

基于 Q-学习算法的交通控制与诱导 协同模式的在线选择

杨庆芳^{1,2}, 杨朝²

(1. 吉林大学 汽车动态模拟国家重点实验室, 长春 130022; 2. 吉林大学 交通学院, 长春 130022)

摘要:采用 Q-学习算法实现了交通控制与诱导协同模式的在线选择。首先,采用 Q-学习算法训练多智能体,根据多智能体内部的推理得到不同交通状态下的最优协同模式,最终实现交通控制与交通诱导协同模式的在线选择与转换。仿真结果表明,本文提出的基于 Q-学习算法的协同模式选择方法在一般交通拥挤状态下具有较好的协同控制效果,对比离线式模式选择方法更能适应交通状态的不断变化,从而达到有效避免严重交通拥堵、改善路网性能的目的。

关键词:交通运输工程; 交通控制与诱导协同; 模式选择; Q-学习算法; 回报函数

中图分类号:U491 文献标志码:A 文章编号:1671-5497(2010)05-1215-05

On-line selection method of the traffic control and route guidance collaboration mode based on Q-learning algorithm

YANG Qing-fang^{1,2}, YANG Chao²

(1. State Key Laboratory of Automobile Dynamic Simulation, Jilin University, Changchun 130022, China; 2. College of Transportation, Jilin University, Changchun 130022, China)

Abstract: The on-line traffic control and route guidance collaboration mode selection was realized by the Q-learning algorithm. Using the multi-intelligence agents trained with the Q-learning algorithm, the optimal collaboration mode was obtained under different traffic conditions according to the inner inference of the multi-intelligence agent. So, the on-line selection and switching of the traffic control and route guidance collaboration mode was accomplished. The simulation results show that the proposed collaboration mode selection method based on the Q-learning is characterized by better collaboration control effect under the ordinary traffic congestion condition and more adaptive to constantly changing traffic condition than the traditional off-line mode selection method. The proposed method is helpful to avoiding the heavy traffic congestion and improving the traffic network performance.

Key words: engineering of communications and transportation; collaboration of traffic control and route guidance; mode selection; Q-learning algorithm; reward function

收稿日期:2009-11-05.

基金项目:“863”国家高技术研究发展计划项目(2007AA12Z242).

作者简介:杨庆芳(1966-),女,教授,博士生导师. 研究方向:智能运输系统. E-mail:yangqf@jlu.edu.cn

近年来,关于交通流诱导与控制策略协同的研究已取得许多成果。国内的徐立群^[1]针对我国城市目前的交通管理系统提出一种依托道路交通控制中心的交通流诱导与控制的协同机制,将交通信息共享作为实现两系统协同的基础;李瑞敏等^[2]对基于多智能体系统的城市交通控制与诱导集成化进行了研究。

国外的 Szepesvari^[3]在一定条件下证明了 Q 学习的收敛速度;Werbos 等^[4]通过将强化学习与最优控制理论和动态规划联系起来而进行了理论上的研究;Celine^[5]采用强化学习算法实现了匝道自适应控制,强化学习算法在交通领域取得了明显的效果。

由于交通系统复杂多变,影响因素众多,精确的数学模型难以求得最优解,并且传统的离线式协同模式选择方法不能应对路网上的偶发性交通拥挤和突发性交通事故。因此,有必要设计一种在线的交通控制与诱导协同模式选择方法。本文基于这种应用需求,提出了一种基于 Q 学习的交通控制与诱导协同模式在线选择方法。在实现智能化模式选择的功能上,选择基于 Agent 的 Q 学习算法,把交通控制与诱导协同的几种模式切换作为 Agent 的动作,把基于速度和饱和度的协同小区交通状态作为 Agent 对环境的认识,把综合交叉口饱和度、通行能力、排队长度、延误时间、出行费用等指标的路网性能函数作为回报函数,设计了智能化协同模式选择的过程,给出了强化学习算法实现协同模式选择的步骤。

