مجله کنترل
ISSN (print) 2008-8345
ISSN (online) 2538-3752
جلد ۱۱، شماره ۲، تابستان ۱۳۹۶، صفحه ۲۱-۹

توسعه کنترلر هوشمند چراغ‌های راهنمایی بر پایه یادگیری تقویتی حالت پیوسته در محیط ترافیکی میکروسکوپیک

محمد اصلانی [۱]، محمد سعدی مسگری [۲]

[۱] گروه مهندسی GIS، دانشگاه صنعتی خواجه نصیرالدین طوسی، maslani@mail.kntu.ac.ir

[۲] دانشیار، گروه مهندسی GIS، قطب علمی مهندسی فناوری اطلاعات مکانی، دانشگاه صنعتی خواجه نصیرالدین طوسی، mesgari@kntu.ac.ir

(تاریخ دریافت مقاله ۱۳۹۶/۲/۳۰، تاریخ پذیرش مقاله ۱۳۹۶/۴/۱۶)

چکیده: افزایش روزافزون تعداد خودروها و در پی آن ترافیک‌های سنگین شهری چالش بزرگی را برای کنترل بهینه ترافیک شهری برای مهندسین ایجاد کرده است. روش مناسب برای کنترل بهینه ترافیک هرچه باشد یقیناً باید وفق پذیر بوده تا بتواند ترافیک شهری را که دارای طبیعت پویا، پیچیده و تغییرپذیر است را به‌خوبی مدیریت نماید. در این راستا تمرکز اصلی تحقیق حاضر کنترل هوشمند و توزیع یافته چراغ‌های راهنمایی بر پایه یادگیری تقویتی است. کنترل هوشمند چراغ‌های راهنمایی بر پایه یادگیری تقویتی نیاز به یادگیری و تصمیم‌گیری در فضای حالت بزرگ (پیوسته) را دارد. همین امر باعث می‌شود که روش‌های رایج یادگیری تقویتی (حالت گسسته) برای چنین مسائلی (با فضای حالت بزرگ) به‌خوبی قابل بسط نباشند. هدف تحقیق حاضر حل این چالش در مسئله کنترل ترافیک میکروسکوپیک است. در همین راستا نوآوری تحقیق حاضر را می‌توان توسعه کنترلر هوشمند چراغ‌های راهنمایی بر پایه یادگیری تقویتی حالت پیوسته برای حل چالش بزرگ بودن فضای حالت برشمرد. یادگیری تقویتی حالت پیوسته از شباهت سنجی حالات برای تخمین ارزش آن‌ها استفاده می‌کند. در این تحقیق به‌منظور اعتبار سنجی، دو روش یادگیری Q و عملگر-نقاد حالت گسسته نیز پیاده‌سازی و عملکرد آن‌ها با روش پیشنهادی مقایسه شدند. نتایج نشان می‌دهند که روش پیشنهادی منجر به کاهش ۱۶٪ و ۱۳٪ زمان سفر در مقایسه با دو روش عملگر-نقاد و یادگیری Q می‌شود.

کلمات کلیدی: یادگیری تقویتی پیوسته، یادگیری Q، عملگر-نقاد، ناحیه بندی فضا و کنترل میکروسکوپیک ترافیک.

Developing Adaptive Traffic Signal Controller based on Continuous Reinforcment Learning in a Microscopic Traffic Environment

Mohammad Aslani, Mohammad Saadi Mesgari

Abstract: The daily increase of a number of vehicles in big cities poses a serious challenge to efficient traffic control. The suitable approach for optimum traffic control should be adaptive in order to successfully content with the urban traffic that has the dynamic and complex nature. Within such a context, the major focus of this research is developing a method for adaptive and distributed traffic signal control based on reinforcement learning (RL). RL as a promising approach for generating, evaluating, and improving traffic signal decision-making solutions is beneficial and synergetic. RL-embedded traffic signal controller has the capability to learn through experience by dynamically interacting with the traffic environment in order to reach its goals. Traffic signal control often requires dealing with continuous state defined by means of continuous variables. Conventional RL methods do not scale well to problems with continuous state space or very large state space because they require storing distinct estimations of each state value in lookup tables. The contribution of the present research is developing adaptive traffic signal controllers based on continuous state RL for handling the big state space challenge arises in traffic control. The

performance of the proposed method is compared with Q-learning and actor-critic and the results reveal that the proposed method outperforms others.

Keywords: Continuous State Reinforcement Learning, Q-Learning, Actor-Critic, Microscopic Traffic Control.

۱- مقدمه

با افزایش روزافزون تقاضا برای حمل‌ونقل در مسیرهای درون‌شهری، ازدحام و ترافیک در شهرها تبدیل به یکی از پدیده‌های رایج روزمره شده است. افزایش زمان سفر و در نتیجه افزایش آلودگی ناشی از سوخت‌های فسیلی از پیامدهای عمده این افزایش تقاضا به شمار می‌روند [۱]. از آنجائی که راه‌حل‌های بلندمدت برای این مشکل مستلزم سرمایه‌گذاری و فرهنگ‌سازی گسترده است، استفاده از راه‌کارهای سریع برای کاهش این مشکل ضروری به نظر می‌رسند. در این میان کنترل بهینه چراغ‌های راهنمایی نقش مهمی را در مدیریت و کاهش ترافیک ایفا می‌نماید [۲]. ازآنجایی‌که در طول یک شبانه‌روز شاهد انواع ترافیک‌ها در یک تقاطع خاص هستیم، استفاده از یک زمان‌بندی ثابت و از پیش‌تعریف‌شده برای یک چراغ راهنمایی، نه‌تنها باعث کنترل ترافیکی مناسبی نمی‌شود، بلکه به‌نوعی عامل افزایش آن نیز خواهد بود. امروزه با پیشرفت فناوری در علوم کامپیوتر و مهندسی برق و کنترل، با نصب چند سنسور در تقاطع‌ها و به‌کارگیری کنترلرهای هوشمند، به‌راحتی می‌توان مدیریت چراغ‌های راهنمایی را به یک سیستم هوشمند سپرد و از به‌هدررفتن زمان افراد و سوخت خودروها در پشت چراغ‌قرمزها و هم‌چنین افزایش آلودگی هوا جلوگیری کرد. چراغ راهنمایی هوشمند، سیستمی است که با توجه به حجم خودروهای ورودی به یک تقاطع هم‌سطح، زمان فازهای مختلف چراغ راهنمایی را به‌صورت عادلانه مدیریت می‌کند. در سال‌های اخیر روش‌های یادگیری ماشین ازجمله منطق فازی [۳, ۴]، شبکه عصبی [۵, ۶] و یادگیری تقویتی [۷-۹] پتانسیل‌های بالایی را برای طراحی کنترلرهای هوشمند چراغ‌های راهنمایی از خود نشان داده‌اند.

در این تحقیق یادگیری تقویتی به دلیل توانایی برخط آن جهت بهبود تدریجی عملکردش، توانایی وفق پذیری‌اش با نوسانات ترافیکی و توانایش در کنترل بدون دانستن مدل محیط ترافیکی مورد استفاده قرار گرفت [۱۰, ۱۱]. به بیان ساده‌تر، کنترلر مبتنی بر یادگیری تقویتی از طریق تعامل هوشمند با محیط ترافیکی که دارای روندهای غیرخطی و اتفاقی[1] است تجربه‌اندوزی کرده و برای رسیدن به اهدافش اعمال لازمه را انتخاب می‌نماید. در این راستا، عبدلهای و کاتان در سال ۲۰۰۳ مزیت‌های به‌کارگیری یادگیری تقویتی، به‌طور خاص یادگیری Q [2]، را برای کنترل یک چراغ راهنمایی با دو فاز بررسی کردند. نتایج کار آن‌ها نشان داد که کنترلر یادگیری Q دارای عملکرد نسبتاً بهتری در مقایسه با کنترلر غیرهوشمند زمان ثابت است [۱۲]. ون و همکاران در سال ۲۰۰۷، کنترلر هوشمندی را بر اساس روش سارسا[3] برای کنترل یک تقاطع ایزوله توسعه دادند. عملکرد کنترلر پیشنهادی با کنترلرهای زمان ثابت و القایی [۱۳] مقایسه شد و نتایج نشان دادند که کنترلر پیشنهادی برای کنترل ترافیک، خصوصاً حالت‌های اشباع ترافیک، دارای عملکرد به‌مراتب بهتری است [۱۴]. در [۱۵]، روش یادگیری Q برای کنترل یک تقاطع ایزوله در شهر تورنتوی کانادا توسعه داده شد. در این تحقیق محققین تأثیر تعاریف مختلف برای حالات محیط را بر روی عملکرد کنترلر بررسی کردند. آن‌ها نشان دادند که کنترلر مبتنی بر یادگیری Q صرف‌نظر از نوع تعریف حالت محیط بهتر از کنترلر زمان ثابت عمل می‌نماید. در [۷]، محققین کار قبلی خود را توسعه دادند و عملکرد دو الگوریتم یادگیری Q و سارسا را در کنترل چراغ‌های راهنمایی یک شبکه ترافیکی در شهر تورنتو بررسی کردند. آن‌ها نشان دادند که روش پیشنهادی به‌راحتی برای تعداد زیادی از چراغ‌های راهنمایی قابل توسعه است. همچنین سه تعریف مختلف از حالت محیط و چهار تعریف از تابع پاداش ارائه و مقایسه شدند. نتایج نشان دادند که تعداد خودروهای ورودی به تقاطع و طول صف به‌عنوان بهترین تعریف حالت محیط و مجموع منفی زمان تأخیر خودروهای منتظر در تقاطع به‌عنوان بهترین تابع پاداش می‌باشند. آن‌ها نشان دادند که یادگیری Q دارای عملکرد بهتری در مقایسه با سارسا است. همچنین در تحقیق ایشان تأثیر عملکرد مناسب کنترلر چراغ راهنمایی بر کاهش آلاینده CO تولیدشده بررسی شد و نتایج نشان دادند که کنترلر هوشمند مبتنی بر یادگیری Q منجر به کاهش ۲۸٪ آلاینده CO در مقایسه با کنترلر غیرهوشمند می‌شود. در تحقیق [۱۶]، از یادگیری Q برای کنترل یک شریان ترافیکی متشکل از ۵ تقاطع استفاده شد. در این تحقیق هر چراغ راهنمایی علاوه بر مشاهده تعداد خودروهای منتظر در تقاطع تحت کنترل خود تعداد خودروهای منتظر در تقاطع‌های مجاور را نیز در تصمیم‌گیری لحاظ می‌نماید. نتایج کار ایشان نشان داد که در وضعیت اشباع ترافیکی، کنترلر مبتنی بر یادگیری تقویتی منجر به تعداد توقف‌های کمتری در مقایسه با کنترلر زمان ثابت شده است. سرعت‌پایین یادگیری از نقاط ضعف این تحقیق بشمار می‌رود. در [۱۷] محققین به توسعه یادگیری Q در سیستم‌های چند عامله هولونی برای کنترل چراغ‌های راهنمایی در یک شبکه ترافیکی متشکل از ۵۰ تقاطع پرداختند. سازمان هولونی به‌صورت مجموعه‌ای از سلسله‌مراتب در هم آمیخته‌ای از عامل‌ها دیده می‌شود. ازجمله ویژگی‌های این ساختار می‌توان به خود متشابهی و پویایی آن اشاره نمود. در مرحله اول تقاطع‌ها

