

توسعه سامانه‌های چند عامله و یادگیری تقویتی در کنترل هوشمند چراغ‌های راهنمایی

محمد اصلانی*^۱، محمد سعدی مسگری^۲

دانشجوی دکتری مهندسی سیستم‌های اطلاعات مکانی - دانشکده مهندسی نقشه‌برداری - دانشگاه صنعتی خواجه

نصیرالدین طوسی

maslani@mail.kntu.ac.ir

دانشیار دانشکده مهندسی نقشه‌برداری - دانشگاه صنعتی خواجه نصیرالدین طوسی

mesgari@kntu.ac.ir

(تاریخ دریافت مهر ۱۳۹۵، تاریخ تصویب اردیبهشت ۱۳۹۶)

چکیده

امروزه یکی از معضلات جوامع شهری، ازدحام و ترافیک خودروها در معابر شهری است که منجر به آسیب‌های مختلف اقتصادی، محیط زیستی و اجتماعی می‌شود. برای جلوگیری از این آسیب‌ها، نیاز به بهبود زیر ساخت‌های فعلی حمل و نقل در شهرهای بزرگ بیش از پیش احساس می‌شود. تمرکز مقاله حاضر بر روی کنترل هوشمند چراغ‌های راهنمایی به عنوان یکی از شاخه‌های سیستم‌های حمل و نقل هوشمند با استفاده از سامانه‌های چند عامله یادگیر است. دو سناریوی متفاوت شامل کنترل یک تقاطع منفرد و کنترل یک شریان متشکل از چهار تقاطع پیاده‌سازی می‌شوند. در سناریوی اول دو روش یادگیری Q و سارسا با یکدیگر مقایسه و نتایج نشان می‌دهند که روش یادگیری Q بهتر از روش یادگیری سارسا عمل می‌نماید. اما در سناریوی دوم به دلیل وجود چندین عامل یادگیر تقویتی و تأثیر رفتار هر عامل یادگیر بر سایر عامل‌ها، نیاز است که اعمال عامل‌ها با یکدیگر سازگار شوند. به کارگیری یک مکانیسم هماهنگی برای پیدا نمودن رفتار بهینه تمام عامل‌ها یکی دیگر از اهداف تحقیق به شمار می‌رود. نتایج این تحقیق نشان می‌دهند که کنترل هوشمند چراغ‌های راهنمایی منجر به کاهش ۸۱٪ طول صف، ۷۸٪ زمان سفر، ۵۷٪ مصرف سوخت و ۷۳٪ آلودگی هوا نسبت به کنترل غیر هوشمند چراغ‌های راهنمایی می‌شود.

واژگان کلیدی: سامانه‌های چند عامله، یادگیری تقویتی، یادگیری Q، سارسا و کنترل هوشمند چراغ‌های راهنمایی

* نویسنده رابط

۱- مقدمه

افزایش جمعیت و در نتیجه آن افزایش فعالیت‌های اجتماعی و اقتصادی در شهرها، منجر به افزایش تقاضا برای حمل و نقل می‌شود [۱]. بر طبق پیش بینی‌های انجام شده تا سال ۲۰۵۰، ۷۰٪ جمعیت دنیا در شهرها زندگی خواهند کرد [۲]. این افزایش تقاضا برای حمل و نقل در شهرها به حدی است که زیرساخت‌های فعلی از جمله خیابان‌ها، چراغ‌های راهنمایی و فتاوری‌های بکار رفته در آن‌ها به خوبی جوابگوی نیازهای حمل و نقل نمی‌باشند.

ازدحام در تقاطع‌ها و صف‌های طولانی خودروها در پشت چراغ‌ها پدیده رایجی است که امروزه در شهرهای بزرگ همه روزه قابل مشاهده است. این ازدحام که به دو صورت روزمره و غیرمترقبه رخ می‌دهد در سطوح و جنبه‌های مختلف هزینه‌های بسیار زیادی را به جوامع تحمیل می‌نماید. ازدحام روزمره ناشی از بیشتر بودن تقاضا نسبت به ظرفیت در ساعات اوج عبور و مرور است در حالی که ازدحام غیرمترقبه ناشی از اتفاقات ناگهانی تاثیرگذار بر ترافیک مانند تصادف، بدی آب و هوا، نقص فنی و مواردی از این دست است. اگرچه احداث راه‌های جدید و افزایش ظرفیت معابر، باعث کاهش معضل ترافیک شهری می‌شوند اما باید در نظر داشت که این راه حل خود هزینه‌های بسیاری را در پی خواهد داشت. از طرف دیگر تجربه نشان داده است که افزایش ظرفیت در یک نقطه ممکن است باعث ایجاد ازدحام در مکان دیگری از شبکه ترافیکی شود و بنابراین پیاده‌سازی چنین راه حلی در مقیاس وسیع نیاز به هزینه و زمان زیادی دارد [۳].

سیستم‌های حمل و نقل هوشمند بدون تحمیل هزینه‌های زیاد می‌توانند نقش مؤثری در بهبود وضعیت ترافیکی ایفا نمایند. سیستم‌های حمل و نقل هوشمند، سیستم‌هایی هستند که فناوری‌های مخابرات، کنترل، سخت‌افزار و نرم‌افزار را در سیستم‌های حمل و نقل زمینی بکار می‌برند. این سیستم‌ها روش انعطاف‌پذیری را برای مدیریت و کنترل مؤثر ترافیک فراهم می‌آورند [۴]. در کنترل ترافیک عملکرد شبکه ترافیکی با کنترل جریان ترافیکی بهبود می‌یابد. کنترل ترافیک شامل اجزایی از جمله کنترل هوشمند چراغ‌های راهنمایی در تقاطع‌ها، کنترل جریان ورودی به بزرگراه‌ها، اعمال محدودیت

سرعت در نقاط مختلف، سیستم‌های مسیریابی و تابلوهای متغیر خبری است [۵].

تمرکز مقاله حاضر بر روی کنترل وفق پذیر چراغ‌های راهنمایی در تقاطع‌ها می‌باشد. در کنترل وفق پذیر چراغ‌ها، پارامترهای زمان بندی با توجه به شرایط ترافیکی موجود (به عنوان مثال، تعداد ماشین‌های منتظر در پشت تقاطع) برای رسیدن به یک سری اهداف خاص (به عنوان مثال، بیشترین تعداد ماشین‌های عبوری از تقاطع) بهینه می‌شوند. روش‌های متفاوتی برای کنترل وفق پذیر چراغ‌ها تاکنون در مهندسی ترافیک ارائه شده است از جمله اسکوت^۱ [۶]، اسکت^۲ [۷]، اپک^۳ [۸] و پرودین^۴ [۹] و رودس^۵ [۱۰]. سامانه‌های اسکت و اسکوت از روش‌های کنترل ترافیک متمرکز استفاده می‌کنند و در این سامانه‌ها تمام چراغ‌ها به کنترل کننده مرکزی متصل هستند. حجم بالای محاسبات و ارتباط تمامی چراغ‌ها با کنترل کننده مرکزی از جمله معایب این دو سامانه است. سامانه‌های اپک، پرودین و رودس به صورت توزیع یافته عمل می‌کنند و اساس عملکرد آن‌ها بر پایه پیش بینی ترافیک است. پیچیدگی‌های بالای محاسباتی از جمله معایب این سامانه‌ها است [۱۱].

در سال‌های اخیر روش‌های موجود در علوم کامپیوتر و به خصوص هوش مصنوعی توجه متخصصین را برای کنترل ترافیک به خود جلب نموده است. در این میان، سامانه‌های چند عامله به دلیل نزدیکی ویژگی‌های آن‌ها به طبیعت مسائل ترافیکی طرفداران زیادی در کنترل ترافیک پیدا کرده‌اند [۱۲]. از جمله این ویژگی‌ها می‌توان به توزیع یافتگی، هوشمندی و خودمختاری اشاره نمود. در تحقیق حاضر که از سیستم‌های چند عامله برای کنترل ترافیک استفاده شده است دو نوع عامل وجود دارند: عامل‌های چراغ راهنمایی (عامل‌های فعال) که دارای توانایی یادگیری می‌باشند و عامل‌های خودرو (عامل‌های غیرفعال) که دارای رفتارهایی از جمله شتاب گرفتن، ترمز کردن و سبقت گرفتن هستند.

با توجه به عدم قطعیت در محیط‌های ترافیکی، پیچیدگی و دینامیک بالای آن‌ها، به دست آوردن یک دانش اولیه از محیط و قرار دادن آن در ذهن عامل از ابتدا

۱ SCOOT
۲ SCAT
۳ OPAC
۴ PRODYN
۵ ROHDES

راهنمایی بر پایه یادگیری تقویتی (تعریف حالات محیط، اعمال و پاداش دهی) در بخش ۶ بحث می‌شود. در بخش ۷ نتایج پیاده سازی ارائه شده و نهایتاً بخش‌های ۸ و ۹ به بحث و نتیجه گیری اختصاص داده شده‌اند.

۲- مروری بر تحقیقات گذشته

تورپ در سال ۱۹۹۷ از یادگیری تقویتی و به طور خاص روش سارسا برای مدیریت چراغ‌های راهنمایی استفاده نمود. هدف عامل یادگیر در این تحقیق حداقل نمودن زمان مورد نیاز هر خودرو برای خارج شدن از شبکه ترافیکی است. عامل یادگیر در مرحله اول برای یک تقاطع ایزوله آموزشی دیده و در مرحله دوم برای کنترل چندین تقاطع بکار گرفته می‌شود. نتایج کار ایشان نشان می‌دهد که به کارگیری یادگیری تقویتی زمان توقف ماشین‌ها را به طور متوسط ۲۹٪ نسبت به حالتی که چراغ‌های راهنمایی به صورت زمان ثابت عمل نمایند کاهش می‌دهد [۱۵]. در این تحقیق مکانیسم مناسبی برای هماهنگی میان عامل‌های یادگیر ارائه نشده است.

بینگام در سال ۲۰۰۱ یک روش نروفازی را برای کنترل بهینه چراغ‌های راهنمایی ارائه نمود. در این روش قسمت بهینه سازی توابع، توسط یادگیری تقویتی انجام می‌شود. از آنجائی که تنها هدف این مقاله نشان دادن عملیاتی بودن این روش بوده است، عملکرد آن با روش دیگری مقایسه نشده است [۱۶]. همچنین یانگوان در سال ۲۰۰۹، تصمیم‌گیری قواعد فازی را بر اساس روش یادگیری Q بکار گرفت [۱۷].

