误升高幅度不大,即 $dif<10$ 时,可修正 $r_t(s,a)=0$;延误升高幅度较大,即 $dif\geqslant 10$ 时,可保持 $r_t(s,a)=-1$。

如式(4)所示,$b>2$ 且 $r_t(s,a)=-1$ 时,说明被选中的方案已连续 3 次以上被选中,同理可知被选中的方案已是比较优秀的方案,或许是由于交通的波动性或交通环境变化导致延误升高。延误升高幅度不大,即 $dif<10$ 时,可保持 $r_t(s,a)=-1$;延误升高幅度较大,即 $dif\geqslant 10$ 时,可修正 $r_t(s,a)=-b+1$,加强环境变化的反馈值。

$$r_t(s,a)=\begin{cases}-1; r_t(s,a)=-1, dif<10, b>2\\ -b+1; r_t(s,a)=-1, dif\geqslant 10, b>2\end{cases} \tag{4}$$

如式(5)所示,$r_t(s,a)=2$、$b=2$ 时,重新设置 $b=1$,防止紧接着出现相同动作时,随着 b 的升高,出现修正 $r_t(s,a)=-1$ 甚至是更小的负值,从而产生强烈振荡,导致出现不收敛的情况。

$$r_t(s,a)=2; r_t(s,a)=1, b=2 \tag{5}$$

1.4　Q 值表的更新

Q 值的更新采用 Bellman 最优方程:

$$Q_{t+1}(s_t,a_t)=(1-\alpha_t)Q_t(s_t,a_t)+\alpha_t(r_{t+1}+\gamma Q_{t\max}(s_{t+1},a_{t+1})) \tag{6}$$

该算法需建立 2 张 Q 值表,一张用于记录常规交通,另一张用于记录异常交通,其他参数统一设置。取 $\alpha=0.5$、$\gamma=0.9$(α 为学习速率,其值越大,保留之前训练的效果越少;γ 为折扣因子,其值越大,之前训练的效果所起的作用越大。可根据具体交叉口特性按需选择)。动作选择策略采用贪婪算法,即 $\varepsilon-$greedy 探索策略,通过设置自增加的 ε 值 $\in$

[0.7,0.9]与随机生成数 $r\in[0,1]$，比较大小进行学习动作选取。选取规则：$r<\varepsilon$ 时，选取当前状态下 Q 值最大的动作；$r\geqslant\varepsilon$ 时，在当前状态下随机选取一个动作执行。ε 自增加规则：迭代次数 $n\leqslant500$ 次时，$\varepsilon=0.7+0.2/500n$；$n>500$ 次时，$\varepsilon=0.9$。

该算法中，Q 值表为 5×5 矩阵(见表 1)，其中 s_1、s_2、s_3、s_4、s_5代表 5 种状态(为便于理解，采用传统强化学习中状态表示字母 s，对应的状态为 P_1、P_2、P_3、P_4、P_5)，a_1、a_2、a_3、a_4、a_5代表 5 种动作(为便于理解，采用传统强化学习中动作表示字母 a，对应的动作为 P_1、P_2、P_3、P_4、P_5)。

表 1 Q 值的初值

状态	a_1	a_2	a_3	a_4	a_5
s_1	0	0	0	0	0
s_2	0	0	0	0	0
s_3	0	0	0	0	0
s_4	0	0	0	0	0
s_5	0	0	0	0	0

根据 Bellman 方程，在迭代中更新各状态下不同行为的 Q 值，让各状态下最优行为的 Q 值取得最大值，这样最优行为的选择概率越来越大，而非最优行为的概率越来越小，在 Q 值矩阵最终收敛后，确保在每种状态下都能大概率地选择最优行为。

相比一般强化学习算法，该算法无论是常规 Q 值表还是异常 Q 值表都能在收敛后迅速响应交通环境的长期缓慢变化。因为收敛后最优行为大概率会被连续选中，若最优行为已与交通环境不再匹配，将因为被连续选中而受到不断加重的惩罚，从而迅速减小被选中的概率，直到达到新的收敛状态。