[1] Stochastic
[2] Q-learning
[3] SARSA

از طریق یک الگوریتم گراف مبنا به تعدادی هولون طبقه‌بندی شدند که عامل‌ها در داخل هر هولون با یکدیگر به تبادل اطلاعات می‌پردازند و هر هولون توسط یک عامل سطح بالاتر کنترل می‌شود. نتایج این تحقیق نشان داد که روش پیشنهادی دارای عملکرد بهتری در مقایسه با کنترلر زمان ثابت است.

در [۱۸, ۱۹] چارچوب یادگیری تقویتی مدل مبنا برای کنترل چندین چراغ راهنمایی پیشنهاد شد. چارچوب پیشنهادی برخلاف تحقیقات فوق خودرو مبنا است بدین معنی که چراغ‌های راهنمایی بر اساس اطلاعات دریافتی از خودروها تصمیم‌گیری می‌نمایند. به عبارت بهتر هر خودرو زمان توقف خود را تخمین زده و به نزدیک‌ترین چراغ راهنمایی منتقل می‌نماید. هدف سیستم کنترل ترافیک مینیمم نمودن زمان توقف تمام خودروها در کل شبکه است. تابع ارزش[1] که زمان توقف مورد انتظار خودروها را تخمین می‌زند توسط چراغ‌های راهنمایی و خودروها تخمین زده می‌شود. نتایج این تحقیق نشان دادند که سیستم پیشنهادی زمان توقف را ۲۲٪ در مقایسه با کنترلر زمان ثابت کاهش می‌دهد. خمیس و گوما در سال ۲۰۱۴ روش پیشنهادی در [۱۸, ۱۹] را با استفاده از تئوری بایزین توسعه دادند. در این تحقیق تابع هدف تلفیق خطی چندین شاخص ازجمله متوسط زمان توقف در طول سفر، متوسط زمان توقف در هر تقاطع، سرعت و امنیت است. روش پیشنهادی در این تحقیق با روش *TC-1* در [۱۸] مقایسه و نتایج نشان دادند که روش پیشنهادی دارای عملکرد به‌مراتب بهتری است. نقطه‌ضعف اساسی در تحقیق ایشان غیرعملی بودن تشخیص متوسط زمان سفر، متوسط زمان توقف برای هر خودرو است. زیرا چنین تشخیصی در وهله اول نیاز به تشخیص هر خودرو به‌صورت مجزا در شبکه دارد. تشخیص هر خودرو و ردیابی آن در شبکه‌های ترافیکی بزرگ در عمل بسیار دشوار و نیازمند زیرساخت‌های بسیار قوی است [۲۰].

اگرچه روش‌های رایج یادگیری تقویتی (اشاره‌شده در تحقیقات فوق) دارای حجم محاسباتی پایینی هستند اما نیاز به تکرارهای بالایی برای یادگیری سیاست بهینه دارند. این روش‌ها برای مسائل کنترل ترافیک با فضای حالت بسیار بزرگ و پیوسته به‌خوبی قابل بسط نمی‌باشند زیرا آن‌ها برای هر زوج حالت-عمل یک مقدار مجزا را تخمین زده و ذخیره می‌کنند و این در حالی است که در اغلب مسائل کنترل ترافیک فضای حالت بسیار بزرگ بوده که این امر باعث عدم کارایی‌شان می‌شود. هدف تحقیق حاضر حل این چالش در مسئله کنترل ترافیک است. در همین راستا نوآوری تحقیق را می‌توان توسعه کنترلر هوشمند چراغ‌های راهنمایی بر پایه یادگیری تقویتی حالت پیوسته برای حل چالش بزرگ بودن فضای حالت برشمرد.

در این تحقیق محیط ترافیکی به‌صورت میکروسکوپیک شبیه‌سازی شده است. در مدل‌سازی میکروسکوپیک ترافیک، رفتار هر یک از رانندگان (خودروها) به‌تنهایی مدل می‌شود و در نتیجه عکس‌العمل و جزئیات رفتاری خودروها و تعامل آن‌ها با یکدیگر در مدل‌سازی لحاظ می‌شود. لازم به ذکر است که عامل‌های خودرو از نوع واکنشی بوده و فاقد توانایی یادگیری می‌باشند. در این مقاله در بخش ۲ به بررسی یادگیری تقویتی حالت پیوسته پرداخته می‌شود. بخش ۳ به پیاده‌سازی اختصاص داده شده است که در آن شبیه‌سازی میکروسکوپیک ترافیک و طراحی کنترلر مبتنی بر یادگیری تقویتی پیوسته تشریح می‌شوند. در بخش ۴ نتایج حاصل از پیاده‌سازی ارائه می‌شوند، در بخش ۵ به اعتبار سنجی نتایج پرداخته می‌شود و نهایتاً بخش ۶ به نتیجه‌گیری اختصاص داده شده است.

۲- یادگیری تقویتی حالت پیوسته

طیف گسترده‌ای از مسائل کنترل ترافیک ازجمله کنترل هوشمند چراغ‌های راهنمایی، کنترل نرخ جریان ورودی به بزرگراه‌ها و اعمال محدودیت سرعت برای خودروها نیاز به تصمیم‌گیری در فضای حالت بزرگ (پیوسته) را دارند [۲۱]. در این مسائل به‌کارگیری کنترلرهای هوشمند برپایه روش‌های یادگیری تقویتی رایج (حالت گسسته) می‌تواند منجر به افزایش تعداد تکرارهای موردنیاز برای یادگیری شود [۲۲]. به بیان ساده‌تر، غالباً روش‌های یادگیری تقویتی حالت گسسته دارای مشکل سرعت همگرایی آهسته هستند زیرا هرچه تعداد حالات افزایش یابد ابعاد جدول ذخیره‌سازی ارزش حالات (جدول Q) به‌صورت نمایی افزایش می‌یابد. همچنین جهت کاهش ابعاد جدول Q، اگر گسسته سازی متغیرهای فضای حالت به‌صورت درشت‌تری انجام پذیرد یادگیری و درنهایت تصمیم‌گیری با دقت کافی انجام نخواهد پذیرفت. راه‌حل پیشنهادی برای این چالش استفاده از یادگیری تقویتی پیوسته است که از شباهت سنجی حالات برای تخمین ارزش آن‌ها استفاده می‌کند [۱۰]. به‌بیان‌دیگر هرچه دو حالت از محیط به یکدیگر شبیه‌تر باشند ارزش آن‌ها نیز به یکدیگر شبیه‌تر خواهد بود [۲۳]. بر این اساس کنترلر یادگیر تقویتی دیگر نیازی به تجربه‌اندوزی مستقیم در تمام حالات محیط را برای تخمین ارزش آن‌ها ندارد که این موضوع می‌تواند سرعت همگرایی را تا حد زیادی افزایش دهد. در یادگیری تقویتی پیوسته، تابع تقریب زن با جدول ذخیره‌سازی ارزش Q جایگزین می‌شود. تابع تقریب زن توانایی به‌کارگیری دانش به‌دست‌آمده را به حالت‌های دیده نشده دارا است. تابع تقریب زن ابتدا توسط تابع ویژگی[2] حالت اولیه را به فضای ویژگی تصویر کرده و سپس ارزش حالات را در فضای ویژگی جدید توسط رابطه ۱ تخمین می‌زند [۱۰].

$$V(s) = \sum_{i=1}^{N} \varphi_i(s).\theta(i) \qquad (۱)$$

در رابطه ۱، θ پارامتر تنظیمی است که نشان‌دهنده بردار وزن برای ارزش حالات و $\varphi(s)$ تابع ویژگی است. یکی از توابع ویژگی معروف ناحیه بندی فضا است که حجم محاسباتی پایین و قدرت نمایش نسبتاً