1 多智能体和 Q-学习算法

1.1 多智能体

城市交通协同管理系统的拓扑结构具有分布式特性,使其很适合应用多智能体的分布式处理和协同技术。采用多智能体技术,系统具有很强的鲁棒性和可靠性,例如,某些 Agent 不能正常工作时,系统的整体性能也不会显著下降或引起系统崩溃,并具有较高的问题求解率。

本文构建了一种基于多智能体的交通控制与交通诱导协同的递阶体系结构,如图 1 所示。按照 Agent 的功能将 Agent 分为协同层 Agent、战略级 Agent 和战术级 Agent 三类。并选择 Q 学习作为 Agent 的学习方法以适应复杂的交通状态变化,完成协同模式的在线选择。



图 1 基于多智能体的控制和诱导协同框架

Fig. 1 Framework of the collaboration of traffic control and route guidance systems based on multi-agent

1.2 Q-学习

Q 学习是一种基于随机动态过程的不依赖精确数学模型的强化学习方法。强化学习也称“再励学习”,是人工智能领域中的介于极端的监督学习和非监督学习之间的自主学习方法。它通过不断尝试错误,从环境中得到奖惩的方法来自主学习到不同状态下哪个或哪些动作具有最大值,即能获得最大奖励的策略,如图 2 所示。



图 2 Agent 与环境的交互关系

Fig. 2 Relationship between Agent and environment

假设 S 表示可观察的环境状态 s 的集合, A 表示智能体可能采取的动作 a 的集合, 回报函数 $r(s, a)$ 表示 Agent 在状态 s 执行动作 a 得到的奖赏或者惩罚。强化学习的任务就是学习控制策略 $\pi: S \rightarrow A$, 即在依次改变环境状态时选择动作, 从而使获得的奖赏值之和最大。

强化学习中的 Q 学习算法通过直接优化一个可迭代计算的动作值函数(Q 函数)找到一个策略,使得期望折算报酬总和最大。这里的动作值函数被定义为:它的值是从状态 s 开始并使用 a 作为第一个动作时可获得的最大期望折算积累回报^[6]。

$$Q(s, a) = r(s, a) + \gamma \max_{a'} Q[\delta(s, a), a'] \quad (1)$$

式中: $\delta(s, a)$ 为新的状态; γ 为折算因子; Q 为在状态 s 执行动作 a 的预期值。

在 Q 学习中通常使用概率的途径来选择行

为。有较高 Q 值的行为被赋予较高的概率,但所有行为的概率都非 0。这种概率的计算式为

$$P(a | s) = \frac{k^{Q(s,a)}}{\sum k^{Q(s,a)}} \quad (2)$$

式中: $P(a | s)$ 为 Agent 在状态 s 时选择行为 a 的概率; $k > 0$ 为一常量, 它确定此选择优先考虑高 Q' 值的程度。

Q-学习是一种无模型增强学习形式, 是根据将状态-行动映射为期望返回值的行动的价值函数 Q 求解具有不完整信息的马尔可夫行动问题的一种简单方式。在学习过程中, Q-学习仅需要评价系统控制性能的再励信号, 采用状态到动作映射的学习过程, 不需要在线学习训练样本, 这种学习方法不仅能逼近最优值函数, 而且容易根据 Q 值求得相应的最优策略, 从而适合于系统的在线实时控制。

强化学习算法计算速度快, 不需要精确的数学模型, 也不需要任何先验知识, 不受维数的限制, 适合用在一体化控制系统中。因此, 利用 Q-学习算法可以有效解决交通控制与诱导协同模式在线选择的问题。

2 协同模式的选择

考虑协同模式影响因素, 得到了基于 Q-学习的协同模式选择思路: Agent 根据路网所处交通状态的不同确定不同的协同模式, 并对所选择的协同模式进行惩罚或奖励, 最后得到一个累积回报最大的策略, 即为一定状态下的最优协同模式。