باکر و همکاران در سال ۲۰۱۰ روش ارائه شده توسط ویرینگ را توسعه دادند [۱۸]. در این تحقیق از گراف‌های هماهنگ کننده برای هماهنگی میان عامل‌های یادگیر استفاده شده است. ایشان الگوریتم ماکس - پلاس را برای پیدا نمودن اعمال بهینه عامل‌ها استفاده نمودند. حجم بالای محاسبات برای هماهنگی میان عامل‌ها از معایب روش پیشنهادی است [۱۹].

خمیس و گوما در سال ۲۰۱۴ روش پیشنهادی ویرینگ [۱۸، ۲۰] را با استفاده از تئوری بایزن توسعه دادند. چراغ‌های راهنمایی یادگیر سعی در بهینه نمودن تلفیق خطی چندین شاخص از جمله متوسط زمان توقف در طول سفر، متوسط زمان توقف در هر تقاطع، سرعت و

امری بسیار دشوار است. بنابراین نیاز به یک مکانیسم یادگیری که از طریق آن، عامل در حین تصمیم‌گیری و تعامل هوشمندانه با محیط دارای عدم قطعیت بتواند دانش لازمه را به دست آورد کاملاً ضروری است. با توجه به ذات مسئله ترافیک مبنی بر عدم در دسترس بودن داده‌های آموزشی، استفاده از روش‌های یادگیری نظارت‌شده برای رسیدن به این مقصود منطقی نیست و نیاز به روشی است که بدون استفاده از داده‌های آموزشی و در عین حال هدایت‌شده خودکار (عدم استفاده از یادگیری نظارت‌نشده) بتواند دانش لازمه را به دست آورد. یادگیری تقویتی^۱ به عنوان یکی از نگرش‌های موفق در یادگیری ماشین می‌تواند چنین امکانی را فراهم نماید. در این روش، عامل سیاست بهینه (زمان بندی مناسب چراغ راهنمایی) را از طریق تعامل با محیط دارای عدم قطعیت به دست می‌آورد [۱۳]. یکی از مزایای یادگیری تقویتی عدم نیاز آن به مدل محیط و دخالت طراح (انسان) است. به عبارت بهتر عامل یادگیر به صورت پیوسته از طریق تعامل با محیط دانش لازمه را جهت عملکرد بهینه کسب می‌نماید. عامل یادگیر به صورت متوالی اعمالی را در محیط انجام می‌دهد و بازخورد آن اعمال را در قالب پاداش دریافت می‌کند. در این تحقیق به منظور بررسی‌های جامع‌تر دو روش سارسا^۲ [۱۳] و یادگیری Q^۳ [۱۴] مورد استفاده قرار می‌گیرند.

به کارگیری یادگیری تقویتی در کنترل ترافیک که یک مسئله چند عامله است با چالش‌های متفاوتی همراه است. مهم‌ترین چالش این است که عامل‌های یادگیر (چراغ‌های راهنمایی) در سطح انفرادی و مستقل در حال یادگیری می‌باشند و در نهایت رفتار بهینه فردی را پیدا می‌کنند که لزوماً منجر به رفتار بهینه کل عامل‌ها نخواهد شد. حل این چالش در کنترل ترافیک از اهداف اصلی این تحقیق به شمار می‌آید.

در این تحقیق در بخش ۲ به مروری بر تحقیقات مرتبط گذشته پرداخته می‌شود. در بخش ۳ شبیه سازی عامل مبنای ترافیک شرح داده می‌شود. روش بکار گرفته شده در این تحقیق (یادگیری تقویتی) در بخش ۴ مورد بررسی قرار خواهد گرفت. در بخش ۵ شبکه ترافیکی مورد مطالعه شرح داده می‌شود. نحوه طراحی کنترلر چراغ

^۱ Reinforcement learning

^۲ SARSA

^۳ Q-Learning

امنیت می‌نمایند. ایشان روش پیشنهادی خود را با روش ویرینگ (TC-1) مقایسه و نشان دادند که روش آن‌ها دارای عملکرد بهتری است. نقطه‌ضعف اساسی در تحقیق ایشان غیرعملی بودن تشخیص متوسط زمان سفر، متوسط زمان توقف برای هر خودرو است. زیرا تعیین زمان سفر توسط سامانه، در وهله اول نیاز به تشخیص هر خودرو به صورت مجزا در شبکه ترافیکی دارد. تشخیص هر خودرو و ردیابی آن در شبکه‌های ترافیکی بزرگ در عمل بسیار دشوار و تقریباً غیرممکن است [۵].

آل - تانتاوی و همکاران در سال ۲۰۱۴ عملکرد دو روش یادگیری Q و سارسا را در کنترل چراغ‌های راهنمایی یک شبکه ترافیکی دوبعدی در مرکز شهر تورنتو مقایسه کردند. همچنین آن‌ها به بررسی تأثیر نوع تعریف حالت محیط و نحوه پاداش‌دهی بر روی عملکرد کنترلر یادگیر پرداختند. برای این منظور سه تعریف مختلف از حالت محیط و چهار تعریف مختلف از تابع پاداش ارائه و مقایسه شدند. در تحقیق ایشان تعداد خودروهای ورودی به تقاطع و طول صف به عنوان بهترین تعریف حالت محیط و مجموع منفی زمان تأخیر خودروهای منتظر در تقاطع به عنوان بهترین تابع پاداش انتخاب شدند. نتایج نشان دادند که یادگیری Q دارای عملکرد بهتری در مقایسه با سارسا است و منجر به کاهش ۲۸٪ متوسط طول صف می‌شود [۲۱].

لی و همکاران در سال ۲۰۱۶ یادگیری تقویتی عمیق را جهت کنترل یک تقاطع ایزوله بکار گرفتند. یک شبکه عصبی با دولایه مخفی که به ترتیب دارای ۱۶ و ۴ نرون است به عنوان تقریب زن غیرخطی Q-value ها مورد استفاده قرار گرفت [۲۲].

قابل ذکر است که تمامی مراجعی که در فوق بررسی شده‌اند جزو برترین کارهایی هستند که تاکنون بر روی کنترل ترافیک انجام پذیرفته است و در آن‌ها به موضوع هماهنگی میان عامل‌ها در یادگیری تقویتی توجه ویژه‌ای نشده است. تمرکز و نوآوری تحقیق حاضر، هماهنگی میان عامل‌های یادگیر (چراغ‌های راهنمایی) است.

۳- شبیه سازی عامل مبنای ترافیک

خودروها و چراغ‌های راهنمایی به عنوان عامل‌های خودمختار و هوشمند در نظر گرفته می‌شوند. در این

تحقیق چراغ‌هایی راهنمایی دارای توانایی یادگیری تقویتی هستند و در طول شبیه سازی رفتار خود را بهبود می‌دهند اما خودروها که از نوع عامل‌های واکنشی بوده فاقد توانایی یادگیری می‌باشند و دارای رفتارهای ثابت و از قبل مشخص شده‌ای هستند. در هر گام شبیه سازی که در این تحقیق برابر ۱ ثانیه در نظر گرفته شده است موقعیت، سرعت، شتاب جدید خودروها و فاز چراغ‌های راهنمایی به روز می‌شوند. با توجه به میزان جریان ترافیک برای هر مسیر ورودی و مدل تولید ترافیک، ماشین‌ها در زمان‌های مختلف وارد مسیرهای ورودی می‌شوند. مدل تولید ترافیک در این تحقیق نمایی است، بدین معنی که بازه‌های زمانی بین دو ورود متوالی خودروها از یک تابع توزیع نمایی به دست می‌آید. عامل‌های خودرو دارای زمان واکنش هستند که این زمان عبارت است از مدت زمانی که طول می‌کشد تا راننده به تغییرات سرعت خودروی جلویی واکنش نشان دهد. زمان واکنش راننده برابر ۱ ثانیه در نظر گرفته شده است. عامل‌های خودرو دارای زمان واکنش در حالت توقف نیز می‌باشند که عبارت است از مدت زمانی که طول می‌کشد تا خودروی متوقف شده به شتاب گرفتن خودروی جلویی یا تغییرات فاز چراغ راهنمایی واکنش نشان دهد. این زمان واکنش فقط برای خودروهایی که متوقف هستند بکار گرفته می‌شود. در این تحقیق زمان واکنش در حالت توقف ۱٫۳۵ ثانیه در نظر گرفته شده است.

همواره خودروها در طول مسیر حرکت خود تمایل دارند تا به سرعت مطلوب برسند اما محیط اطراف آن‌ها (خودروی جلویی، خودروهای مجاور، چراغ راهنمایی و تابلوهای محدودیت سرعت) مانع از رسیدن آن‌ها به سرعت مطلوبشان می‌شود. در هر گام شبیه سازی موقعیت و سرعت هر خودرو توسط دو مدل تعقیب خودرو و تغییر خط به روز می‌شود. مدل تعقیب خودرو بکار رفته در این تحقیق توسط گیس توسعه داده شده است [۲۳]. این مدل از دو جزء افزایش شتاب و کاهش شتاب تشکیل می‌شود. مدل افزایش شتاب زمانی بکار می‌رود که راننده تمایل به افزایش سرعت داشته باشد و مدل کاهش شتاب زمانی بکار می‌رود که با توجه به محدودیت‌های اعمال شده توسط محیط راننده مجبور به کاهش شتاب (ترمز گرفتن) شود. برای جزئیات بیشتر در این زمینه خوانندگان می‌توانند به مرجع [۲۴] مراجعه نمایند.

میزان خوب یا بد بودن انجام آن عمل در حالت قبلی محیط است. هدف نهایی عامل یادگیر ماکزیمم نمودن پاداش‌های دراز مدت خود می‌باشد (رابطه ۱).

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \dots, 0 < \gamma < 1 \quad (1)$$

در رابطه ۱، γ ضریب تخفیف، r پاداش دریافتی و R مجموع کاهش یافته پاداش‌های دراز مدت است. رابطه ۱ نشان می‌دهد که پاداش‌هایی که در آینده دور به دست می‌آیند تأثیر کمتری نسبت به پاداش‌های نزدیک‌تر خواهند داشت. هرچه ضریب تخفیف به صفر نزدیک‌تر باشد عامل نزدیک بین‌تر و هرچه مقدار آن به یک نزدیک‌تر باشد عامل آینده نگرتر خواهد بود.