[1] Value function

[2] Feature function

مناسب ازجمله ویژگی‌های مثبت این روش به‌حساب می‌آیند [۲۴]. در این روش فضای حالت توسط مجموعه‌ای از شبکه‌ها به ناحیه‌های کوچک‌تری تقسیم می‌شود که به هر شبکه یک *Tiling* و به هر ناحیه داخل شبکه یک *Tile* گفته می‌شود. در صورت به‌کارگیری یک‌لایه *Tiling*، در هر حالت محیط فقط ارزش یک *Tile* به‌روز می‌شود اما در صورت استفاده از چندین لایه *Tiling* در هر حالت محیط ارزش چندین *Tile* به‌روز می‌شود که این امر باعث افزایش سرعت همگرایی می‌شود. در روش تقسیم‌بندی فضا مقدار عضویت هر حالت از محیط به *Tile* های مختلف توسط تابع زیر تعیین می‌شود (رابطه ۲). در این رابطه، s حالت محیط و $\varphi_i(s)$ مقدار عضویت حالت محیط s به $tile_i$ است.

$$\varphi_i(s) = \begin{cases} 0 : if\ s \notin tile_i \\ 1 : if\ s \in tile_i \end{cases} \quad (۲)$$

در طول یادگیری مقدار بهینه بردار θ باید به‌گونه‌ای تعیین شود که تابع هزینه زیر مینیمم شود (رابطه ۳).

$$C = E[(r + \gamma \sum_{i=1}^{N} \varphi_i(s').\theta(i) - \sum_{i=1}^{N} \varphi_i(s).\theta(i))^2] \quad (۳)$$

در رابطه ۳، C تابع هزینه، E امید ریاضی، r پاداش لحظه‌ای، γ نرخ تخفیف و s' نیز حالت بعدی محیط از دید عامل است. با اعمال روش گرادیان نزولی بر روی تابع هزینه به‌روزرسانی بردار θ در طول یادگیری بر اساس رابطه ۴ صورت می‌پذیرد [۱۰].

$$w_{t+1} = \gamma\lambda w_t + \varphi(s)$$
$$\theta_{t+1} = \theta_t - \alpha\left(r_t + \gamma\theta_t^T.\varphi(s') - \theta_t^T.\varphi(s)\right)w_{t+1} \quad (۴)$$

در رابطه ۴، λ میزان تأثیرپذیری ارزش حالات ابتدایی اپیزود از ارزش حالات و سیگنال‌های انتهایی اپیزود است و α نیز نرخ یادگیری است. در این تحقیق ارزش حالات–اعمال بر پایه یادگیری Q حالت پیوسته به‌روزرسانی می‌شوند و خوانندگان برای جزئیات بیشتر دراین‌باره می‌توانند به منابع [۱۰, ۲۲] مراجعه نمایند.

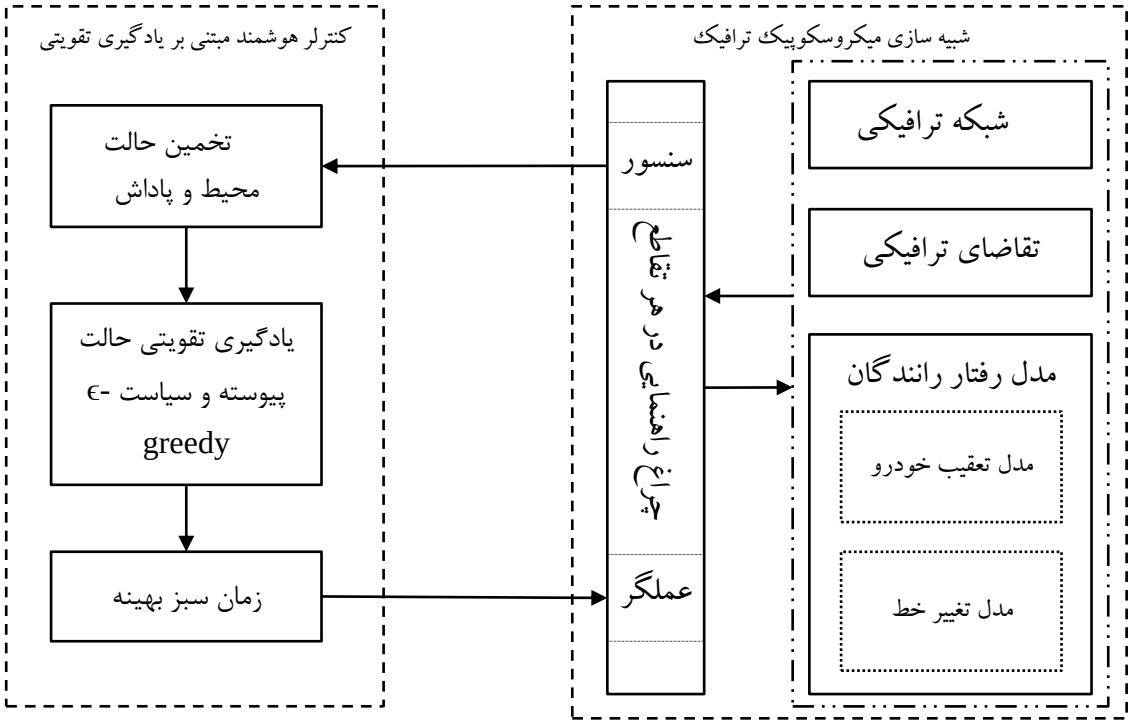
۳- پیاده سازی

توسعه شهرنشینی و پیشرفت فن‌آوری منجر به افزایش روزافزون تعداد سفرهای درون‌شهری و خودروها و در نتیجه آن ترافیک شهری شده است. یکی از روش‌های کاهش معضل ترافیک احداث راه‌های جدید است. اما این راه‌حل خود مستلزم هزینه‌های بالای اجرایی است. راه‌حل دیگر که به‌صرفه‌تر و سریع‌تر نیز است توسعه سیستم‌های حمل‌ونقل هوشمند و به‌طور خاص چراغ‌های راهنمای هوشمند است. در چراغ‌های راهنمایی هوشمند پارامترهای زمان‌بندی چراغ‌های هوشمند با توجه به شرایط ترافیکی موجود (به‌عنوان‌مثال، تعداد ماشین‌های منتظر در تقاطع) برای رسیدن به یک سری اهداف خاص (به‌عنوان‌مثال، بیشترین تعداد ماشین‌های عبوری از تقاطع) بهینه می‌شوند. در این تحقیق به توسعه کنترلرهای هوشمند چراغ‌های راهنمایی برپایه یادگیری تقویتی در محیط ترافیکی میکروسکوپیک پرداخته‌شده است. در راستای رسیدن به این هدف، در این بخش ابتدا مختصراً شبیه‌سازی میکروسکوپیک انجام‌شده، که به‌نوعی نقش محیط را برای چراغ‌های راهنمایی بازی می‌کند، تشریح می‌شود و سپس به توسعه کنترلر وفق پذیر چراغ راهنمایی بر پایه یادگیری تقویتی پیوسته پرداخته می‌شود (شکل ۱). لازم به ذکر است که تمامی پیاده‌سازی‌ها در این تحقیق در زبان $C++$ انجام‌شده‌اند.

۳-۱- شبیه‌سازی ترافیک

در این تحقیق یک شبیه‌سازی میکروسکوپیک ترافیک به‌منظور بررسی عملکرد کنترلرهای چراغ راهنمایی توسعه داده شد. در این راستا می‌توان به تحقیقات انجام‌شده توسط ویرینگ در سال ۲۰۰۰ اشاره نمود که به‌منظور توسعه کنترلرهای هوشمند مبتنی بر یادگیری تقویتی ابتدا به توسعه شبیه‌سازی میکروسکوپیک پرداخت. در تحقیق ایشان برخلاف تحقیق حاضر خودروها فاقد رفتار تغییر خط، افزایش و کاهش شتاب هستند و فقط همانند مستطیل‌هایی در یک راستا و به‌صورت گسسته حرکت می‌کنند [۱۸]. در سال ۲۰۱۴ خمیس و گوما شبیه‌سازی میکروسکوپیک انجام‌شده توسط ویرینگ را به‌گونه‌ای توسعه دادند که خودروها توانایی افزایش و کاهش شتاب را به‌صورت پیوسته داشته باشند. اما همچنان در شبیه‌سازی ایشان همچنان خودروها فاقد توانایی سبقت گرفتن و تغییر خط خود هستند [۲۰]. در سال ۲۰۱۴، ال تانتاوی و همکاران برای ارزیابی کنترلرهای پیشنهادی خود به انجام شبیه‌سازی ترافیکی در بخشی از شهر تورنتو با استفاده از نرم‌افزار *Paramics* پرداختند. در شبیه‌سازی ترافیکی انجام‌شده همانند مقاله حاضر خودروها توانایی افزایش شتاب، کاهش شتاب، تغییر خط و سبقت گرفتن را دارا هستند [۷].