2 仿真试验

采用 VSC++编程实现状态判定及方案控制和选取，在 Vissim 平台构建仿真环境进行仿真，分别采集该算法及固定配时方案下延误、流量等数据，比较分析该算法的有效性。

2.1 模型构建

选取黄山天湖路－科学大道路口及周边相关路口(见图 3)，在 Vissim 中构建道路环境，构建一个南北双向六车道、东西双向三车道的交叉口。

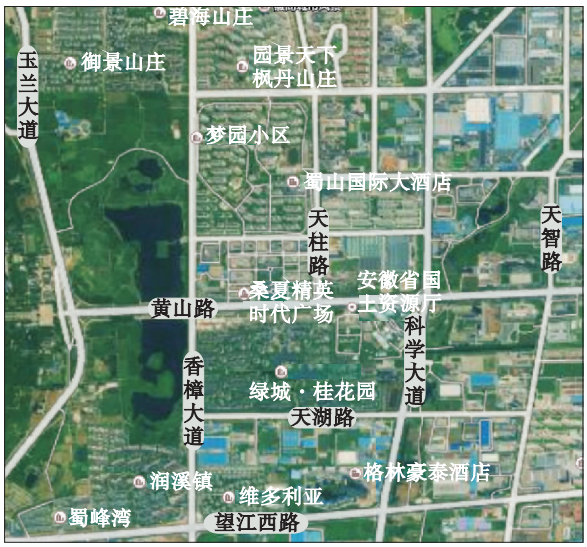

图 3 路口关系示意图

设置 3 个相位对该交叉口进行控制，南向北放行左直右为第一相位 Phase1，北向南放行左直右为第二相位 Phase2，东西向无左转放行直右为第三相位 Phase3。状态空间共 5 套方案(P_1、P_2、P_3、P_4、P_5)，方案的切换通过 VSC++代码控制，信号方案的切换依据 Q 值表规则。最后通过延误检测器统计得到的平均延误对两方案进行评价。

2.2 试验结果对比分析

时段内交通流量统计结果及车道数见表 2。

表 2 各进口交通流量统计结果及车道数

进口	交通流量/(veh·h^{-1})	车道数/条
东进口	300	2
南进口	800	4
西进口	300	2
北进口	400	4

单点交叉口的算法程序 Q 值数表初始值设置见表 1，通过运行算法代码对 Q 值表进行更新，学习结果同样是 5×5 矩阵(见表 3)。

交叉口方案在迭代中依据学习规则进行不断切换，通过设置不同学习次数 N 的值，可获取不同收敛情况的 Q 值表。N 取值越大，Q 值表收敛情况

表 3 训练收敛中的 Q 值表

状态	a_1	a_2	a_3	a_4	a_5
s_1	5.691 076	2.180 493	3.640 528	2.951 589	4.173 591
s_2	0.000 000	3.581 156	−0.299 468	0.000 000	3.953 558
s_3	6.498 757	0.000 000	0.000 000	2.909 429	0.801 303
s_4	5.075 686	0.000 000	0.000 000	0.000 000	−0.050 000
s_5	5.944 814	0.000 000	0.000 000	0.000 000	2.695 478

越好，但相应耗时会增加。试验中选取 N 值为 540 次，每迭代 30 次统计一次总平均延误 D_i，迭代 540 次，共统计 18 次总平均延误进行对比。

$$D_i = \sum_{m=30i+1}^{30i+30} \bar{d} \tag{7}$$

式中：D_i 为第 i 个 30 次迭代车均延误的和；m 为迭代次数（$m \in [1,540]$）；i 为统计次数标记（$i \in [1,18]$）；$\bar{d}$ 为每次迭代的车均延误。