بر این اساس ارزش هر حالت از محیط عبارت است از متوسط پاداش دراز مدتی که عامل با شروع زندگی از آن حالت و دنبال نمودن سیاست π به دست می‌آورد. این مقدار بر طبق معادله بلمن به صورت زیر قابل تعریف است (رابطه ۲).

$$V^\pi(s, a) = \sum_a \pi(s, a) \sum_{s'} P_{ss'}^a [R_{ss'}^a + \gamma V^\pi(s')] \quad (2)$$

در رابطه ۲، a عمل عامل، s حالت فعلی محیط، s' حالت بعدی محیط، $\pi(s, a)$ احتمال انجام عمل a در حالت s است. همچنین $R_{ss'}^a$ و $P_{ss'}^a$ به صورت زیر تعریف می‌شوند (رابطه ۳).

$$\begin{aligned} R_{ss'}^a &= E\{r_{t+1} | s_t = s, a_t = a, s_{t+1} = s'\} \\ P_{ss'}^a &= E\{s_{t+1} = s' | s_t = s, a_t = a\} \end{aligned} \quad (3)$$

$P_{ss'}^a$ احتمال انتقال از حالت s به s' تحت انجام عمل a و $R_{ss'}^a$ پاداش دریافتی به دلیل انتقال از حالت s به s' تحت انجام عمل a است. حل یک مسئله یادگیری تقویتی به معنی پیدا نمودن سیاستی است که ارزش تمام حالت‌های محیط تحت آن سیاست بیشینه می‌شود (رابطه ۴).

$$\pi^* = \operatorname{argmax}_\pi (V^\pi(S)) \quad \forall (S) \quad (4)$$

از آنجائیکه در هر حالت از محیط عامل باید عمل بهینه را انتخاب نماید نیاز است که علاوه بر تعریف $V^\pi(s)$ ، $Q^\pi(s, a)$ نیز تعریف شود. $Q^\pi(s, a)$ نشان دهنده ارزش انجام دادن عمل a در حالت محیط s تحت سیاست π است (رابطه ۵).

$$Q^\pi(s, a) = \sum_{s'} P_{ss'}^a [R_{ss'}^a + \gamma V^\pi(s')] \quad (5)$$

مدل تغییر خط بکار رفته در این تحقیق نیز توسط گپیس توسعه داده شده است [۲۳] و یک پروسه تصمیم‌گیری برای مشخص نمودن تغییر خط یک خودرو است. دلایل مختلفی برای تغییر خط یک خودرو می‌توانند وجود داشته باشند. به عنوان مثال هنگامی که راننده قصد دارد که با سرعت زیادی حرکت کند اما در خط موجود به دلیل ترافیک یا سرعت پایین خودروی جلویی، چنین امکانی فراهم نباشد راننده مجبور است که مسیر خود را تغییر دهد. در این حالت ابتدا راننده بهبود سرعت حرکت و کاهش طول صف خودروهای جلویی را با تغییر خط حرکت خود بررسی و در صورت بهبود هر دو و وجود فضای کافی اقدام به تغییر خط می‌نماید.

در این تحقیق سامانه‌های چند عامله ناهمگن برای شبیه سازی و کنترل ترافیک مورد استفاده قرار می‌گیرند. در سامانه‌های چند عامله ناهمگن عامل‌ها (خودروها و چراغ‌های راهنمایی) دارای اهداف و اعمال غیر یکسان هستند. معماری تمامی عامل‌ها از نوع واکنشی است. در این نوع معماری فرآیند تصمیم‌گیری به صورت یک نگاهت مستقیم از وضعیت موجود به عملیات، پیاده سازی می‌شود و بر اساس یک مکانیزم تحریک-پاسخ که عمل تحریک بر اساس داده‌های حس شده است پایه گذاری می‌گردد. لازم به ذکر است که در این تحقیق از AIMSUN برای شبیه سازی ترافیکی و از زبان برنامه نویسی C++ به منظور توسعه یادگیری تقویتی استفاده شده است [۲۴].

۴- یادگیری تقویتی

یادگیری تقویتی یکی از روش‌های رایج و پرطرفدار در حوزه یادگیری ماشین و هوش مصنوعی است [۲۵]. این روش در روانشناسی، علوم اعصاب، کنترل بهینه و هوش مصنوعی به صورت گسترده بکار رفته است. در این روش عامل تصمیم گیر یادگیر در هر گام زمانی با محیط پیرامون خودش تعامل می‌کند به این معنی که در هر گام زمانی عامل یک نمایشی از حالت محیط را دریافت نموده $(S(t))$ و با توجه به سیاستی که نشان دهنده احتمال انجام اعمال مختلف در حالت‌های مختلف محیط است عملی را انجام می‌دهد. به محض انجام دادن یک عمل در محیط، پاسخی از محیط تحت عنوان پاداش (جریمه) به سمت عامل بازگردانده می‌شود که میزان این پاداش متناسب با

در رابطه ۵، a عمل عامل، s حالت فعلی محیط، s' حالت بعدی محیط، $\pi(s,a)$ احتمال انجام عمل a در حالت s است. روش های پیدا کردن سیاست بهینه به سه دسته کلی تقسیم می شوند: ۱- برنامه نویسی پویا^۲ ۲- مونت کارلو^۳ ۳- اختلاف زمانی^۴ [۲۶]. در روش برنامه نویسی پویا مدل محیط ($P_{ss'}^a$ و $R_{ss'}^a$) از ابتدا مشخص است اما در دو روش مونت کارلو و اختلاف زمانی نیازی به دانستن مدل محیط از ابتدا نیست. روش مونت کارلو در مقایسه با اختلاف زمانی دارای سرعت همگرایی پایینی است. در این تحقیق با توجه به اینکه مدل محیط از ابتدا مشخص نیست و نیاز به سرعت همگرایی بالایی است، از روش اختلاف زمانی استفاده شد. تاکنون روش های مختلفی بر پایه اختلاف زمانی در ادبیات یادگیری تقویتی ارائه شده اند. از میان آن ها، یادگیری Q و سارسا به دلیل محبوبیت آن ها در این تحقیق مورد استفاده قرار گرفته اند.

۴-۱- سارسا

الگوریتم سارسا یکی از الگوریتم های مشهور یادگیری تقویتی می باشد که نیازی به دانستن مدل محیط ندارد. در این الگوریتم عامل تحت سیاست π در حالت s عمل a را انجام داده و به حالت s' منتقل می شود و سپس عمل a' را در حالت s' انجام می دهد. ارزش عمل a در حالت s تحت رابطه ۶ به روز می شود.

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma Q(s',a') - Q(s,a)] \quad (6)$$

در رابطه ۶، r پاداش لحظه ای، γ ضریب تخفیف، α ضریب یادگیری و $Q(s,a)$ ارزش انجام عمل a در حالت s را نشان می دهند.

```

Initialize matrix Q
For each episode
  Select a random initial state (s)
  Select an action (a) among all possible actions in the
  current state (s) according to the policy of the agent
  Do while the goal state hasn't been reached (end of
  episode)
    • Take action (a), observe reward (r) and next state (s)
    • Select next action (a') among all possible actions in
    the new state (s')
    •  $Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma Q(s',a') - Q(s,a)]$ 
    •  $s \leftarrow s', a \leftarrow a'$ 
  End Do
End For

```

شکل ۱- الگوریتم سارسا [۲۶]

این روش از نوع On policy است بدین معنا که سیاستی که عامل با آن زندگی می کند با سیاستی که ارزش آن را تخمین می زند یکسان است. از آنجائی که سیاست بهینه همواره سیاست حریصانه است نیاز است که سیاست عامل در طول زندگی تدریجاً از تصادفی به حریصانه انتقال یابد. الگوریتم این روش در شکل ۱ ارائه شده است.

۴-۲- یادگیری Q

این الگوریتم یادگیری همانند سارسا نیازی به دانستن مدل محیط ندارد. در این الگوریتم عامل در حالت s عمل a را انجام می دهد و به حالت s' می رود ولی بر خلاف روش سارسا، در حالت s' فرض می شود که عامل عملی را انجام می دهد که دارای بیشترین ارزش است. یادگیری Q یک روش Off policy است بدین معنی که سیاستی که در زندگی دنبال می کند ممکن است با آنچه بهبود می دهد متفاوت باشد. برای یادگیری تابع Q می توان از جدولی استفاده کرد که هر سطر آن یک زوج $\langle s,a \rangle$ به همراه تقریبی است که یادگیر از مقدار واقعی Q به دست آورده است. ارزش عمل a در حالت s به صورت زیر به روز می شود (رابطه ۷).

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma \max_{a'} Q(s',a') - Q(s,a)] \quad (7)$$

در رابطه ۷، r پاداش لحظه ای، γ ضریب تخفیف، α ضریب یادگیری و $Q(s,a)$ ارزش انجام عمل a در حالت s را نشان می دهند. الگوریتم این روش در شکل ۲ ارائه شده است.

```

Initialize matrix Q
For each episode
  Select a random initial state
  Do while the goal state hasn't been reached
    • Select an action (a) among all possible actions for the
    current state (s)
    • Take action (a), observe reward (r) and next state (s)
    •  $Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma \max_{a'} Q(s',a') - Q(s,a)]$ 
    •  $s \leftarrow s'$ 
  End Do
End For

```

شکل ۲- الگوریتم یادگیری Q [۱۴]

^۱ Dynamic Programming

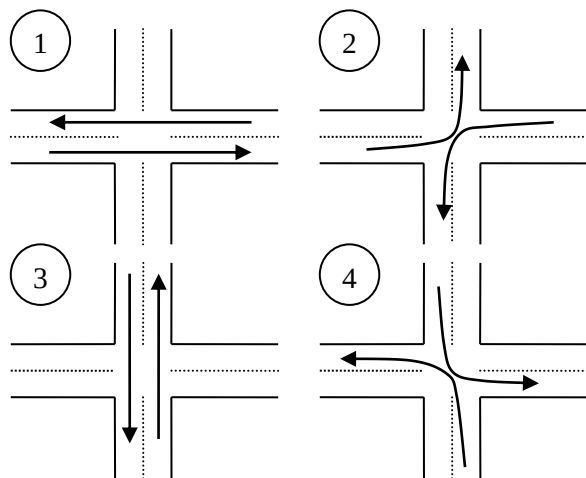
^۲ Monte Carlo

^۳ Temporal Difference

۵- شبکه ترافیکی مورد بررسی

۵-۱- تقاطع ایزوله

در تقاطع ایزوله مورد بررسی چهار مسیر به تقاطع منتهی و چهار مسیر از تقاطع خارج می‌شوند. طول هر کدام از مسیرهای متصل به تقاطع (ورودی و خروجی) ۳۰۰ متر و هر مسیر دارای ۳ خط عبور می‌باشد. چراغ هوشمند قرار گرفته در تقاطع دارای چهار فاز است که ترتیب این فازها در شکل ۳ آمده است.