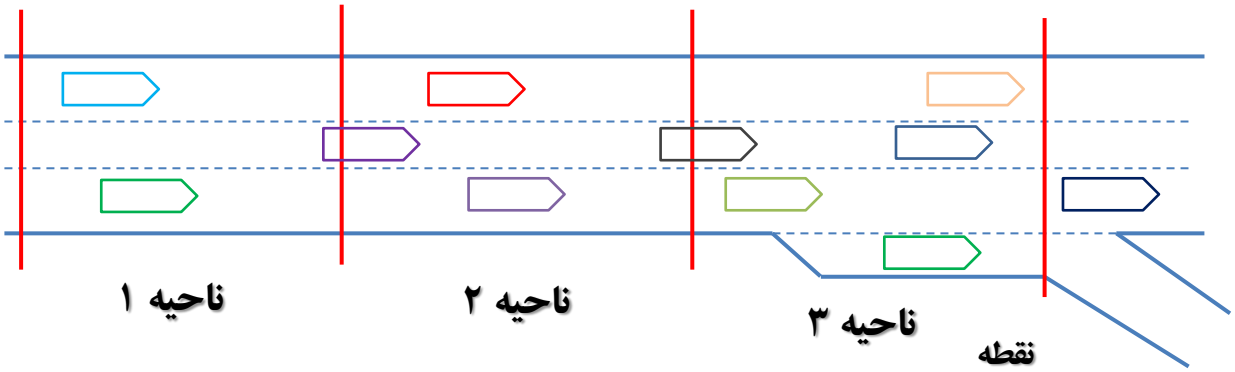
هر شبیه‌سازی میکروسکوپیک ترافیک از چهار بخش اصلی شبکه ترافیکی، تقاضای ترافیکی، موجودیت‌های متحرک (نظیر خودرو، موتورسیکلت و اتوبوس) و موجودیت‌های کنترل ترافیک (نظیر تابلوهای محدودیت سرعت و چراغ‌های راهنمایی) تشکیل‌شده است (شکل ۱). شبکه ترافیکی نشان‌دهنده هندسه و توپولوژی خیابان‌ها و تقاطع‌ها است. به عبارت بهتر شبکه ترافیکی نشان‌دهنده ابعاد هر خیابان (طول و عرض) و تعداد خطوط آن است. در این تحقیق کنترل ترافیک برای یک شبکه ترافیکی ۳×۳ همگن (متقارن) متشکل از ۹ تقاطع و ۲۴ خیابان انجام می‌پذیرد. تمامی خیابان‌ها دوطرفه و دارای دو خط با ظرفیت ۴۰ خودرو



شکل ۱. اجزاء مختلف در پیاده‌سازی

```
If (it is necessary to change lanes) then
      Apply Lane-Changing Model
endif
If (the vehicle has not changed lanes) then
      Apply Car-Following Model
Endif
```

شکل ۲. الگوریتم به‌کارگیری دو مدل تغییر خط و تعقیب خودرو



شکل ۳. ناحیه‌های ترافیکی در هر خیابان

در هر طرف می‌باشند. طول هر خیابان نیز ۲۵۰ متر با حداکثر سرعت مجاز ۵۰ کیلومتر بر ساعت است. همچنین فرض می‌شود که ۳۳٪ از خودروها مسیر مستقیم، ۳۳٪ گردش‌به‌چپ و ۳۳٪ گردش‌به‌راست را در صورت مواجه با تقاطع انجام می‌دهند. همواره خودروها در طول مسیر حرکت خود تمایل دارند تا به‌سرعت مطلوب برسند اما محیط پیرامونشان (خودروی جلویی، خودروهای مجاور، چراغ راهنمایی و تابلوهای محدودیت سرعت) مانع از رسیدن آن‌ها به‌سرعت مطلوب می‌شود. در هر گام شبیه‌سازی موقعیت و سرعت هر خودرو توسط دو مدل تعقیب خودرو و تغییر خط به‌روز می‌شوند (شکل ۲) [۲۵, ۲۶]. مدل تغییر خط یک‌روند تصمیم‌گیری است که دو مسئله امکان تغییر خط و نیاز به تغییر خط (رفتن به خط گردش و رسیدن به‌سرعت رانندگی بهتر) را در نظر می‌گیرد.

```
Z←Current zone of the vehicle
TL← Turning lane
V←Current speed of the vehicle
V_D← Desired speed of the driver
L← Current driving lane
If (Z==1)
    V′←calculate the future speed of the vehicle in the desired driving lane
    If (|V′-VD|<|V-VD|)
        If (there is enough gap in the desired driving lane)
            Change the driving lane
        Endif
    Endif
Endif
If (Z==2)
    If (TL <> L)
        If (there is enough gap in the desired driving lane)
            Change the driving lane
        Endif
    Endif
Endif
If (Z==3)
    If (TL <> L)
        If (there is enough gap in the desired driving lane)
            Change the driving lane
        Elseif
            Decrease the speed by using normal deceleration
        Endif
    Endif
Endif
```

شکل ۴. الگوریتم تغییر خط

مدل تغییر خط یک‌روند تصمیم‌گیری است که دو مسئله امکان تغییر خط و نیاز به تغییر خط (رفتن به خط گردش و رسیدن به‌سرعت رانندگی بهتر) را در نظر می‌گیرد. برای دستیابی به یک نمایش دقیق‌تری از رفتار رانندگان در روند تصمیم‌گیری تغییر خط، هر خیابان به سه ناحیه تقسیم می‌شود (شکل ۳) که در هر ناحیه انگیزه راننده جهت تغییر خط متفاوت است (شکل ۴). در ناحیه ۱، انگیزه تغییر خط، رسیدن به شرایط ترافیکی بهتر (سرعت رانندگی نزدیک‌تر به‌سرعت مطلوب) است. در ناحیه ۲، انگیزه تغییر خط نزدیک شدن به خط گردش[1] است و در ناحیه ۳ نیز انگیزه نزدیک شدن به خط گردش است با این تفاوت که در ناحیه ۳ در صورت نبودن فضا در خط مقصد راننده باید سرعت خود را جهت پیدا نمودن فضای مناسب در خط مقصد کاهش دهد (شکل ۴).

در این تحقیق مدل تعقیب خودروی ارائه‌شده توسط گیپس مورد استفاده قرار گرفت [۲۷, ۲۸]. این مدل از دو جزء کاهش شتاب[2] و افزایش شتاب[3] تشکیل‌شده است. جزء اول دربرگیرنده محدودیت‌های سرعت اعمال‌شده توسط خودروی جلویی و جزء دوم دربرگیرنده قصد خودرو برای رسیدن به‌سرعت مطلوبش است. ایده مدل‌سازی کاهش شتاب بر اساس اجتناب از تصادف است. بدین معنی که راننده همواره سعی می‌کند که فاصله‌اش را از خودروی جلوی در حدی نگه دارد که در صورت توقف ناگهانی از جانب خودروی جلویی بتواند بدون هیچ‌گونه برخوردی با خودروی جلویی به توقف کامل برسد. رابطه ۵ مدل کاهش شتاب بکار رفته در این تحقیق را نشان می‌دهد [۲۶]. در این رابطه n شماره خودروی جلوی، $n+1$ شماره خودروی عقبی، t زمان فعلی شبیه‌سازی، V_{n+1} سرعت خودروی $n+1$ ام، V_n سرعت خودروی n، T گام زمانی و همچنین زمان واکنش راننده، b_{n+1} شتاب کاهشی خودروی عقبی، $x_n(t)$ محل خودروی n در زمان t، $x_{n+1}(t)$ محل خودروی $n+1$ ام در زمان t و L_n کمترین سرفاصله[4] است. لازم به ذکر است که زمان واکنش راننده عبارت است از مدت‌زمانی که طول می‌کشد تا راننده به تغییرات سرعت خودروی جلویی واکنش نشان دهد. در این تحقیق میزان این زمان برابر ۱ ثانیه در نظر گرفته شد.

$$V_{n+1}^{D}(t+T) \le T\, b_{n+1} + (T^2\, b_{n+1}^2(t) - b_{n+1}[2(x_n(t) - x_{n+1}(t) - L_n)] - T\, V_{n+1}(t) - \frac{V_n^2(t)}{b_n})^{0.5} \quad (۵)$$

رابطه ۶ نیز برای افزایش شتاب مورد استفاده قرار گرفت [۲۶]. در این رابطه a_{n+1} حداکثر شتاب افزایشی خودروی $n+1$ و $V_{n+1}^{*}(t)$ سرعت مطلوب خودروی $n+1$ در زمان t هستند.

[1] Turning lane
[2] Deceleration
[3] Acceleration
[4] Minimum Headway

$$V_{n+1}^{A}(t+T) = V_{n+1}(t) + 2.5\, a_{n+1} T\left(1 - \frac{V_{n+1}(t)}{V_{n+1}^{*}(t)}\right)\sqrt{0.025 + \frac{V_{n+1}(t)}{V_{n+1}^{*}(t)}} \quad (۶)$$

سرعت نهایی خودرو در زمان $t+T$ مینیمم سرعت خودرو در دو حالت افزایش شتاب و کاهش شتاب است (رابطه ۷).

$$V_{n+1}^{Final}(t+T) = Min\{V_{n+1}^{A}(t+T), V_{n+1}^{D}(t+T)\} \quad (۷)$$

لازم به ذکر است که رفتار و ویژگی خودروها توسط پارامترهای حداکثر سرعت، حداکثر شتاب افزایشی و حداکثر شتاب کاهشی قابل توصیف هستند. در این تحقیق حداکثر سرعت، حداکثر شتاب افزایشی و حداکثر شتاب کاهشی هر خودرو به ترتیب از توابع توزیع گوسین با میانگین‌های ۱۱۰ کیلومتر بر ساعت، ۳ و ۶ متر بر مجذور ثانیه و انحراف از معیارهای ۱۰ کیلومتر بر ساعت، ۰.۲ متر بر مجذور ثانیه و ۰.۵ متر بر مجذور ثانیه انتخاب می‌شوند.