该算法与固定配时方案的总平均延误对比见图 4。

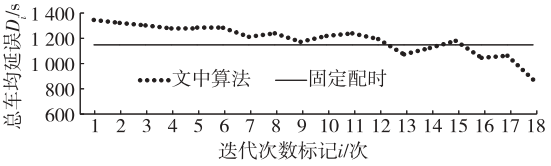

图 4　文中算法与固定配时延误对比曲线

由表 2 和图 4 可知：固定信号配时与交通环境较匹配。该算法在学习 480 次后才在总平均延误上达到固定信号配时的控制效果，主要原因在于该算法的信号配时方案是不断地在原固定信号配时方案上下进行探索，易导致仿真中交叉口车流出现波动现象，而偶尔的波动就会给交通延误指标造成较大影响；迭代 480 次后，该算法在延误指标上趋于稳定并优于固定信号配时方案；迭代 540 次后，总平均延误为 871.992 25 s。固定配时方案的总平均延误为 1 147.871 719 s，相对于固定配时，该算法的总平均延误减少 24%，明显优于固定信号配时。

3　结语

运用 Q 学习算法设计单点信号配时方案选择算法，以交叉口平均延误的相关规则作为评价回报值，利用 Qlearning 进行 Q 值矩阵的收敛性学习，实现对交叉口信号配时方案选择的学习。通过对 Vissim 进行二次开发，将该算法与交叉口原固定信号配时进行仿真对比，结果表明该算法经过学习后的总平均延误优于固定信号配时，对单点交叉口方案选择具有一定的适用性。

参考文献：

[1] 朱海峰，刘畅，温熙华，等.均衡流量和饱和度的交通瓶颈控制[J].控制理论与应用，2019，36(5)：816－824.

[2] 朱海峰，刘畅，刘彦斌，等.基于通行能力系数优化的道路交叉口单点动态控制研究[J].公路与汽运，2019(5)：29－34.

[3] 王祉祈，赵颢，马健霄，等.基于 Q－learning 算法的单点信号控制研究[J].物流工程与管理，2021，43(4)：93－95＋109.

[4] 沈玲宏，赵颢，张梦凡，等.小型平立复合式交叉口交通设计及适用性[J].物流工程与管理，2020，42(4)：133－136.

[5] 刘皓，吕宜生.基于深度强化学习的单路口交通信号控制[J].交通工程，2020，20(2)：54－59.

[6] 舒凌洲，吴佳，王晨.基于深度强化学习的城市交通信号控制算法[J].计算机应用，2019，39(5)：1495－1499.

[7] 李珣，刘瑶，周健，等.基于改进遗传算法的交通信号配时优化模型[J].工业仪表与自动化装置，2017(4)：125－130.

[8] 王铁鹏.单交叉口配时优化的函数逼近型强化学习模型[D].长沙：长沙理工大学，2017.

[9] 李志强.Q 学习在单路口交通信号控制中的应用研究[D].长沙：长沙理工大学，2010.

收稿日期：2021－06－09

（上接第 37 页）

北京：北京交通大学，2014.

[3] 张玉茹，赵戊辰，李晖，等.智能停车场停车诱导方法研究[J].哈尔滨商业大学学报（自然科学版），2015，31(6)：732－734＋740.

[4] 史未名.停车诱导信息板版面设计研究[D].北京：北京工业大学，2010.

[5] BRAAKSMA J P，COOK J W.Human orientation in transportation terminals[J].Transportation Engineering Journal，1980，106(2)：189－203.

[6] SENEVIRATNE P N，MARTEL N.Criteria for evaluating quality of service in air terminals[J].Transportation Research Record：Journal of the Transportation Research Board，1994，1461：24－30.

[7] 陈振武，陈小鸿，熊文.基于视域叠加分析的导向标志设计评价[J].城市轨道交通研究，2009，12(4)：19－24.

[8] 姜军，陆建，李娅.基于驾驶人视认特性的城市道路指路标志设置[J].东南大学学报（自然科学版），2010，40(5)：1089－1092.

[9] 许金良，王荣华，冯志慧，等.基于动视觉特性的高速公路景观敏感区划分[J].交通运输工程学报，2015(2)：1－9.

[10] 韩磊.基于驾驶员视觉特性的草原公路交通标志信息量研究[D].呼和浩特：内蒙古农业大学，2020.

收稿日期：2021－03－17