شکل ۳- ترتیب فازهای چراغ راهنمایی

در فاز ۱ مسیر حرکت از غرب به شرق و بلعکس، در فاز ۲ مسیر حرکت از غرب به شمال و از شرق به جنوب، در فاز ۳ مسیر حرکت از شمال به جنوب و بلعکس و در فاز ۴ از شمال به شرق و از جنوب به غرب آزاد (سبز) می‌شود. چراغ موجود در تقاطع از روش چرخه‌ای تبعیت می‌نماید. بدین معنی که ترتیب فازها مشخص و ثابت است و فقط زمان فازها تغییر می‌کند و در نتیجه طول کل چرخه متغیر خواهد بود. در ابتدای هر فاز و براساس تعداد ماشین‌های منتظر در هر مسیر ورودی، برای فاز جاری مدت زمانی به عنوان زمان سبز انتخاب می‌شود. بعد از اتمام زمان سبز در هر فاز، ۵ ثانیه به عنوان زمان زرد برای تخلیه کامل تقاطع لحاظ شده و سپس فاز بعدی آغاز می‌شود. جریان ترافیکی برای همه مسیرهای ورودی یکسان فرض شده است. برای شبیه سازی ۶۰٪ از ماشین‌ها مسیر مستقیم، ۲۰٪ گردش به چپ و ۲۰٪ گردش به راست را انجام می‌دهند. جریان ورودی در تمام مسیرها برابر $500 \frac{Veh}{h}$ است. شبیه سازی ترافیکی برای ۱۰۰۰ ساعت (۴۱٫۶ روز) انجام گرفته است.

۵-۲- شریان متشکل از چندین تقاطع

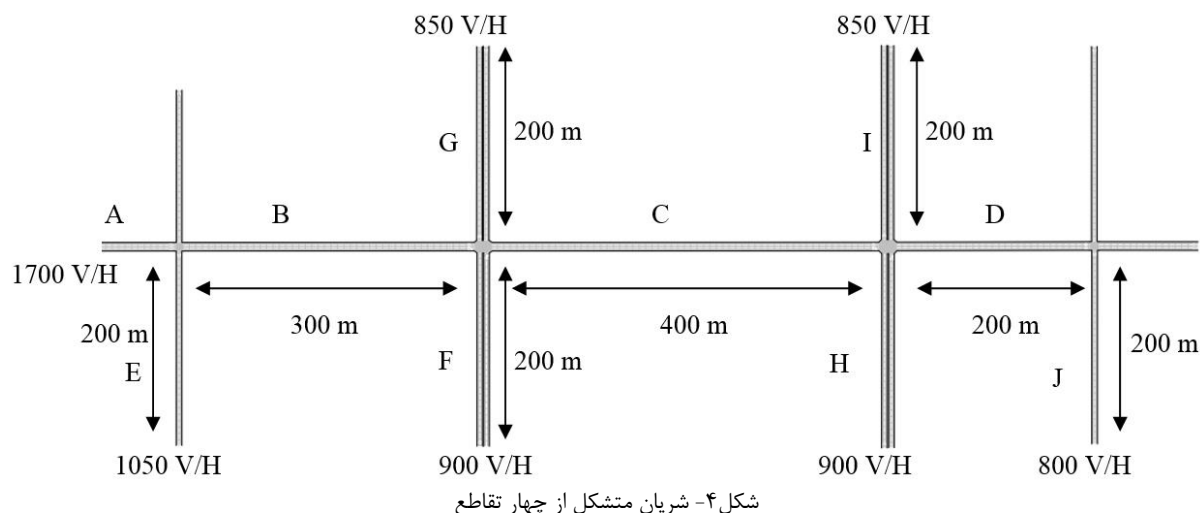
شکل ۴ شریان ترافیکی مورد بررسی را به همراه طول مسیرها و جهات آنها نشان می‌دهد. در این شریان چهار تقاطع وجود دارند که تقاطع‌های اول و چهارم فاز ۲ و تقاطع‌های دوم و سوم فاز ۳ هستند. در چراغ‌های دو فاز، در فاز ۱ مسیرهای غرب به شرق و شمال و در فاز ۲ مسیرهای جنوب به شمال و شرق سبز می‌شوند. در چراغ‌های ۳ فاز، در فاز ۱ مسیرهای غرب به شرق و جنوب و شرق و در فاز ۲ مسیرهای شمال به جنوب و شرق و در فاز ۳ مسیرهای جنوب به شمال سبز می‌شوند. شریان اصلی غرب به شرق دارای ۳ خط عبور و سایر مسیرها دارای ۲ خط عبور هستند. فاصله بین تقاطع‌های اول و دوم ۳۰۰ متر، تقاطع‌های دوم و سوم ۴۰۰ متر و تقاطع‌های سوم و چهارم ۲۰۰ متر هستند. در شکل ۴ جریان‌های ورودی به تمام مسیرها نشان داده شده است. همان‌طور که مشخص است جریان مسیرهای ورودی مختلف با یکدیگر یکسان نمی‌باشند که این موضوع بر پیچیدگی‌های ترافیکی اضافه می‌نماید. همچنین فرض شده است در تقاطع‌ها، ۶۰٪ از ماشین‌ها مسیر مستقیم، ۲۰٪ گردش به راست و ۲۰٪ گردش به چپ را انتخاب می‌کنند. در این سناریو همانند سناریوی اول شبیه سازی برای ۱۰۰۰ ساعت انجام گرفته است.

۶- طراحی کنترلر چراغ‌های راهنمایی بر

اساس یادگیری تقویتی

۶-۱- تعریف اعمال عامل‌ها

برای معین کردن اعمال عامل بر روی محیط برای هر فاز یک حداقل و حداکثر زمان سبز مشخص می‌گردد و بین این دو مقدار بازه‌های ده‌ثانیه‌ای برای گسسته سازی اعمال در نظر گرفته می‌شوند. حداکثر و حداقل زمان سبز برای هر فاز به ترتیب ۹۰ و ۱۰ در نظر گرفته شدند (۹ عمل متفاوت). ذکر این نکته ضروری است که این مقادیر بر اساس اطلاعات گذشته حجم ترافیک تقاطع و با آزمون و خطا قابل تعیین هستند و به این ترتیب نقش دانش مهندسی ترافیک بارز می‌گردد. در یادگیری تقویتی برای هر زوج حالت-عمل یک مقدار ارزش تعیین می‌شود $(Q(s, a))$. ابعاد جدول Q که نشان دهنده تعداد زوج‌های حالت-عمل هستند در جدول ۱ برای یک تقاطع منفرد و در جدول ۲ برای یک شریان آورده شده‌اند.



۶-۲- تعریف حالت محیط

نحوه تعریف درک عامل از حالت محیط در یادگیری آن نقش بسزایی بازی می‌کند. مسلماً از دیدگاه‌های گوناگون می‌توان حالت‌های متفاوتی برای تقاطع در نظر گرفت. هرچه تعداد حالت‌های محیط بیشتر باشند فضای حالت برای عامل یادگیر بزرگ‌تر شده و یادگیری در اصطلاح دچار نفرین بعدی^۱ می‌شود و در نتیجه نیاز به تعداد تکرارهای بیشتری برای همگرایی به سیاست بهینه خواهد بود. همچنین در صورتی که تعداد حالت‌های محیط کمتر از مقدار مورد نیاز تعریف شوند عملکرد عامل یادگیر بعد از یادگیری به خوبی انجام نخواهد پذیرفت. نکته حائز اهمیت دیگر این است که نحوه درک عامل از محیط باید به گونه‌ای تعریف شود که محیط از دید عامل دارای ویژگی مارکوف باشد [۱۳]. نکته دیگر، عملیاتی بودن تعریف حالت‌ها است. در این تحقیق تعداد ماشین‌های موجود در هر مسیر ورودی منتهی به تقاطع به عنوان حالت محیط در نظر گرفته می‌شوند. مهم‌ترین مزیت این تعریف این است که بار ترافیکی به نوعی در تعریف وضعیت محیط گنجانده می‌شود. دیگر مزیت این تعریف، مدیریت کردن خودروهای پرسرنشین است. به عنوان مثال می‌توان به اتوبوس‌ها یا خودروهای پرسرنشین ضریب بالاتری را نسبت داد تا وسایل نقلیه عمومی در اولویت بالاتری قرار گیرند. در این تحقیق حالت محیط برای تقاطع منفرد توسط یک بردار با ۵ المان نمایش داده می‌شود. به گونه‌ای که المان اول نشان دهنده

فاز جاری چراغ و ۴ المان باقی مانده نشان دهنده تعداد ماشین‌ها در ۴ مسیر ورودی هستند. دو الگوریتم یادگیری Q و سارسا، نیاز به گسسته سازی حالات محیط دارند. برای این منظور، تعداد ماشین‌های به دسته‌های خطی با طول ۱۰ افراز شدند (جدول ۱). همان‌طور که مشخص است در مجموع محیط از دید عامل دارای ۱۵۳۶۶۴ وضعیت مختلف است. برای کنترل یک شریان متشکل از چندین تقاطع وضعیت محیط برای هر چراغ راهنمایی به دو روش متفاوت تعریف می‌شوند. در روش اول وضعیت محیط برای هر تقاطع یک بردار با $N+1$ المان است که المان اول آن بیان کننده فاز جاری چراغ و N المان بعدی نشان دهنده تعداد ماشین‌های قرار گرفته در N مسیر ورودی به تقاطع است. در این روش چراغ‌های راهنمایی هر تقاطع به صورت مستقل و بدون در نظر گرفتن سایر عامل‌ها یادگیری را انجام و به رفتار بهینه همگرا می‌شوند. اما لزومی ندارد که رفتار بهینه هر عامل یادگیر منجر به همگرایی به رفتار بهینه برای کل سیستم (چندین عامل) شود. برای حل این چالش، در روش دوم وضعیت محیط برای هر چراغ با $N+1+1$ المان تعریف می‌شود که $N+1$ المان اول آن همانند روش اول بوده و المان آخر نشان دهنده تعداد ماشین‌هایی است که در خیابان اتصالی به تقاطع بعدی حضور دارند (جدول ۲).