سرعت مطلوب یک خودرو بر اساس سه فاکتور حداکثر سرعت خودرو، حداکثر سرعت مجاز خیابان‌ها و میزان تبعیت از حداکثر سرعت مجاز خیابان‌ها تعیین می‌شود. به‌عنوان‌مثال فرض نمایید که حداکثر سرعت خودرو ۱۵۰ کیلومتر بر ساعت، حداکثر سرعت مجاز خیابان نیز ۶۰ کیلومتر بر ساعت و میزان تبعیت از حداکثر سرعت مجاز ۱.۲ باشد. سرعت مطلوب خودرو توسط رابطه ۸، ۷۲ کیلومتر بر ساعت محاسبه می‌شود. در این تحقیق میزان تبعیت از حداکثر سرعت مجاز خیابان‌ها به‌تصادف برای هر خودرو از یک تابع توزیع گوسین با میانگین ۱.۱ و انحراف از معیار ۰.۱ انتخاب می‌شود.

$$V^{*} = min(150, 1.2 \times 60) = 72\ (km/h) \quad (۸)$$

۳-۲- کنترلر هوشمند چراغ راهنمایی بر پایه یادگیری تقویتی پیوسته

در شروع هر فاز، چراغ راهنمایی وضعیت ترافیکی محلی (حالت محیط) را که توسط بردار $[Ph, A_1, A_2, \ldots, A_w]$ نمایش داده می‌شود را مشاهده می‌کند. در این بردار، Ph شماره فاز جاری چراغ راهنمایی، w تعداد خیابان‌های ورودی به تقاطع و $A_1, A_2, \ldots A_w$ تعداد خودروهای منتظر در خیابان‌های ورودی به تقاطع هستند. برای مثال فضای حالت برای یک تقاطع که دارای ۴ خیابان ورودی است ۵ بعدی می‌باشد. بعد از مشاهده وضعیت محیط، چراغ راهنمایی یک مدت‌زمان سبز را از میان مقادیر [۹۰ و ۸۰ و ۷۰ و ۶۰ و ۵۰ و ۴۰ و ۳۰ و ۲۰] ثانیه انتخاب می‌کند. بعد از انتخاب زمان سبز، چراغ راهنمایی تا انتهای فاز که مجموع زمان سبز و زرد (زمان زرد ۵ ثانیه) است صبر کرده و سپس سیگنال پاداش را که نشان‌دهنده میزان مطلوبیت عمل انتخاب‌شده از محیط است را دریافت می‌نماید. این سیگنال پاداش توسط رابطه ۹ تعریف می‌شود.

$$Reward = \sum_{i=1}^{w} A_{i,k} - \sum_{i=1}^{w} A_{i,k+1} \quad (۹)$$

در رابطه ۹، w تعداد خیابان‌های ورودی به تقاطع، $A_{i,k}$ تعداد خودروهای منتظر در تقاطع در گام زمانی k و $A_{i,k+1}$ تعداد خودروهای منتظر در تقاطع در گام زمانی $k+1$ است. پاداش دریافتی برای به‌روزرسانی ارزش اعمال که نشان‌دهنده کیفیت آن‌ها در راستای برآورده نمودن اهداف مسئله (کاهش زمان سفر) است مورداستفاده قرار می‌گیرد. ارزش اعمال در ابتدا صفر است که نشان‌دهنده عدم دانش چراغ‌های راهنمایی از محیط ترافیکی است. این ارزش‌ها در طول زمان به‌روزرسانی می‌شوند. از آنجائی که چراغ‌های راهنمایی در ابتدا هیچ‌گونه دانشی از محیط اطراف خود ندارند نیاز به کنکاش اعمال مختلف در حالت‌های مختلف محیط را دارند، بنابراین چراغ‌های راهنمایی با انتخاب اعمال مختلف و دریافت سیگنال‌های پاداش به تعامل با محیط اطراف خود می‌پردازند. ارزش اعمال-حالات محیط توسط روابط ۱ و ۴ به‌روزرسانی می‌شوند.

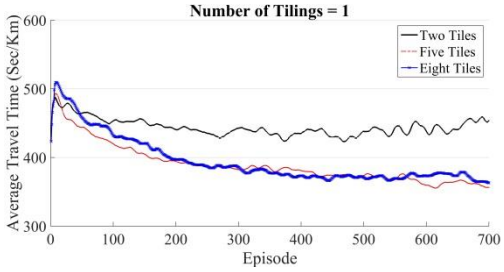
در یادگیری تقویتی پیوسته از تقریب زن‌های خطی بر پایه *Tile coding* استفاده شد. عملکرد روش *Tile coding* وابستگی زیادی به تعداد *Tile* ها و *Tiling* ها دارد و عدم انتخاب مناسب آن‌ها تأثیر بسزایی در کاهش عملکرد کنترلر هوشمند دارد. در این تحقیق مقادیر مختلف *Tile* ها و *Tiling* ها موردبررسی قرار می‌گیرند و نتایج به ازای مقادیر مختلف ارائه و بررسی می‌شوند (به بخش ۴ مراجعه شود). شبیه‌سازی میکروسکوپیک برای ۷۰۰ ساعت انجام و هر ساعت یک اپیزود در نظر گرفته می‌شود. هرچه تعداد اپیزودها افزایش یابد چراغ‌های راهنمایی کمتر به کنکاش محیط پرداخته و بیشتر به دانش به‌دست‌آمده خود اکتفا کرده و اعمالی را انتخاب می‌کنند که دارای ارزش بیشتری هستند. در حقیقت چراغ‌های راهنمایی یادگیر نیاز به برقراری تعادل میان اکتشاف و بهره‌برداری دارند. در این آزمایش از روش ε-*greedy* برای این منظور استفاده شد [۱۰]. پارامتری که عملکرد یادگیری تقویتی را تحت تأثیر قرار می‌دهد، نرخ یادگیری است. درصورتی‌که نرخ یادگیری کوچک باشد سرعت یادگیری کُند و درصورتی‌که بزرگ باشد یادگیری منجر به واگرایی سیستم می‌شود. نرخ یادگیری در این تحقیق برابر ۰.۱ انتخاب شد. همچنین برای مقایسه عادلانه روش‌های مختلف از سه شاخص ارزیابی متوسط زمان سفر[1] هر خودرو *(Sec/km)*، متوسط زمان توقف[2] هر خودرو *(Sec/km)* و متوسط تعداد توقف‌ها *(#/veh/km)* استفاده شد. متوسط زمان سفر هر خودرو عبارت است از میانگین زمانی که یک خودرو نیاز دارد تا یک کیلومتر را در شبکه طی نماید و بدین‌صورت محاسبه می‌شود که ابتدا برای هر خودرو زمان سفر تعیین‌شده و بر کل تعداد خودروها میانگین‌گیری می‌شود. متوسط زمان توقف برای هر خودرو عبارت است از مجموع زمان توقف هر خودرو در هر کیلومتر. متوسط تعداد توقف‌ها عبارت است از متوسط تعداد

[1] Travel Time
[2] Stop Time

توقف‌هایی که هر خودرو در هر کیلومتر دارد. شایان‌ذکر است که ماکزیمم کردن پاداش عامل‌ها با مینیمم کردن معیارهای ارزیابی همبستگی دارد.

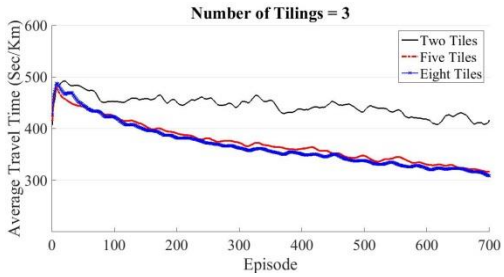
۴- نتایج

تعداد *Tiling* ها و *Tile* ها در عملکرد کنترلر یادگیر بسیار تأثیرگذار هستند بنابراین در این بخش عملکرد کنترلرهای یادگیر مختلف به ازای مقادیر مختلف *Tiling* ها و *Tile* ها ارائه و مقایسه می‌شوند تا درنهایت مقادیر بهینه آن‌ها تعیین شوند. در شکل ۵، متوسط زمان سفر به ازای یک *Tiling* و مقادیر مختلف *Tile* ها (۲، ۵ و ۸) نشان داده شده است. در این شکل نمودارهای یادگیری از میانگین‌گیری ۵ بار اجرای الگوریتم و میانگین‌گیری در بازه‌های ۹ ساعته به‌دست‌آمده‌اند. در تمام این نمودارها محور افقی نشان‌دهنده اپیزود و محور قائم نشان‌دهنده شاخص ارزیابی است.



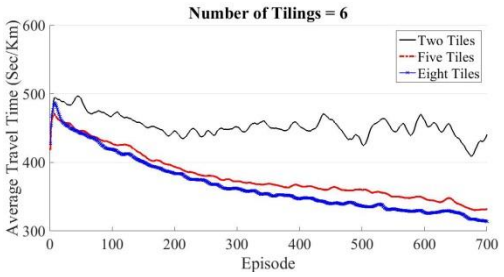
شکل ۵. نمودار یادگیری کنترلرهای یادگیر تقویتی پیوسته به ازای یک *Tiling* و تعداد *Tile* های ۲، ۵ و ۸

همان‌طور که مشخص است کنترلر یادگیر به ازای دو *Tile* در تمام شاخص‌های ارزیابی فاقد رفتار یادگیری است و نمودارِ نشان‌دهنده عملکرد آن دارای نوسانات زیادی است. همچنین کنترلرهای یادگیر به ازای پنج و هشت *Tile* دارای عملکردهای نزدیک به هم هستند. همچنین اختلاف محسوس میان نمودارهای یادگیری میان دو *Tile* و پنج و هشت *Tile* نشان از تأثیر زیاد تعداد *Tile* ها بر روی عملکرد یادگیری دارد.

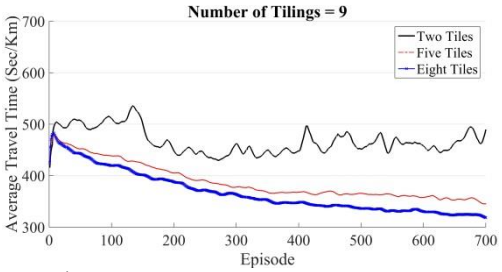


شکل ۶. نمودار یادگیری کنترلرهای یادگیر تقویتی پیوسته به ازای سه *Tiling* و تعداد *Tile* های ۲، ۵ و ۸

در شکل ۶، عملکرد کنترلرهای یادگیر به ازای سه *Tiling* و مقادیر مختلف *Tile* ها (۲، ۵ و ۸) نشان داده شده است. افزایش تعداد *Tile* ها از پنج به هشت تأثیر قابل‌توجه ای در یادگیری نداشته است. در شکل ۷، نمودارهای یادگیری کنترلرهای وفق پذیر به ازای شش *Tiling* نشان داده‌شده‌اند. همانند اشکال ۵ و ۶ هیچ‌گونه یادگیری به ازای دو *Tile* صورت نپذیرفته است. اما اختلاف عملکرد یادگیری به ازای پنج و هشت Tile در مقایسه با اشکال ۵ و ۶ محسوس‌تر است.