2.1 状态描述

Agent 需要知道当前路网的交通状态, 依据交通状态选择最佳的协同模式。决定道路交通状态的参数有交叉口饱和度、路段车辆行驶的平均速度、道路的占有率, 而对协同小区来说, 还会有关键交叉口的饱和程度、主干道的拥挤程度、发生拥挤的交叉口数和发生交通事故的路段或交叉口等, 所有这些状态都在一定程度上影响了小区的交通状态。如此多的状态全放到 Q-学习中会耗费 Q-学习很多时间。本文选取影响协同小区交通状态的两类因素, 并以类 state 封装所有的这两个类别。本文采取以区域的关键交叉口来表征小区的交通状态, 选择饱和度和平均车速两个量来进行判别。

交通环境的状态是连续的, 强化学习解决的却是离散空间的求解, 为此, 需要把交通状态离散

化。在本文的离散过程中, 将选择 0.1、5 为区间值离散饱和度和平均速度。离散后的饱和度和速度值如表 1 所示。

表 1 离散的交通状态值

Table 1 Values of discrete traffic state

平均饱和度	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
离散值	1	2	3	4	5	6	7	8	9
平均速度	15	20	25	30	35	40	45	50	55
/($\text{km} \cdot \text{h}^{-1}$)									60
离散值	1	2	3	4	5	6	7	8	9
									10

由饱和度的 9 种状态和平均车速的 10 种状态可以得到 $10 \times 9 = 90$ 种可能存在的状态, $S = \{s_{11}, s_{12}, \dots, s_{ij}, \dots\}, i = 1, 2, \dots, 9, j = 1, 2, \dots, 10$ 。

几点说明如下:

(1) 表 1 中的饱和度和平均速度是整个协同小区内所有关键交叉口及主要路段的平均值。

(2) 文献[7]已经对关键交叉口的交通状态判别做了深入研究, 本文在其基础上, 把研究范围扩大到协同小区。

$$S = \sum_{i=1}^m s_i / m \quad (3)$$

式中: S 为某协同小区的总饱和度, 本文采用的是各个关键交叉口的饱和度之和的平均值; m 为协同小区内关键交叉口的个数; s_i 为关键交叉口 i 的总饱和度, 是指饱和程度最高的相位所达到的饱和度值, $s_i = \max x_j, j = 1, \dots, n, n$ 为交叉口 i 的相位数, x_j 为相位饱和度, 是实际到达流量与允许通过能力的比值。

(3) 假设所有关键交叉口对整个协同小区的交通状态的影响是相同的。

2.2 动作集合

Agent 相应控制器的指令如下: 中心控制级 Agent 进行协同模式选择都由控制器 B 完成, B 发出的行动指令为 a_0, a_1 , 其中 a_0 表示进行协同模式的转换, 例如: 由独立式转向偏重式; a_1 表示保持原来的协同模式不变。 a 的值取 0、1、2、3、4、5, 分别代表递阶式、偏重控制式、协作式、一体化式、偏重诱导式、独立式六种协同方式。协同方式转换的控制过程如下: 当 a 选定一个值之后, 控制器通过指令 a_0 和 a_1 来控制协同模式的转换, 当控制器选择 a_1 时, a 值保持不变, 当控制器选择 a_0 时, a 值被赋予新的值, 之后在新的 a 值下重复上面的过程。

2.3 确定回报

Agent 的主要任务就是完成在不同交通状态下的协同模式的选择, 得到一个动作序列, 使路网的性能函数值最小^[8], 并且最终要选出在一定交通状态下的最优协同模式。因此, 如何评价最优模式成为 Agent 主要关心的问题。本文设计的路网综合运行指标 PI 为

$$PI = \sum_{i=1}^n [w_i^{(1)} D(\xi) + w_i^{(2)} S(\xi) + w_i^{(3)} Q(\xi) + w_i^{(4)} \frac{1}{C(\xi)}] + \sum_{i=1}^m [k_i^1 T(\xi) + k_i^2 M(\xi)] \quad (4)$$

式中: $w_i^{(1)} \sim w_i^{(4)}$ 分别为第 i 个路口的车辆延误时间、停车次数、排队长度、交叉口通行能力的加权系数; $D(\xi)$ 为交叉口平均延误时间; $S(\xi)$ 为交叉口平均停车次数; $Q(\xi)$ 为交叉口排队长度; $C(\xi)$ 为交叉口的通行能力; k_i^1 、 k_i^2 分别为第 i 条路径上车辆的行程时间和出行费用的加权系数; ξ 为优选参数; $T(\xi)$ 为路段的行程时间; $M(\xi)$ 为路段的出行费用。