^۱ Curse of dimensionality

جدول ۱- نحوه گسسته سازی حالت تقاطع

ابعاد جدول Q	تعداد حالت	مرزبندی حالت‌های محیط
۱۴×۱۴×۱۴×۱۴×۴×۹	۱۴	{۳۰-۴۰-۵۰-۶۰-۷۰-۸۰-۹۰-۱۰۰-۱۱۰-۱۲۰-۱۳۰-۱۴۰-۱۵۰}
	۱۴	{۳۰-۴۰-۵۰-۶۰-۷۰-۸۰-۹۰-۱۰۰-۱۱۰-۱۲۰-۱۳۰-۱۴۰-۱۵۰}
	۱۴	{۳۰-۴۰-۵۰-۶۰-۷۰-۸۰-۹۰-۱۰۰-۱۱۰-۱۲۰-۱۳۰-۱۴۰-۱۵۰}
	۱۴	{۳۰-۴۰-۵۰-۶۰-۷۰-۸۰-۹۰-۱۰۰-۱۱۰-۱۲۰-۱۳۰-۱۴۰-۱۵۰}
	۴	{۱ و ۲ و ۳ و ۴}
		مسیر ورودی ۱
		مسیر ورودی ۲
		مسیر ورودی ۳
		مسیر ورودی ۴
		فاز

جدول ۲- نحوه گسسته سازی وضعیت محیط برای چندین تقاطع

تعداد کل حالت‌ها	تعداد حالت	مرزبندی حالت‌های محیط	متغیر تشکیل دهنده حالت	شماره عامل	روش	
بدون مکانیسم هماهنگی	۱۸×۸×۲×۹	{۲۰-۳۰-۴۰-۵۰-۶۰-۷۰-۸۰-۹۰-۱۰۰-۱۱۰-۱۲۰-۱۳۰-۱۴۰-۱۵۰-۱۶۰-۱۷۰-۱۸۰}	مسیر ورودی A	عامل ۱		
			مسیر ورودی E			
			فاز			
	۱۵×۸×۸×۳×۹	۱۵	{۲۰-۳۰-۴۰-۵۰-۶۰-۷۰-۸۰-۹۰-۱۰۰-۱۱۰-۱۲۰-۱۳۰-۱۴۰-۱۵۰}	مسیر ورودی B	عامل ۲	
		۸	{۱۰-۲۰-۳۰-۴۰-۵۰-۶۰-۷۰}	مسیر ورودی F		
		۸	{۱۰-۲۰-۳۰-۴۰-۵۰-۶۰-۷۰}	مسیر ورودی G		
		۳	{۱ و ۲ و ۳}	فاز		
	۱۸×۸×۸×۳×۹	۱۸	{۲۰-۳۰-۴۰-۵۰-۶۰-۷۰-۸۰-۹۰-۱۰۰-۱۱۰-۱۲۰-۱۳۰-۱۴۰-۱۵۰-۱۶۰-۱۷۰-۱۸۰}	مسیر ورودی C	عامل ۳	
		۸	{۱۰-۲۰-۳۰-۴۰-۵۰-۶۰-۷۰}	مسیر ورودی H		
		۸	{۱۰-۲۰-۳۰-۴۰-۵۰-۶۰-۷۰}	مسیر ورودی I		
		۳	{۱ و ۲ و ۳}	فاز		
		۱۰×۸×۲×۹	۱۰	{۲۰-۳۰-۴۰-۵۰-۶۰-۷۰-۸۰-۹۰-۱۰۰}		
۸	{۱۰-۲۰-۳۰-۴۰-۵۰-۶۰-۷۰}		مسیر ورودی J			
۲	{۱ و ۲}		فاز			
با مکانیسم هماهنگی	۱۸×۸×۸×۲×۹	{۲۰-۳۰-۴۰-۵۰-۶۰-۷۰-۸۰-۹۰-۱۰۰-۱۱۰-۱۲۰-۱۳۰-۱۴۰-۱۵۰-۱۶۰-۱۷۰-۱۸۰}	مسیر ورودی A	عامل ۱		
			مسیر ورودی E			
			مسیر خروجی B			
			فاز			
	۱۵×۸×۸×۱۱×۳×۹	۱۵	{۲۰-۳۰-۴۰-۵۰-۶۰-۷۰-۸۰-۹۰-۱۰۰-۱۱۰-۱۲۰-۱۳۰-۱۴۰-۱۵۰}	مسیر ورودی B	عامل ۲	
		۸	{۱۰-۲۰-۳۰-۴۰-۵۰-۶۰-۷۰}	مسیر ورودی F		
		۸	{۱۰-۲۰-۳۰-۴۰-۵۰-۶۰-۷۰}	مسیر ورودی G		
		۱۱	{۲۰-۴۰-۶۰-۸۰-۱۰۰-۱۲۰-۱۴۰-۱۶۰-۱۸۰-۲۰۰}	مسیر خروجی C		
		۳	{۱ و ۲ و ۳}	فاز		
	۱۸×۸×۸×۶×۳×۹	۱۸	{۲۰-۳۰-۴۰-۵۰-۶۰-۷۰-۸۰-۹۰-۱۰۰-۱۱۰-۱۲۰-۱۳۰-۱۴۰-۱۵۰-۱۶۰-۱۷۰-۱۸۰}	مسیر ورودی C	عامل ۳	
		۸	{۱۰-۲۰-۳۰-۴۰-۵۰-۶۰-۷۰}	مسیر ورودی H		
		۸	{۱۰-۲۰-۳۰-۴۰-۵۰-۶۰-۷۰}	مسیر ورودی I		
		۶	{۲۰-۴۰-۶۰-۸۰-۱۰۰}	مسیر خروجی D		
		۳	{۱ و ۲ و ۳}	فاز		
	۱۰×۸×۲×۹	۱۰	{۲۰-۳۰-۴۰-۵۰-۶۰-۷۰-۸۰-۹۰-۱۰۰}	مسیر ورودی D	عامل ۴	
		۸	{۱۰-۲۰-۳۰-۴۰-۵۰-۶۰-۷۰}	مسیر ورودی J		
		۲	{۱ و ۲}	فاز		

۳-۶- تعریف پاداش

اختلاف تعداد خودروها در همه مسیرهای ورودی قبل و بعد از هر فاز عنوان تابع پاداش در نظر گرفته می‌شوند. به عنوان نمونه، اگر بعد از اتمام فاز اول ۳۰ ماشین در تمام مسیرهای ورودی و بعد از اتمام فاز دوم ۱۰ ماشین در تمام مسیرهای ورودی باشند پاداش ۲۰+ به عامل داده می‌شود و این پاداش به معنی عملکرد مناسب چراغ است. اما اگر بعد از اتمام فاز دوم ۵۰ ماشین در تمام مسیرهای ورودی باشند پاداش ۲۰- (جریمه) دریافت می‌کند که به معنی عملکرد نامناسب چراغ است.

۴-۶- تعریف سیاست

در این تحقیق از سیاست حریصانه کنترل شده^۱ استفاده شده است. بر طبق این سیاست عامل در اغلب مواقع در هر وضعیت محیط عملی را که دارای بالاترین ارزش است را انتخاب می‌کند و در ϵ دفعات سایر اعمال را به صورت تصادفی انتخاب می‌کند. رابطه ۸ نحوه عملکرد سیاست حریصانه کنترل شده را نشان می‌دهد.

$$P(s, a_i) = \begin{cases} 1 - \epsilon + \frac{\epsilon}{|A_s|}, & \text{if } a_i = \operatorname{argmax}_{a' \in A_s} Q(s, a') \\ \frac{\epsilon}{|A_s|}, & \text{else} \end{cases} \quad (8)$$

در رابطه ۸، $P(s, a_i)$ احتمال انتخاب عمل a_i در حالت s ، $|A_s|$ تعداد اعمال ممکن، $Q(s, a')$ ارزش عمل a' در حالت s و ϵ احتمال انتخاب اعمال غیر بهینه بر اساس دانش فعلی عامل هستند. مقدار ϵ_0 برابر ۰٫۷ که در طول یادگیری به صورت آهسته کاهش می‌یابد.

۷- نتایج پیاده‌سازی

۷-۱- کنترل ترافیک یک تقاطع با استفاده از روش یادگیری Q و سارسا

برای بررسی عملکرد هر یک از سیستم‌های کنترلی از سه معیار استفاده شده است. معیار اول متوسط طول صف است. این معیار عبارت است متوسط تعداد ماشین‌هایی که در هر خط در هر مسیر ورودی در زمان مشخصی قرار می‌گیرند. معیار دوم زمان سفر برای کل شبکه است. این

معیار عبارت است از میانگین زمانی که یک خودرو نیاز دارد تا یک کیلومتر را در شبکه طی نماید. نحوه محاسبه زمان سفر به این صورت است که زمان سفر برای هر ماشین محاسبه شده و برای کل خودروها میانگین‌گیری و به واحد زمان بر کیلومتر تبدیل می‌شود. معیار سوم، انحراف از معیار زمان سفر است که نشان دهنده عدالت و مساوات الگوریتم بین تمام خودروها می‌باشد. شکل ۵ عملکرد دو الگوریتم یادگیری Q و سارسا را برای سه معیار ذکر شده برای ۱۰۰۰ ساعت شبیه سازی ترافیکی نشان می‌دهد. خطوط نازک از میانگین‌گیری نتایج ۵ بار پیاده‌سازی و خطوط ضخیم از میانگین‌گیری خطوط نازک در بازه‌های ۱۱ ساعته به دست آمده‌اند.