شکل ۷. نمودار یادگیری کنترلرهای یادگیر تقویتی پیوسته به ازای شش *Tiling* و تعداد *Tile* های ۲، ۵ و ۸



شکل ۸ نمودار یادگیری کنترلرهای یادگیر تقویتی پیوسته به ازای نُه *Tiling* و تعداد *Tile* های ۲، ۵ و ۸

در شکل ۸ نیز نمودارهای یادگیری به ازای نُه *Tiling* نشان داده‌شده‌اند. واضح است که کنترلر یادگیر به ازای هشت *Tile* دارای بهترین عملکرد است. با بررسی اشکال ۵ تا ۸ می‌توان فهمید که افزایش تعداد *Tile* ها در صورتی منجر به بهبود عملکرد خواهد شد که تعداد *Tiling* ها نیز بالا باشد. در شکل ۵ به دلیل اینکه فقط از یک *Tiling* استفاده‌شده بود الگوریتم به ازای هشت *Tile* عملکرد مناسبی ندارد، زیرا در هر حالت محیط فقط ارزش یک *Tile* به‌روز می‌شود اما در اشکال ۷ و ۸ چون در هر حالت–عمل ارزش چندین *Tile* به‌روزرسانی می‌شود الگوریتم به ازای هشت *Tile* دارای عملکرد مناسبی است. در جدول ۱ متوسط عملکرد کنترلرهای یادگیر مختلف در ۵۰ اپیزود انتهایی نشان داده‌شده‌اند. در این جدول منظور از *Tile coding (x,y)*، روش *Tile coding* با تعداد *x Tiling* و *y Tile* است. همان‌طور که مشخص است اختلاف نسبتاً زیادی بین بهترین کنترلر به ازای یک *Tiling* با بهترین کنترلر ها به ازای سه، شش و نُه *Tiling* وجود دارد. دلیل این امر انعطاف‌پذیری پایین و سرعت یادگیری پایین به علت استفاده کردن از یک *Tiling* است. همچنین افزایش تعداد *Tiling* ها از یک به سه منجر به افزایش نسبتاً قابل‌ملاحظه‌ای در سرعت یادگیری شده است اما با افزایش تعداد *Tiling* ها به شش و نُه تغییر چندانی در سرعت یادگیری اتفاق نیافتاده و فقط تعداد پارامترهای θ افزایش‌یافته است.

جدول ۱. متوسط عملکرد کنترلرهای یادگیر به ازای مقادیر مختلف *Tile* ها و *Tiling* ها در ۵۰ اپیزود انتهایی

کنترلر یادگیر	متوسط زمان سفر (sec/km)	متوسط زمان توقف (sec/km)	متوسط تعداد توقف‌ها (#/veh/km)
Tile Coding (۱,۲)	۴۵۱.۶۲±۱۵.۱۱	۳۶۲.۸۸±۱۴.۸۴	۵.۰۵۴±۰.۱۲۳
Tile Coding (۱,۵)	۳۶۱.۹۹±۶.۷۴	۲۷۴.۸۵±۶.۵۷	۴.۳۶۴±۰.۰۶۴
Tile Coding (۱,۸)	۳۶۶.۱۸±۷.۷۶	۲۷۹.۱۱±۷.۵۷	۴.۳۸۰±۰.۰۷۷
Tile Coding (۳,۲)	۴۱۵.۴۳±۱۷.۴۷	۳۲۷.۲۴±۱۷.۰۷	۴.۹۹۶±۰.۱۵۲
Tile Coding (۳,۵)	۳۲۰.۷۶±۴.۴۳	۲۳۴.۲۸±۴.۲۸	۴.۱۸۶±۰.۰۴۲
Tile Coding (۳,۸)	۳۱۷.۱۷±۵.۷۵	۲۳۰.۷۹±۵.۶۰	۴.۱۴۶±۰.۰۴۹
Tile Coding (۶,۲)	۴۲۵.۰۲±۱۹.۱۲	۳۳۶.۷۴±۱۸.۷۷	۵.۱۵۹±۰.۱۵۸
Tile Coding (۶,۵)	۳۳۳.۶۲±۵.۸۵	۲۴۶.۸۲±۵.۷۰	۴.۲۹۶±۰.۰۵۶
Tile Coding (۶,۸)	۳۱۹.۷۰±۶.۵۶	۲۳۳.۳۱±۶.۳۹	۴.۱۶۳±۰.۰۶۴
Tile Coding (۹,۲)	۴۷۳.۰۵±۲۳.۳۱	۳۸۴.۳۷±۲۳.۰۰	۵.۳۹۹±۰.۱۹۷
Tile Coding (۹,۵)	۳۵۲.۶۶±۶.۳۹	۲۶۵.۵۶±۶.۲۱	۴.۴۱۲±۰.۰۶۰
Tile Coding (۹,۸)	۳۲۴.۰۴±۴.۳۴	۲۳۷.۶۱±۴.۲۱	۴.۱۸۰±۰.۰۴۷

همچنین با مقایسه انحراف از معیار بهترین کنترلرها به ازای *Tiling* های مختلف مشخص می‌شود که افزایش تعداد *Tiling* ها منجر به کاهش نوسانات نمودار یادگیری (انحراف از معیار) شده است. ازلحاظ متوسط عملکرد، کنترلر یادگیر بر پایه (۸ و ۳)*Tile Coding* دارای بهترین عملکرد و از منظر انحراف از معیار کنترلر یادگیر بر پایه (۸ و ۹)*Tile Coding* دارای بهترین عملکرد است.

۵- بحث و تحلیل نتایج

روش‌های یادگیری تقویتی گسسته جزء روش‌های بسیار رایج در کنترل وفق پذیر چراغ‌های راهنمایی هستند که در تحقیقات مختلفی [۸, ۱۷] مورد استفاده قرارگرفته‌اند. در این بخش به‌منظور اعتبار سنجی نتایج، دو روش یادگیری Q و عملگر-نقاد حالت گسسته [۱۰] پیاده‌سازی و نتایج آن با روش پیشنهادی مقایسه شدند. روش یادگیری Q یک روش *off-policy* است بدین معنی که سیاستی که با آن زندگی می‌کند ممکن است با آنچه بهبود می‌دهد متفاوت باشد. برای یادگیری تابع Q می‌توان از جدولی استفاده کرد که هر سطر آن یک زوج $<s,a>$ به همراه تقریبی است که یادگیر از مقدار واقعی Q به دست آورده است. نحوه به‌روزرسانی ارزش حالات-اعمال در یادگیری Q مطابق رابطه ۱۰ است.

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma \max_a Q(s',a') - Q(s,a)] \quad (10)$$

در رابطه ۱۰، r پاداش لحظه‌ای، γ ضریب تخفیف، α ضریب یادگیری و $Q(s,a)$ ارزش انجام عمل a در حالت s را نشان می‌دهند. ضریب تخفیف و مقدار ضریب یادگیری برای روش یادگیری Q در این تحقیق به ترتیب برابر ۰.۹۹ و ۰.۰۱ انتخاب شدند. روش عملگر-نقاد یک روش *on-policy* است بدین معنی که سیاستی که عامل با آن زندگی می‌کند با سیاستی که ارزش آن را تخمین می‌زند یکی است [۲۹]. در این روش یادگیری، ساختار حافظه جداگانه‌ای هم برای سیاست و هم برای تابع ارزش در نظر گرفته می‌شود. در این روش یادگیری ساختار سیاست به‌عنوان عملگر و ساختار تابع ارزش به‌عنوان نقاد شناخته می‌شوند. ارزش حالت‌های مختلف محیط در روش عملگر-نقاد توسط روابط ۱۱ و ۱۲ به‌روز می‌شوند.

$$V(s) \leftarrow V(s) + \alpha[r_{t+1} + \gamma V(s_{t+1}) - V(s_t)]e_{t+1}(s) \quad (11)$$

$$e_{t+1}(s) = \begin{cases} \gamma\lambda e_t(s) & \text{, if } s \neq s_t \\ \gamma\lambda e_t(s) + 1 & \text{, if } s = s_t \end{cases} \quad 0 \leq \gamma, \lambda \leq 1 \quad (12)$$

در روابط ۱۱ و ۱۲، α نرخ یادگیری در عملگر، γ نرخ تخفیف، λ میزان تأثیرپذیری ارزش حالت‌های ابتدایی اپیزود از ارزش حالت‌ها و سیگنال‌های انتهایی محیط هستند. مقدار نرخ یادگیری، ضریب تخفیف و λ در عملگر به ترتیب برابر ۰.۱، ۰.۹ و ۰.۸۵ انتخاب شدند.

جدول ۲. نحوه گسسته سازی حالت هر تقاطع

ابعاد جدول Q	تعداد حالات	مرزبندی حالات محیط	
۶×۶×۶×۶×۴×۸	۶	{۵-۲۰-۳۵-۵۰-۶۵}	خیابان ورودی اول
	۶	{۵-۲۰-۳۵-۵۰-۶۵}	خیابان ورودی دوم
	۶	{۵-۲۰-۳۵-۵۰-۶۵}	خیابان ورودی سوم
	۶	{۵-۲۰-۳۵-۵۰-۶۵}	خیابان ورودی چهارم
	۴	{۴ و ۳ و ۲ و ۱}	فاز

تقویت کردن و یا ضعیف کردن تمایل برای انتخاب هر عمل در عملگر می‌تواند توسط افزایش یا کاهش $P(s_t,a_t)$ در زمان‌های مختلف انجام شود (رابطه ۱۳).