基于以上的性能函数设计的回报函数为

$$r = \begin{cases} 1, & p_j^k - p_j^{k-1} < 0 \\ 0, & p_j^k - p_j^{k-1} > 0 \\ 10, & \text{最优策略下的奖赏值} \end{cases} \quad (5)$$

式中: p 为路网性能函数; r 为在状态 s_j 下采用动作 a_j 得到的回报值。

2.4 算法流程

Step1 对于每对 (s, a) , 初始化 Q 表中对应的 $Q'(s, a)$ 的入口为 0.1。

Step2 观察当前交通状态 s_j , 重复做以下几个步骤, 直到得到 Q_{\max} 。

(1) 按一定的规则选择协同模式 a 的值, 并执行相应的协同策略。

(2) 计算回报值 r 。

(3) 转移到新的状态。

(4) 按照下式更新 $Q'(s, a)$ ^[9]:

$$Q'(s, a) = (1 - \alpha) \cdot Q(s, a) + \alpha \cdot (r + \gamma \cdot Q_{\max}) \quad (6)$$

Step3 得到 Q_{\max} 及相应的最优协同策略。

3 仿真验证

本文设计了一个对比实验, 对离线和在线两种形式的协同模式选择效果进行评价。以长春市人民大街、自由大路、亚泰大街、南湖大路等路段组成的交通小区为背景建立仿真路网。

为简化计算, 仿真过程作了以下两点假设:

(1) 只研究主干路, 忽略一些对结果分析影响不大的次干路, A 路口简化为十字交叉口。次干路与主干路交叉口、次干路与次干路交叉口均不设信号灯。

(2) 假设诱导服从率为 0.4。

各个主干路均为双向 8 车道的路段, 并进行了主要的渠化, 分为左、直、右三个进口道, 每条车道的饱和流量 1800 pcu/h。一个信号周期内总损失时间为 12 s。路网结构及各个路段的长度如图 3 所示, 仿真时长为 60 min, 采样间隔为 180 s。

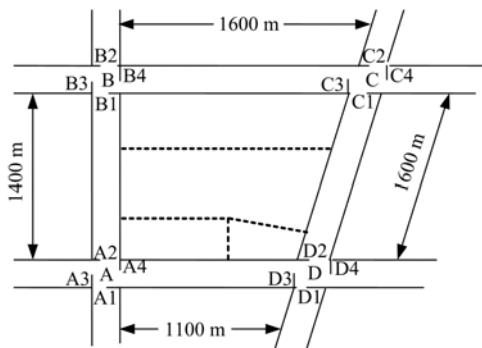


图 3 仿真路网结构

Fig. 3 Structure of emulating traffic network

方案一是离线式的模式选择方法, 本文选择的是偏重控制式。方案二用 Q-学习训练一个控制器, 采用 VC++ 6.0 的编程环境, 编写了算法程序, 并与 Vissim 相接, 实现 Q-学习的在线模式选择。Q-学习的参数如下: 学习速率 $\alpha = 0.5$, 折算因子 $\gamma = 0.9$, Q 值更新次数为 100, 计算过程中运行次数为 500, 存放 Q 值的大表 Qtable 是一个 $S \times A = 90 \times 6$ 的矩阵, 初始值为 0.1, Q-学习的目标是: 路网的性能函数达到最小值。最后, 本文选取了路网的三个性能指标(平均延误时间、平均行程时间和平均排队长度)对以上两种方案进行对比验证。受篇幅的限制, 这里只给出实验对比的结果, 见表 2。