نمودارها در اپیزود ۹۰۰ دچار شکستگی شده‌اند و دلیل این موضوع تغییر سیاست عامل از حریصانه کنترل شده به حریصانه^۲ است. به عبارت دیگر از اپیزود ۰ تا ۹۰۰ عامل تجربه‌اندوزی کرده و از اپیزود ۹۰۰ تا ۱۰۰۰ از تجربه به دست آمده استفاده می‌نماید. برای مقایسه بهتر دو الگوریتم یادگیری Q و سارسا متوسط مقادیر میانگین طول صف، زمان سفر و انحراف از معیار زمان سفر برای اپیزودهای ۹۰۰ تا ۱۰۰۰ محاسبه شده و در جدول ۳ آورده شده‌اند.

جدول ۳- مقایسه عملکرد دو روش یادگیری Q و سارسا برای اپیزودهای ۹۰۰ تا ۱۰۰۰

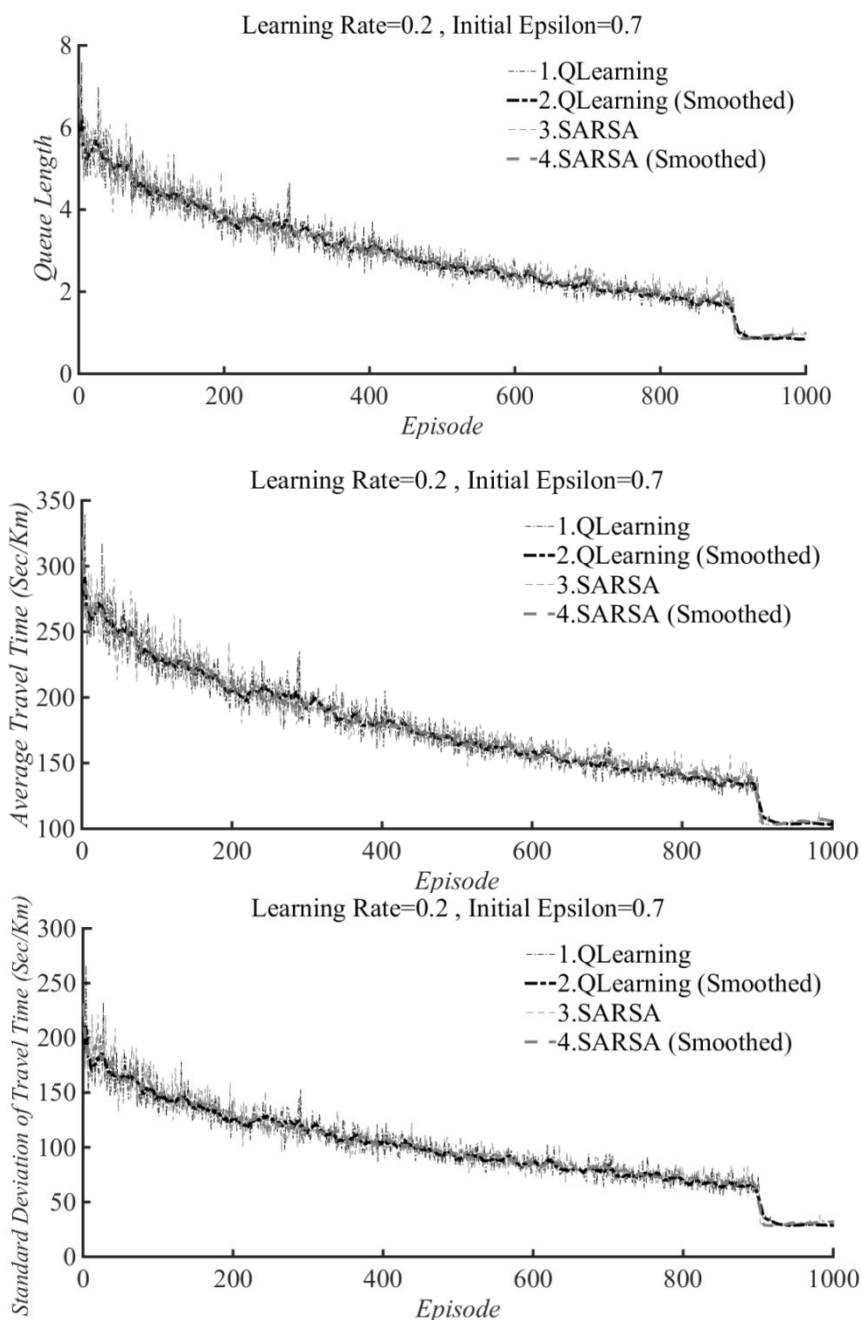
	یادگیری Q	سارسا
متوسط طول صف	۰٫۹۰	۰٫۹۳
زمان سفر (ثانیه/کیلومتر)	۱۰۵٫۴۶	۱۰۵٫۶۶
انحراف از معیار زمان سفر (ثانیه/کیلومتر)	۳۱٫۱۲	۳۱٫۱۱

جدول ۴- متوسط طول صف در دو روش یادگیری Q و سارسا برای مسیرهای ورودی مختلف

	یادگیری Q	سارسا
متوسط طول صف برای مسیر ورودی غرب به شرق	۰٫۹۰۴	۰٫۹۲۹
متوسط طول صف برای مسیر ورودی جنوب به شمال	۰٫۹۱۳	۰٫۸۹۶
متوسط طول صف برای مسیر ورودی شرق به غرب	۰٫۹۰۴	۰٫۹۳۴
متوسط طول صف برای مسیر ورودی شمال به جنوب	۰٫۹۰۲	۰٫۹۵۶

^۲ Greedy

^۱ ϵ -Greedy



شکل ۵- طول صف، زمان سفر و انحراف معیار زمان سفر در دو الگوریتم یادگیری Q و سارسا

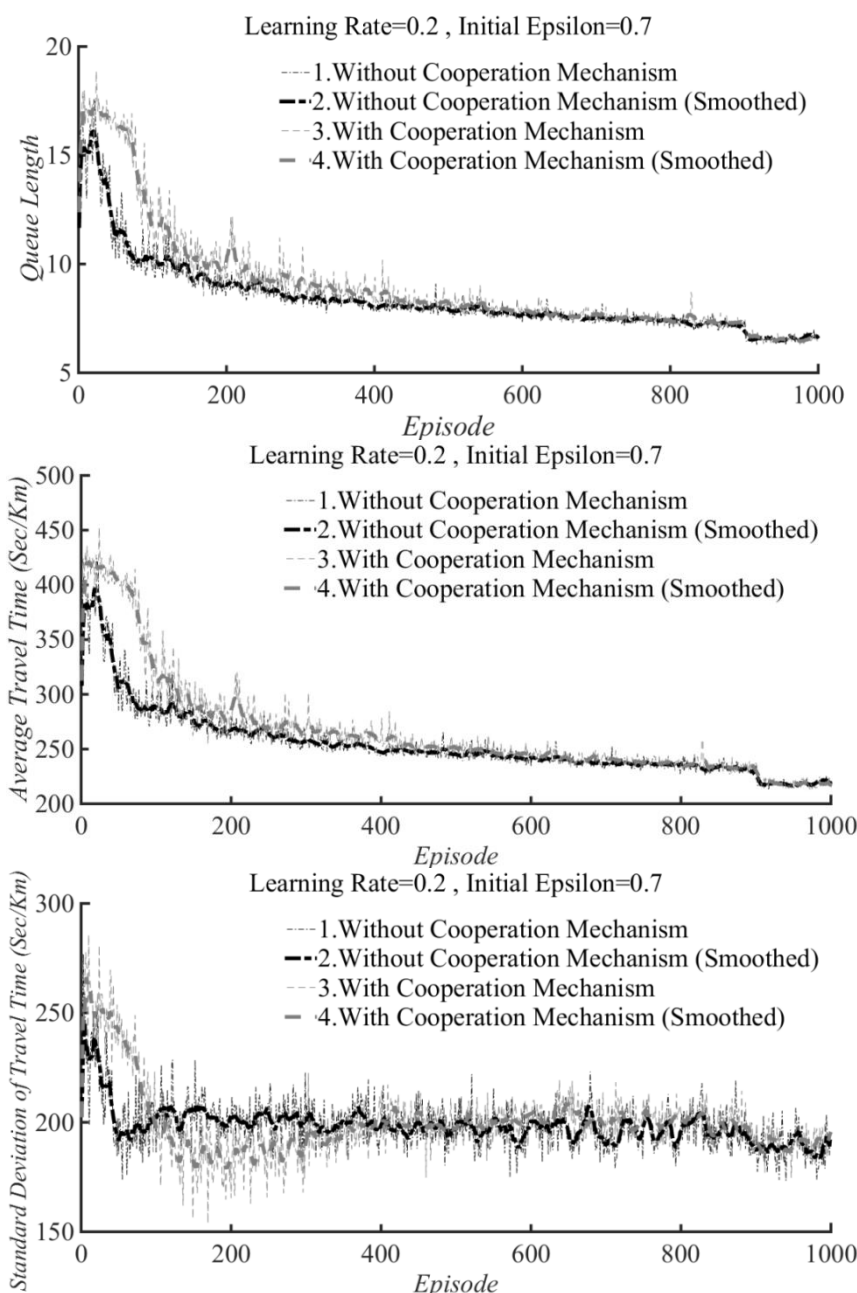
۷-۲- کنترل ترافیک یک شریان متشکل از چهار تقاطع با استفاده از یادگیری Q

به دلیل اینکه روش یادگیری Q برای کنترل یک تقاطع بهتر از روش سارسا عمل نموده است در کنترل یک شریان متشکل از چندین تقاطع نیز از روش یادگیری Q استفاده شد. شکل ۶ نتایج عملکرد کنترل یک شریان را برای سه معیار میانگین طول صف، زمان سفر و انحراف از معیار زمان سفر برای دو حالت یادگیری مستقل و یادگیری هماهنگ در طول ۱۰۰۰ ساعت شبیه سازی ترافیکی نشان می دهد. در

همان طور که از جدول ۳ مشخص است متوسط طول صف، زمان سفر و انحراف از معیار زمان سفر برای روش یادگیری Q کمتر از روش سارسا است که این موضوع نشان دهنده عملکرد بهتر روش یادگیری Q است. برای ارائه دقیق تر نتایج پیاده سازی، متوسط طول صف در هر مسیر ورودی برای اپیزود ۹۰۰ تا ۱۰۰۰ در جدول ۴ به تفکیک آمده است. همان طور که مشخص است به طور متوسط عملکرد یادگیری Q کمی بهتر از سارسا می باشد.