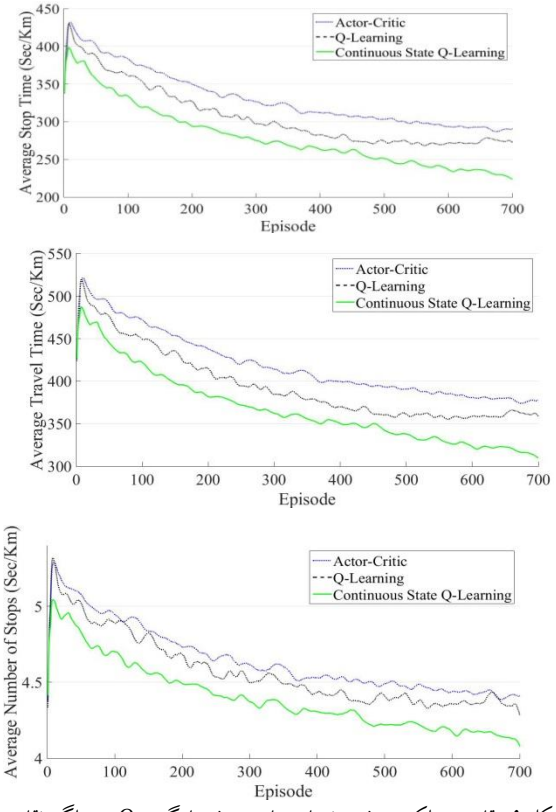
$$P(s,a) \leftarrow P(s,a) + \beta[r_{t+1} + \gamma V(s_{t+1}) - V(s_t)]e_{t+1}(s) \quad (۱۳)$$

در رابطه ۱۳، β پارامتر طول گام[1] بوده و دارای یک مقدار مثبت (۰.۱) می‌باشد. این دو روش نیاز به گسسته سازی حالات محیط دارند برای این منظور تعداد ماشین‌های موجود در خیابان‌های منتهی به هر تقاطع مطابق جدول ۲ گسسته سازی می‌شوند.

همچنین از سیاست ϵ-*greedy* جهت برقراری تعادل میان اکتشاف و بهره‌برداری استفاده شد که مقدار ϵ به‌صورت خطی در طول یادگیری از ۰.۸ تا صفر کاهش می‌یابد. شکل ۹ عملکرد کنترلرهای مبتنی بر یادگیری Q و عملگر-نقاد حالت گسسته را با بهترین کنترلر مبتنی بر یادگیری Q حالت پیوسته ((۸ و ۳)*Tile Coding*) در سه معیار متوسط زمان سفر، متوسط زمان توقف و متوسط تعداد توقف‌ها مقایسه می‌نماید.

همان‌طور که از شکل ۹ مشخص است روش پیشنهادی دارای عملکرد بهتری است. به‌منظور ارزیابی دقیق‌تر، متوسط عملکرد این سه کنترلر یادگیر در ۵۰ اپیزود انتهایی بر اساس سه معیار متوسط زمان سفر، متوسط زمان توقف و متوسط تعداد توقف‌ها در جدول ۳ نشان داده‌شده‌اند. همان‌طور که مشخص است استفاده از کنترلر یادگیر پیشنهادی منجر به بهبود ۱۶٪ زمان سفر در مقایسه با کنترلر عملگر-نقاد و بهبود ۱۳٪ زمان سفر در مقایسه با کنترلر یادگیر Q شده است.

به‌منظور فراهم آوردن درک ملموس‌تری از میزان تأثیر عملکرد کنترلر چراغ راهنمایی، میزان مصرف سوخت (*lit*) و انتشار آلاینده‌های *CO*، *HC* و NO_x (*kg*) برای کنترلرهای هوشمند مختلف بررسی می‌شوند. هدف از انجام چنین بررسی برقراری ارتباط میان مهندسی ترافیک، یادگیری ماشین و محیط‌زیست است. در شبیه‌سازی میکروسکوپیک ترافیک، به‌منظور محاسبه میزان مصرف سوخت و آلایندگی برای هر خودرو چهار حالت توقف کامل، کاهش شتاب، افزایش شتاب و حرکت با سرعت ثابت در نظر گرفته‌شده است و برای هر حالت میزان مصرف سوخت و میزان تولید آلاینده‌ها (CO، NO_x و HC) به‌صورت مجزا معرفی می‌شود. در طول شبیه‌سازی بر اساس سرعت هر خودرو و مسافتی که در شبکه طی می‌نماید میزان آلاینده‌ها و مصرف سوخت محاسبه می‌شوند [۳۰]. شکل ۱۰ میزان مصرف سوخت و انتشار آلاینده‌های مختلف را برای کنترلرهای مبتنی بر یادگیری Q، عملگر-نقاد حالت گسسته و بهترین کنترلر مبتنی بر یادگیری Q حالت پیوسته نشان می‌دهد.

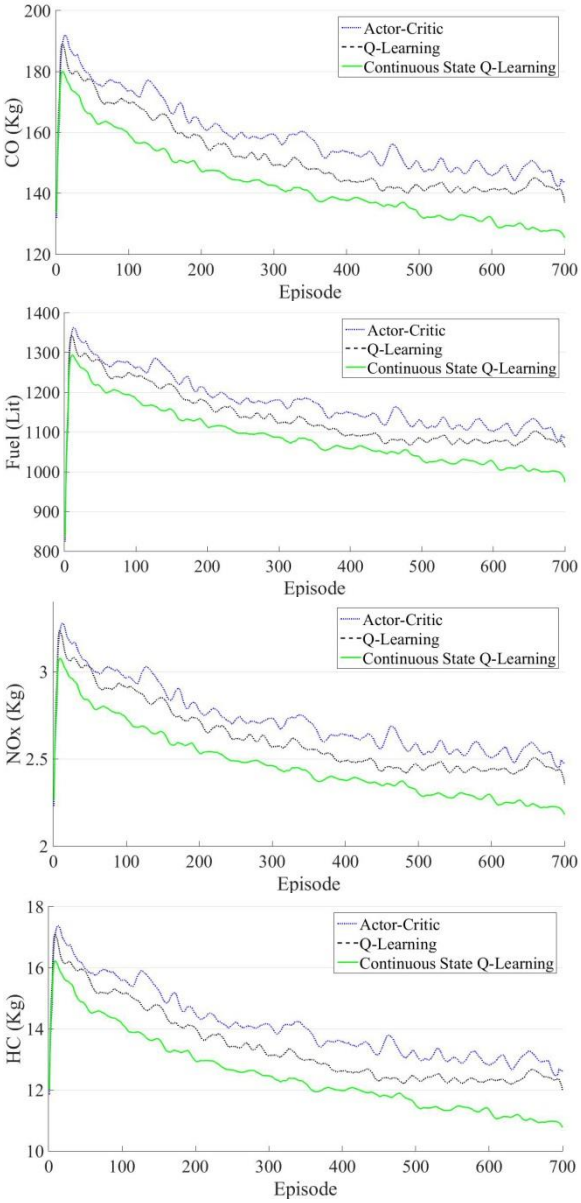


شکل ۹. مقایسه عملکرد روش پیشنهادی با دو روش یادگیری Q و عملگر-نقاد

[1] Step-Size

جدول ۳. مقایسه روش پیشنهادی با دو روش یادگیری Q و عملگر-نقاد

کنترلر	متوسط زمان سفر (sec/km)	متوسط زمان توقف (sec/km)	متوسط تعداد توقف‌ها (#/veh/km)
عملگر-نقاد	۳۷۷.۰۹±۵.۰۲	۲۸۹.۹۹±۵.۰۷	۴.۴۲±۰.۰۶
یادگیری Q	۳۶۳.۳۷±۸.۰۵	۲۷۶.۳۳±۷.۷۹	۴.۳۸±۰.۰۹
روش پیشنهادی	۳۱۷.۱۷±۵.۷۵	۲۳۰.۷۹±۵.۶۰	۴.۱۵±۰.۰۴۹
٪ بهبود روش پیشنهادی در مقایسه با عملگر-نقاد	۱۵.۹	۲۰.۴	۶.۱
٪ بهبود روش پیشنهادی در مقایسه با یادگیری Q	۱۲.۷	۱۶.۵	۵.۴



شکل ۱۰. مقایسه عملکرد روش پیشنهادی با دو روش یادگیری Q و عملگر-نقاد با معیارهای میزان مصرف سوخت و آلایندگی

لازم به ذکر است در این جدول منظور از CO، HC و NO_x به ترتیب میزان متوسط آلاینده CO، HC و NO_x تولیدشده در هر ساعت است. همان‌طور که مشخص است مجموع میزان سوخت مصرفی و آلاینده‌های مختلف در طول یادگیری دارای روند کاهشی بوده که این موضوع نشان‌دهنده یادگیری چراغ‌های راهنمایی و نزدیک شدن آن‌ها به سیاست بهینه است. همچنین کنترلر پیشنهادی بر اساس تمام معیارهای میزان مصرف سوخت و انتشار آلایندگی دارای عملکرد بهتری است.

به‌منظور ارزیابی دقیق‌تر، متوسط عملکرد این سه کنترلر یادگیر در ۵۰ اپیزود انتهایی بر اساس معیارهای فوق در جدول ۴ نشان داده‌شده‌اند. همان‌طور که مشخص است استفاده از کنترلر یادگیر پیشنهادی منجر به بهبود ۱۰٪ میزان مصرف سوخت و ۱۲٪ انتشار آلاینده NO_x در مقایسه با کنترلر عملگر-نقاد و بهبود ۸٪ میزان مصرف سوخت و ۱۲٪ انتشار آلاینده NO_x در مقایسه با کنترلر یادگیر Q شده است.