由表 2 可以看出, 在方案二的情况下路网的三项指标均好于方案一。验证结果表明, 本文提

表 2 小区路网性能参数对比

Table 2 Comparison of performance parameters of the traffic network

	平均延误时间/s	平均行程时间/s	平均排队长度/m
方案一	34.9	140.5	129.0
方案二	29.9	137.0	125.6

出的在线模式选择方法对改善交通系统运行状态是有效的。

图4是两种方案下的模式选择结果对比。可以看出,在多数情况下,两个方案的选择结果是不同的,方案二的选择变化比较多,也就是说,方案二的选择方式更适应多变的交通环境,对交通状态的变化反应更敏感,因此带来了更好的协同效果。

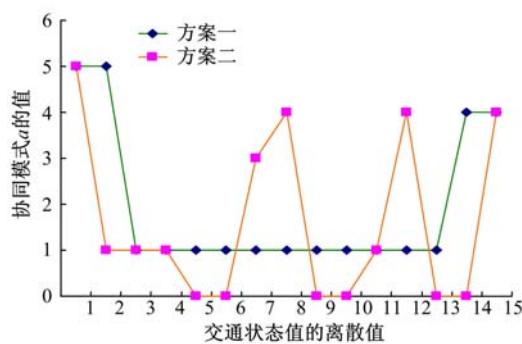


图4 两种方案下的协同模式选择结果对比

Fig. 4 Comparison results with two different collaboration mode

4 结束语

本文提出了一种基于Q-学习的交通控制与诱导协同模式在线选择方法,它克服了现有离线选择方法不能满足交通复杂多变特性、需要准确的数学模型等缺陷。本文的方法实现了协同模式的在线选择,仿真实验结果表明,该方法对改善交通状态是有效的,为交通控制与诱导协同运作的实施提供了一种新的方法和思路。

参考文献:

- [1] 徐立群.城市交通流诱导与控制一体化理论和模型研究[D].长春:吉林大学交通学院,2000.
- Xu Li-qun. Urban traffic flow guidance and control theory and model of integration [D]. Changchun: College of Transportation, Jilin University, 2000.
- [2] 李瑞敏,史其信.基于多智能体系统的城市交通控制与诱导集成化研究[J].公路交通科技,2004(5): 109-112.
- Li Rui-min, Shi Qi-xin. Research on integration of urban traffic control and route guidance based on multi-agent[J]. Journal of Highway and Transportation Research and Development, 2004 (5): 109-112.
- [3] Szepesvari C. The asymptotic convergencerate of Q-Learning[C] // Proceedings of Neural Information Processing Systems. Cambridge, MA: The MIT Press, 1997.
- [4] Werbos P J. A Menu of Designs for Reinforcement Learning Over Time[M]. Cambridge, MA: The MIT Press, 1990.
- [5] Celine Jacob. Optimal. Integrated and adaptive traffic corridor control: a machine learning approach [D]. Canada: Department of Civil Engineering University of Toronto, 2005.
- [6] 于江涛.多智能体模型、学习和协作研究与应用[D].杭州:浙江大学控制科学与工程学系,2003.
- Yu Jiang-tao. Research and application on multi-agent system modeling, learning and cooperation [D]. Hangzhou: Department of Control Science and Engineering, Zhejiang University, 2003.
- [7] 赵强.基于关键交叉口交通状态判别的配时参数计算[D].长春:吉林大学交通学院,2007.
- Zhao Qiang. The calculation of signal timing parameters based on the traffic state identification of critical intersections[D]. Changchun: College of Transportation, Jilin University, 2007.
- [8] 于德新,杨兆升,王媛,等.基于多智能体的城市道路交通控制系统及其协调优化[J].吉林大学学报:工学版,2006,36(1):113-118.
- Yu De-xin, Yang Zhao-sheng, Wang Yuan, et al. Urban road traffic control system and its coordinate optimization based on multi-agent system[J]. Journal of Jilin University(Engineering and Technology Edition), 2006,36(1):113-118.
- [9] 杨煜普,欧海涛.基于再励学习与遗传算法的交通信号自组织控制[J].自动化学报,2002,28(4):564-568.
- Yang Yu-pu, Ou Hai-tao. Self-organized control of traffic signals based on reinforcement learning and genetic algorithm[J]. Acta Automatica Sinica, 2002, 28(4):564-568.