یادگیری هماهنگ هر عامل وضعیت خیابان متصل شده به عامل بعدی را نیز در نظر می گیرد.

یادگیری مستقل هر عامل یادگیر بدون در نظر گرفتن عامل مجاور و به صورت مستقل یادگیری را انجام می دهد اما در



شکل ۶- سه معیار طول صف، زمان سفر و انحراف معیار زمان سفر در دو روش یادگیری مستقل و یادگیری هماهنگ

یادگیری مستقل است و همین امر باعث کندتر شدن یادگیری می شود. اما از اپیزود ۵۰۰ تا ۹۰۰ تقریباً هر دو روش تقریباً یکسان عمل نموده اند و از اپیزود ۹۰۰ تا ۱۰۰۰ که سیاست عامل ها به حریصانه تبدیل شده است روش یادگیری هماهنگ کمی بهتر از یادگیری مستقل عمل نموده است.

در شکل ۶، خطوط نازک از میانگین گیری نتایج ۴ بار پیاده سازی و خطوط ضخیم از میانگین گیری خطوط نازک در بازه های ۱۱ ساعته به دست آمده اند. همان طور که مشخص است در اپیزودهای ۰ تا ۵۰۰ روش یادگیری مستقل با توجه به دو معیار طول صف و متوسط زمان سفر بهتر از یادگیری هماهنگ عمل نموده است که این موضوع نیز قابل انتظار می باشد. زیرا در یادگیری هماهنگ فضای حالت برای هر عامل بزرگ تر از فضای حالت در

جدول ۵- مقایسه عملکرد یادگیری مستقل و یادگیری هماهنگ برای
اپیزودهای ۹۰۰ تا ۱۰۰۰

یادگیری هماهنگ	یادگیری مستقل	
۶,۵۸۶	۶,۵۹۳	متوسط طول صف
۲۱۸,۷۶۵	۲۱۸,۸۴۵	زمان سفر (ثانیه/کیلومتر)
۱۹۰,۱۱۰	۱۸۸,۷۸۷	انحراف از معیار زمان سفر (ثانیه/کیلومتر)

بیشتری را به مسیر B به نسبت مسیر G می‌دهد اما در تقاطع سوم دادن زمان سبز بیشتر به مسیر C به نسبت مسیر I به معنی دریافت سود کمتر برای عامل خواهد بود. بنابراین مقدار متوسط طول صف فقط در مسیر ورودی G بالا می‌باشد.

جدول ۶- مقایسه متوسط طول صف برای دو حالت یادگیری مستقل و یادگیری هماهنگ در مسیرهای ورودی مختلف

یادگیری هماهنگ	یادگیری مستقل	
۲,۵۸۰	۲,۳۲۲	متوسط طول صف برای مسیر ورودی A
۵,۲۱۹	۵,۳۳۶	متوسط طول صف برای مسیر ورودی B
۵,۴۷۸	۵,۵۸۸	متوسط طول صف برای مسیر ورودی C
۱,۵۵۸	۱,۶۷۱	متوسط طول صف برای مسیر ورودی D
۳,۱۷۶	۳,۰۱۵	متوسط طول صف برای مسیر ورودی E
۵,۶۴۹	۵,۷۷۰	متوسط طول صف برای مسیر ورودی F
۲۹,۵۵۶	۲۹,۴۲۲	متوسط طول صف برای مسیر ورودی G
۵,۲۵۱	۵,۱۵۶	متوسط طول صف برای مسیر ورودی H
۵,۵۹۶	۵,۴۹۴	متوسط طول صف برای مسیر ورودی I
۱,۷۹۷	۲,۱۵۱	متوسط طول صف برای مسیر ورودی J

۸- بحث و تحلیل نتایج

به منظور اعتبارسنجی، نتایج به دست آمده از این تحقیق با روش زمان ثابت مقایسه شدند. در روش زمان ثابت [۲۷] چراغ‌های راهنمایی توانایی یادگیری نداشته و در طول شبانه‌روز بدون در نظر گرفتن نوسانات ترافیکی زمان بندی ثابتی را برای فازهای مختلف ارائه می‌دهند. جدول ۷ متوسط طول صف، زمان سفر و انحراف از معیار زمان سفر را برای کنترل یک تقاطع منفرد با استفاده از روش زمان ثابت و یادگیری Q مقایسه می‌کند. میزان بهبود در این جدول مطابق رابطه ۹ محاسبه شده است.

$$(9) \quad \text{بهبود} \% = \frac{\text{روش یادگیری} - \text{روش زمان ثابت}}{\text{روش زمان ثابت}} \times 100$$

جدول ۷- مقایسه عملکرد کنترل یک تقاطع منفرد با روش یادگیری Q و زمان ثابت

میزان بهبود	یادگیری Q	زمان ثابت	
۸۵,۵٪	۰,۹۰۶	۶,۳۳۲	متوسط طول صف
۶۳,۷۷٪	۱۰۵,۴۶۳	۲۹۱,۰۶	زمان سفر (ثانیه/کیلومتر)
۸۴,۰۶٪	۳۱,۱۱۷	۱۹۵,۲۴۷	انحراف از معیار زمان سفر (ثانیه/کیلومتر)

برای مقایسه دقیق‌تر دو روش یادگیری مستقل و یادگیری هماهنگ متوسط طول صف، زمان سفر و انحراف از معیار زمان سفر برای بازه‌های ۹۰۰ تا ۱۰۰۰ در جدول ۵ آورده شده‌اند. جدول ۵ نشان می‌دهد که در نظر گرفتن وضعیت خیابان متصل به تقاطع بعدی باعث کاهش زمان سفر و متوسط طول صف و افزایش انحراف از معیار زمان سفر شده است. به عبارت بهتر هنگامی که یادگیری هماهنگ صورت می‌پذیرد مساوات کمتری میان خودروها از نظر زمان سفر برقرار می‌شود. دلیل این موضوع این است که در یادگیری هماهنگ چون وضعیت شریان اصلی (مسیرهای B، C و D) که دارای بالاترین حجم ترافیکی است توسط چند عامل یادگیر در نظر گرفته می‌شود، عامل‌ها به صورت خودکار در طول یادگیری سعی می‌کنند تا ماشین‌های بیشتری را از شریان اصلی عبور دهند تا پاداش بیشتری دریافت نمایند و نتیجتاً به مسیرهای ورودی فرعی توجه کمتری می‌کنند که این موضوع باعث افزایش انحراف از معیار زمان سفر می‌شود.

برای ارائه دقیق‌تر نتایج پیاده‌سازی متوسط طول صف در هر مسیر ورودی برای اپیزودهای ۹۰۰ تا ۱۰۰۰ در جدول ۶ به تفکیک ذکر شده است. همان‌طور که مشخص است متوسط طول صف برای مسیر ورودی G با سایر مسیرهای ورودی اختلاف زیادی دارد. علت این موضوع این است که چون جریان ترافیکی مسیر ورودی G به نسبت سایر مسیرها کمتر است و با توجه به اینکه هرچه عامل یادگیر ماشین‌های بیشتری را از تقاطع عبور دهد پاداش بیشتری می‌گیرد، بنابراین چراغ راهنمایی برای مسیرهای ورودی B و F به دلیل بالاتر بودن جریان ترافیکی زمان سبز بیشتر و برای مسیر ورودی G زمان سبز کمتری را در نظر می‌گیرد که همین امر باعث افزایش بی رویه طول صف در مسیر ورودی G شده است.

از طرف دیگر کمترین نرخ پر شدن مسیر B و C در طول زمان‌های مختلف به ترتیب برابر $315 \frac{Veh}{h}$ و $255 \frac{Veh}{h}$ است بنابراین در تقاطع دوم الگوریتم یادگیری زمان سبز

جدول ۸- مقایسه عملکرد کنترل یک تقاطع منفرد با روش یادگیری Q و زمان ثابت

میزان بهبود	یادگیری Q	زمان ثابت	
۹۰,۷۰٪	۳۷,۵۳	۱۹,۶۸	متوسط سرعت در هر ساعت (کیلومتر/ساعت)
۸۳,۲۴٪	۳۷,۷۳	۲۲۵,۱۷	متوسط زمان تأخیر در هر ساعت (ثانیه/کیلومتر)
۸۶,۱۵٪	۲۹,۹۳	۲۱۶,۱۲	متوسط زمان توقف در هر ساعت (ثانیه/کیلومتر)
۴۳,۰۸٪	۱۱۷,۷۷	۲۰۶,۹۲	متوسط سوخت مصرف شده در هر ساعت (لیتر)
۵۳,۱۳٪	۱۴,۳۸	۳۰,۶۸	متوسط آلایند CO تولید شده در هر ساعت (کیلوگرم)
۵۹,۵۰٪	۱,۱۳	۲,۷۹	متوسط آلایند HC تولید شده در هر ساعت (کیلوگرم)
۵۵,۳۲٪	۰,۲۱	۰,۴۷	متوسط آلایند NOx تولید شده در هر ساعت (کیلوگرم)

خودرو چهار حالت توقف کامل، کاهش شتاب، افزایش شتاب و حرکت با سرعت ثابت در نظر گرفته شده است و در هر حالت مقادیر متفاوتی برای آلایندگی خودروها در نظر گرفته شد. جدول ۹- مقایسه عملکرد کنترل یک شریان با روش یادگیری

میزان بهبود	یادگیری هماهنگ	زمان ثابت	هماهنگ و زمان ثابت
۸۱,۶۴٪	۶,۵۸۶	۳۵,۸۷۸	متوسط طول صف
۷۸,۹۳٪	۲۱۸,۷۶۵	۱۰۳۸,۲۴	زمان سفر (ثانیه/کیلومتر)
۶۷,۶۱٪	۱۹۰,۱۱۰	۵۸۶,۹۵۷	انحراف از معیار زمان سفر (ثانیه/کیلومتر)
۱۳۴,۳۱٪	۲۴,۱۱	۱۰,۲۹	متوسط سرعت در هر ساعت (کیلومتر/ساعت)
۸۴,۳۳٪	۱۵۲,۲	۹۷۱,۲۹۹	متوسط زمان تأخیر در هر ساعت (ثانیه/کیلومتر)
۸۶,۰۹٪	۱۳۴,۶۴	۹۶۷,۹۳	متوسط زمان توقف در هر ساعت (ثانیه/کیلومتر)
۵۷,۹۸٪	۷۷۹,۲۲	۱۸۵۴,۵۵	متوسط سوخت مصرف شده در هر ساعت (لیتر)
۷۲,۵۶٪	۹۸,۷۵	۳۵۹,۸۳۱	متوسط آلایند CO تولید شده در هر ساعت (کیلوگرم)
۷۴,۰۳٪	۷,۸۲	۳۰,۱۱	متوسط آلایند HC تولید شده در هر ساعت (کیلوگرم)
۷۵,۱۱٪	۱,۷۹	۷,۱۹	متوسط آلایند NOx تولید شده در هر ساعت (کیلوگرم)

همان طور که از جدول ۸ مشخص است میزان مصرف سوخت ۴۳٪ و میزان تولید آلاینده های مختلف ۵۳٪ کاهش یافته است. این موضوع نشان دهنده اهمیت محیط زیستی و اقتصادی روش بکار گرفته شده است. جدول ۹ نیز مقادیر شاخص های مختلف ترافیکی را برای کنترل یک شریان در دو روش زمان ثابت و یادگیری Q نشان می دهد. همان طور که از جدول ۹ مشخص است میزان مصرف سوخت ۵۷٪ و متوسط میزان تولید آلاینده های مختلف ۷۳٪ کاهش یافته است.