۶- نتیجه‌گیری

طراحی و پیاده‌سازی کنترلرهای چراغ‌های راهنمایی به دلیل نوسانات و پیچیدگی‌های ترافیکی ساده و عاری از چالش نیست. طراحی کنترلرهای چراغ راهنمایی بر اساس یادگیری تقویتی، توانایی وفق پذیری و انعطاف‌پذیری آن‌ها را با شرایط مختلف ترافیکی فراهم می‌آورد. در این تحقیق کنترلر یادگیر تقویتی بر اساس تقریب زن خطی *Tile Coding* به‌منظور کنترل چراغ راهنمایی طراحی و پیاده‌سازی شد. در یادگیری تقویتی حالت پیوسته از مفهوم تعمیم استفاده می‌شود بدین معنی که برخلاف یادگیری تقویتی گسسته عامل برای تقریب زدن ارزش تمام حالات یا حالات-اعمال نیازی به تجربه‌اندوزی مستقیم ندارد و ارزش یک حالت-عمل از روی شباهت سنجی با سایر حالات-اعمال مشابه تخمین زده می‌شود. هرچه دو حالت از محیط به هم شبیه‌تر باشند ارزش آن‌ها نیز به هم نزدیک‌تر خواهد بود.

جدول ۴. مقایسه روش پیشنهادی با دو روش یادگیری Q و عملگر–نقاد

کنترلر	میزان مصرف سوخت در هر ساعت (lit)	CO (kg)	HC (kg)	NO_x (kg)
عملگر–نقاد	۱۱۰۷±۴۳	۱۴۶.۷±۶.۴	۱۲.۹±۰.۵۹	۲.۵۲±۰.۱۱
یادگیری Q	۱۰۸۷±۳۲	۱۴۳.۰±۴.۵	۱۲.۵±۰.۴۰	۲.۴۷±۰.۰۸
روش پیشنهادی	۱۰۰۰±۲۴	۱۲۷.۹±۳.۲	۱۱.۰±۰.۲۹	۲.۲۲±۰.۰۶
٪ بهبود روش پیشنهادی در مقایسه با عملگر–نقاد	۹.۶	۱۲.۸	۱۴.۹	۱۱.۹
٪ بهبود روش پیشنهادی در مقایسه با یادگیری Q	۸.۰	۱۰.۶	۱۲.۲	۱۱.۳

همچنین در این تحقیق به‌منظور اعتبار سنجی، عملکرد یادگیری تقویتی حالت پیوسته با دو روش یادگیری Q و عملگر–نقاد حالت گسسته مقایسه شد و نتایج نشان دادند که یادگیری تقویتی حالت پیوسته به‌مراتب دارای عملکرد بهتری است.

مراجع

[1] A. Stevanovic, J. Stevanovic, K. Zhang, and S. Batterman, "Optimizing traffic control to reduce fuel consumption and vehicular emissions: Integrated approach with VISSIM, CMEM, and VISGAOST," Transportation Research Record: Journal of the Transportation Research Board, vol. 2128, no. 2128, pp. 105–113, 2009.

[2] S. F. Smith, G. J. Barlow, X.-F. Xie, and Z. B. Rubinstein, "Smart Urban Signal Networks: Initial Application of the SURTRAC Adaptive Traffic Signal Control System," Proceedings of the Twenty-Third International Conference on Automated Planning and Scheduling, Rome, Italy, pp. 434-442, 2013.

[3] M. C. Choy, D. Srinivasan, and R. L. Cheu, "Cooperative, hybrid agent architecture for real-time traffic signal control," IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, vol. 33, no. 5, pp. 597-607, 2003.

[4] P. G. Balaji and D. Srinivasan, "Type-2 fuzzy logic based urban traffic management," Engineering Applications of Artificial Intelligence, vol. 24, no. 1, pp. 12-22, 2011.

[5] D. Srinivasan, M. C. Choy, and R. L. Cheu, "Neural Networks for Real-Time Traffic Signal Control," IEEE Transactions on Intelligent Transportation Systems, vol. 7, no. 3, pp. 261-272, 2006.

[6] K.-H. Chao, R.-H. Lee, and M.-H. Wang, "An Intelligent Traffic Light Control Based on Extension Neural Network," KES '08 Proceedings of the 12th international conference on Knowledge-Based Intelligent Information and Engineering Systems, Part I, I. Lovrek, R. J. Howlett, and L. C. Jain (eds), Springer ,Berlin, Heidelberg, pp. 17-24, 2008.

[7] S. El-Tantawy, B. Abdulhai, and H. Abdelgawad, "Design of Reinforcement Learning Parameters for Seamless Application of Adaptive Traffic Signal Control," Journal of Intelligent Transportation Systems: Technology , Planning, and Operations vol. 18, no. 3, pp. 227-245, 2014.

[8] M. Abdoos, N. Mozayani, and A. L. C. Bazzan, "Traffic Light Control in Non-stationary Environments based on MultiAgent Q-learning," 14th International IEEE Conference on Intelligent Transportation Systems, Washington, DC, USA, pp. 1580-1585, 2011.

[9] C. Jacob and B. Abdulhai, "Automated Adaptive Traffic Corridor Control Using Reinforcement Learning: Approach and Case Studies," Transportation Research Record: Journal of the Transportation Research Board, vol. 1959, no. 1, pp. 1-8, 2006.

[10] R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction. MIT Press, Cambridge, 1998.

[11] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement Learning: A Survey," Journal of Artifcial Intelligence Research, vol. 4, pp. 237-285, 1996.

[12] B. Abdulhai and L. Kattan, "Reinforcement learning: Introduction to theory and potential for transport applications," Canadian Journal of Civil Engineering, vol. 30, no. 6, pp. 981-991, 2003.

[13] R. P. Roess, E. S. Prassas, and W. R. McShane, Traffic Engineering. Pearson Higher Education, New Jersey, 2010.

[14] K. Wen, S. Qu, and Y. Zhang, "A stochastic adaptive control model for isolated intersections," IEEE International Conference on Robotics and Biomimetics, ROBIO 2007, pp. 2256-2260, 2007.

[15] S. El-Tantawy and B. Abdulhai, "An agent-based learning towards decentralized and coordinated traffic signal control," 13th International IEEE Conference on Intelligent Transportation Systems (ITSC), pp. 665-670, 2010.

[16] J. C. Medina, A. Hajbabaie, and R. F. Benekohal, "Arterial traffic control using reinforcement learning agents and information from adjacent intersections in the state and reward structure," 13th International IEEE Conference on Intelligent Transportation Systems (ITSC), Funchal, 2010.

[17] M. Abdoos, N. Mozayani, and A. L. C. Bazzan, "Holonic multi-agent system for traffic signals control," Engineering Applications of Artificial Intelligence, vol. 26, no. 5–6, pp. 1575-1587, 2013.

[18] M. Wiering, "Multi-agent reinforcement learning for traffic light control," 17th International Conference on Machine Learning, Stanford, CA, pp. 1151-1158, 2000.

[19] M. Wiering, J. Vreeken, J. V. Veenen, and A. Koopman, "Simulation and optimization of traffic in a city," IEEE Intelligent Vehicles symposium, pp. 453 - 458, 2004.

[20] M. A. Khamis and W. Gomaa, "Adaptive multi-objective reinforcement learning with hybrid exploration for traffic signal control based on cooperative multi-agent framework," Engineering Applications of Artificial Intelligence, vol. 29, pp. 134-151, 2014.

[21] M. Abdoos, N. Mozayani, and A. L. C. Bazzan, "Hierarchical control of traffic signals using Q-learning with tile coding," Applied Intelligence, vol. 40, no. 2, pp. 201-213, 2014.

[22] C. Szepesvári, Algorithms for Reinforcement Learning. Morgan & Claypool Publishers, 2010.

[23] L. Busoniu, R. Babuska, B. D. Schutter, and D. Ernst, Reinforcement Learning and Dynamic Programming Using Function Approximators. CRC Press, Florida, 2010.

[24] J. S. Albus, "A New Approach to Manipulator Control: the Cerebellar Model Articulation Controller (CMAC)," Journal of Dynamic Systems, Measurement, and Control, vol. 97, pp. 220-227, 1975.

[25] A. Kesting, M. Treiber, and D. Helbing, "General Lane-Changing Model MOBIL for Car-Following Models," Transportation Research Record: Journal of the Transportation Research Board, vol. 1999, pp. 86-94, 2004.

[26] J. Casas, J. L. Ferrer, D. Garcia, J. Perarnau, and A. Torday, "Traffic Simulation with Aimsun," Fundamentals of Traffic Simulation, J. Barceló (ed), Springer New York, New York, NY, pp. 173-232, 2010.

[27] P. G. Gipps, "A behavioural car-following model for computer simulation," Transportation Research Part B: Methodological, vol. 15, no. 2, pp. 105–111, 1981.

[28] P. G. Gipps, "Multsim: a model for simulating vehicular traffic on multi-lane arterial roads," Mathematics and Computers in Simulation, vol. 28, no. 4, pp. 291-295, 1986.

[29] I. Grondman, L. Busoniu, G. A. D. Lopes, and R. Babuska, "A Survey of Actor-Critic Reinforcement Learning: Standard and Natural Policy Gradients," IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 42, no. 6, pp. 1291-1307, 2012.

[30] AIMSUN, "Microsimulator and Mesosimulator in Aimsun 6 User's Manual," 2009.

[DOR: 20.1001.1.20088345.1396.11.2.4.3]