۹- نتیجه گیری

روش های کنترل ترافیک متمرکز نمی توانند به خوبی بر پیچیدگی های ترافیکی موجود در شبکه های شهری غلبه کنند. گرایش به سمت کنترل غیرمتمرکز ترافیک در دهه های اخیر به طور ملموسی پدیدار شده است. سامانه های چند عامله با توجه به ذات توزیع شده کنترل ترافیک می توانند در این زمینه مفید باشند؛ اما بدیل

همان طور که از جدول ۷ مشخص است به کارگیری روش یادگیری Q منجر به بهبود قابل ملاحظه ای در متوسط طول صف و زمان سفر شده است. به منظور ارزیابی های بیشتر جدول ۸ عملکرد روش یادگیری Q و زمان ثابت را با توجه به شاخص های زیر مقایسه می نماید.

- متوسط سرعت: نشان دهنده متوسط سرعت کلیه خودروها در طول مسیر حرکتشان است.
- زمان توقف: متوسط زمان توقف هر ماشین در هر کیلومتر می باشد.
- زمان تأخیر: نشان دهنده زمان تأخیر برای هر ماشین در هر کیلومتر است. این مقدار از اختلاف بین زمان سفر در شرایط ایده آل (بدون ترافیک) و زمان سفر در شرایط ترافیکی موجود محاسبه می شود.
- مصرف سوخت: میزان کل سوخت مصرف شده برای تمام خودروها است. برای هر خودرو چهار حالت توقف کامل، کاهش شتاب، افزایش شتاب و حرکت با سرعت ثابت در نظر گرفته شده است و در هر حالت میزان مصرف سوخت متفاوتی برای خودروها در نظر گرفته شده است.
- آلایند CO، HC و NOx: نشان دهنده میزان آلایند های مختلف تولید شده توسط تمام خودروها است. برای محاسبه میزان تولید آلایند نیز همانند سوخت مصرفی برای هر

دارد و بنابراین عمل هر عامل یادگیر باید سازگار با اعمال سایر عامل‌ها باشد تا بتواند عملکرد کلی سیستم ترافیکی را بهبود دهد. در این تحقیق به نوعی هماهنگی میان عامل‌ها در حالت محیط گنجانده شد به این معنی که هر عامل یادگیر علاوه بر در نظر گرفتن وضعیت خیابان‌های متصل به تقاطع خود وضعیت خیابان‌ها متصل به تقاطع بعدی را نیز در حالت محیط لحاظ می‌کند. روش پیشنهادی در این تحقیق با روش کنترل زمان ثابت مقایسه شد و مشخص گردید که روش یادگیری هماهنگ منجر به کاهش ۸۱٪ متوسط طول صف، ۷۸٪ زمان سفر، ۵۷٪ مصرف سوخت و ۷۳٪ آلودگی هوا می‌شود.

وجود پیچیدگی‌های بالا در ترافیک در نظر گرفتن رفتارهای از پیش تعریف شده برای عامل‌ها بسیار دشوار است. بنابراین نیاز به یک روش یادگیری که عامل‌ها به صورت تدریجی در طول یادگیری به رفتار بهینه دست پیدا کنند احساس می‌شود. یادگیری تقویتی به عنوان یکی از روش‌های مناسب برای ایجاد یادگیری در عامل‌ها می‌تواند در این زمینه کارا باشد. یکی از چالش‌های بزرگ در به کارگیری یادگیری تقویتی در سامانه‌های چند عامله ایجاد هماهنگی میان عامل‌ها برای پیدا نمودن رفتارهایی است که منجر به رسیدن به بهینه کلی شوند. زیرا در یادگیری تقویتی چند عامله، تأثیر عمل هر عامل یادگیر بر روی محیط به عمل سایر عامل‌های یادگیر نیز بستگی

مراجع

- [1] B. Bhatta, (2010). "Analysis of Urban Growth and Sprawl from Remote Sensing Data." Verlag Berlin Heidelberg: Springer.
- [2] A. Pizam, (1999). "Life and tourism in the year 2050." International Journal of Hospitality Management. Vol. 18, No. 4, PP. 331-343.
- [3] S. Timotheou, C. G. Panayiotou, and M. M. Polycarpou, (2015). "Transportation Systems: Monitoring, Control, and Security." in Intelligent Monitoring, Control, and Security of Critical Infrastructure Systems, E. Kyriakides and M. Polycarpou, Eds., ed Berlin Heidelberg: Springer, PP. 125-166.
- [4] M. A. Chowdhury and A. W. Sadek, (2003). "Fundamentals of Intelligent Transportation Systems Planning ". Norwood, MA: Artech House.
- [5] M. A. Khamis and W. Gomaa, (2014). "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, No. PP. 134-151.
- [6] P. B. Hunt, D. I. Robertson, R. D. Bretherton, and R. I. Winton, "SCOOT - a traffic responsive method of coordinating signals," Crowthorne, U.K.1981.
- [7] A. G. Sims and K. W. Dobinson, (1980). "The Sydney coordinated adaptive traffic (SCAT) system philosophy and benefits." IEEE Transactions on Vehicular Technology. Vol. 29, No. 2, PP. 130 - 137
- [8] N. H. Gartner, (1983). "OPAC: A demand-responsive strategy for traffic signal control." Transportation Research Record: Journal of the Transportation Research Board. Vol. 906, No. PP. 75-81.
- [9] J. J. Henry, J. L. Farges, and J. Tufal, (1983). "The PROLYN real-time traffic algorithm." in Proceedings of the 5th IFAC/IFIP/IFORS Symposium on Control in Transportation Systems, Baden, Germany.
- [10] K. L. Head, P. B. Mirchandani, and D. Sheppard, (1992). "Hierarchical framework for real-time traffic control." Transportation Research Record. Vol. 1360, No. PP. 82-88.
- [11] A. L. C. Bazzan, (2009). "Opportunities for multiagent systems and multiagent reinforcement learning in traffic control." Autonomous Agents and Multi-Agent Systems. Vol. 18, No. 3, PP. 342-375.
- [12] L. B. d. Oliveira and E. Camponogara, (2010). "Multi-agent model predictive control of signaling split in urban traffic networks." Transportation Research Part C: Emerging Technologies. Vol. 18, No. 1, PP. 120-139.
- [13] R. S. Sutton and A. G. Barto, (1998). "Introduction to Reinforcement Learning." Cambridge, MA: MIT Press.
- [14] C. Watkins and P. Dayan, (1992). "Q-learning." Machine learning. Vol. 8, No. 3, PP. 279-292.
- [15] T. Thorpe, (1997), "Vehicle Traffic Light Control Using SARSA." MSc, Computer Science Department, Colorado State University, Fort Collins, Colorado.
- [16] E. Bingham, (2001). "Reinforcement learning in neurofuzzy traffic signal control." European Journal of Operational Research. Vol. 131, No. 2, PP. 232-241.

- [17] L. Yongquan and C. Xiangjun, (2009). "Study on Traffic Signal Control Based on Q-Learning." in Sixth International Conference on Fuzzy Systems and Knowledge Discovery, Tianjin, pp. 581 - 585.
- [18] M. Wiering, (2000). "Multi-agent reinforcement learning for traffic light control." in 17th International Conference on Machine Learning, Stanford, CA, pp. 1151-1158.
- [19] B. Bakker, S. Whiteson, L. Kester, and F. C. A. Groen, (2010). "Traffic Light Control by Multiagent Reinforcement Learning Systems." in Interactive Collaborative Information Systems, R. Babuška and F. C. A. Groen, Eds., ed Berlin Heidelberg: Springer, PP. 475-510.
- [20] M. Wiering, J. Vreeken, J. V. Veenen, and A. Koopman, (2004). "Simulation and optimization of traffic in a city." in IEEE Intelligent Vehicles symposium, pp. 453 - 458.
- [21] S. El-Tantawy, B. Abdulhai, and H. Abdelgawad, (2014). "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.
- [22] L. Li, Y. Lv, and F. Y. Wang, (2016). "Traffic signal timing via deep reinforcement learning." IEEE/CAA Journal of Automatica Sinica. Vol. 3, No. 3, PP. 247-254.
- [23] P. G. Gipps, (1981). "A behavioural car-following model for computer simulation." Transportation Research Part B: Methodological. Vol. 15, No. 2, PP. 105–111.
- [24] J. Casas, J. L. Ferrer, D. Garcia, J. Perarnau, and A. Torday, (2010). "Traffic Simulation with Aimsun." in Fundamentals of Traffic Simulation, J. Barceló, Ed., ed New York, NY: Springer New York, PP. 173-232.
- [25] R. S. Sutton and A. G. Barto, (1981). "Toward a Modern Theory of Adaptive Networks: Expectation and Prediction." Psychological Review. Vol. 88, No. 2, PP. 135-170.
- [26] R. S. Sutton and A. G. Barto, (1998). "Reinforcement Learning: An Introduction." Cambridge: MIT Press.
- [27] R. P. Roess, E. S. Prassas, and W. R. McShane, (2010). "Traffic Engineering." New Jersey: Pearson Higher Education.