توسعه یادگیری تقویتی پیوسته در مسائل مکانی توزیع یافته (مورد مطالعاتی: کنترل هوشمند چراغ های راهنمایی)

محمد اصلانی[۱] محمد سعدی مسگری[۲]

۱- دانشکده نقشه برداری- دانشگاه صنعتی خواجه نصیرالدین طوسی- تهران- ایران

maslani@mail.kntu.ac.ir

۲- دانشیار- دانشکده نقشه برداری- دانشگاه صنعتی خواجه نصیرالدین طوسی- تهران - ایران

mesgari@kntu.ac.ir

چکیده: سامانه های چند عامله به عنوان شاخه ای از هوش مصنوعی در سال های اخیر به عنوان یک نگرش برای مطالعه، بررسی و تحلیل پدیده هایی که دارای خصوصیاتی همچون توزیع یافتگی، پیچیدگی، پایین به بالا بودن و پویایی هستند در زمینه های مختلف از جمله ترافیک، حمل و نقل، اقتصاد، محیط زیست و مواردی از این دست به طور گسترده بکار گرفته شده اند. چالش اصلی در سامانه های چند عامله بدست آوردن رفتار مناسب برای تک تک عامل ها برای رسیدن به رفتار سطح بالای بهینه برای کل سامانه است. یادگیری تقویتی به عنوان رویکردی مناسب که به صورت خودکار و تدریجی می تواند رفتار بهینه را برای تمام عامل ها در تعامل با محیط بدست آورد،برای حل این چالش مناسب است. در یادگیری تقویتی عامل ها در طول زمان از طریق تعامل با محیط یاد میگیرند که در شرایط (حالات) مختلف چه اعمالی را انجام دهند که منجر به دریافت بیشترین سود شود. روش های رایج یادگیری تقویتی در مسائل دنیای واقعی که دارای تعداد حالات محیط بسیار بالا یا بی نهایت هستند عملکرد مناسبی ندارند زیرا این روش ها مقداری مجزا را برای ارزش هر زوج حالت-عمل در حافظه اختصاص می دهند وعامل برای بدست آوردن مقدار دقیق ارزش هر زوج حالت-عمل نیاز دارد که به دفعات ارزش آنها را مشاهده نماید. نوآوری تحقیق حاضر،حل چالش فوق از طریق یادگیری تقویتی پیوسته در مسائل مکانی با فضای حالت-عمل بزرگ و پیوسته است. در رویکرد یادگیری تقویتی پیوسته از مفهوم تعمیم برای تخمین ارزش حالت-عمل استفاده می شود. در این روش عامل نیازی به تجربه اندوزی مستقیم در تمام حالات محیط را ندارد و ارزش یک حالت با شباهت سنجی از ارزش سایر حالات مشابه، تخمین زده می شود. این روش ها برای شباهت سنجی نیاز به کد گذاری حالات محیط دارند که در این تحقیق ناحیه بندی فضا که دارای حجم محاسباتی پایینی است مورد استفاده قرار گرفت. در این تحقیق کنترل ترافیک (به طور خاص مدیریت چراغ های راهنمایی) که دارای پویایی و پیچیدگی بالایی است به عنوان مورد مطالعاتی مطلوب انتخاب شد.

واژه‌های کلیدی:سامانه های چند عامله، یادگیری تقویتی پیوسته، ناحیه بندی فضا و کنترل ترافیک.

تاریخ ارسال مقاله: ۱۳۹۵/۰۱/۱۰

تاریخ پذیرش مشروط مقاله:۱۳۹۵/۰۴/۲۳

تاریخ پذیرش مقاله : ۱۳۹۵/۰۸/۱۶

نام نویسنده‌ی مسئول: محمد اصلانی

نشانی نویسنده‌ی مسئول : ایران- تهران- خیابان ولی‌عصر- تقاطع میرداماد- دانشکده مهندسی عمران و نقشه برداری

۱- مقدمه

در چند دهه گذشته سامانه های اطلاعات مکانی با بکارگیری رویکردها، الگوریتم ها و فناوریهای جدید از سایر علوم مخصوصاً هوش مصنوعی از یک سامانه ساده برای ذخیره سازی، بازیابی و نمایش اطلاعات مکانی به سامانه ای پیچیده برای تجزیه و تحلیل های مکانی تبدیل شده است[۱, ۲]. از سوی دیگر به دلیل پیچیدگی و پویایی بالای مسائل مکانی و طبیعت پایین به بالای اغلب آنها، مفاهیم سامانه های چند عامله در علوم مکانی رشد روز افزونی داشته است[۳-۵]. سامانه های چند عامله روشی برای بررسی و تحلیل پدیده هایی که دارای خاصیت توزیع یافتگی هستند ارائه می دهند. یک سامانه چند عامله از گروهی از عامل های خودمختار و در حال تعامل با یکدیگر تشکیل شده که در یک محیط مشترک برای رسیدن به اهداف خود در حال فعالیت و تعامل با یکدیگر هستند. هر یک از عامل ها محیط اطراف را از طریق سنسورهایشاندرک کرده و از طریق عملگرهای خود عملی را در محیط انجام می دهند[۶, ۷]. در اغلب مسائل مکانی عامل ها با محیط های نسبتاً پیچیده و پویا سروکار دارند، بنابراین عامل ها باید تواناییتصمیم گیری بهینه و خودمختار را داشته باشند. در واقع عامل ها باید بتوانند بدون کنترل خارجی، در اثر تعامل با محیط، دانش لازمه را بدست آورده و رفتار بهینه و عقلایی از خود بروز دهند. قرار دادن دانش کامل و رفتار بهینه در ذهن عامل از ابتدا توسط طراح به دلیل پیچیدگی بالا و ناایستا بودن محیط امری دشوار می باشد. یادگیری تقویتی به عنوان رویکردی که در آن عامل از طریق تعامل با محیط اطرافش در طول زمان رفتار بهینه را می آموزد می تواند برای حل این چالش مفید واقع شود[۸]. اساس یادگیری تقویتی بر این اصل استوار است که اگر عمل انجام شده منجر به بهبود عملکرد شود میل به انجام آن عمل تقویت می گردد[۹]. در یادگیری تقویتی در هر گام زمانی عامل حالت فعلی محیط را درک کرده و براساس دانش کسب کرده از تکرارهای قبلی،عملی را براساس سیاست خود انتخاب و اجرا می کند. عمل انجام شده باعث می شود که حالت محیط تغییر و عامل یک مقدار اسکالر به عنوان پاداش که نشان دهنده میزان خوب بودن عمل انجام شده است دریافت نماید. هدف عامل در طول زندگی بیشینه نمودن مجموع مقادیر پاداش های دریافتی از محیط است.

روش های یادگیری تقویتی رایجبرای مسائلی مناسب هستند که فضای حالات-اعمال آن ها گسسته و کوچک باشد زیرا این روش ها برای هر زوج حالت-عمل یک مقدار مجزا را تخمین زده و ذخیره می کنند و این در حالی است کهدر اغلب مسائل مکانی فضای حالت محیط بسیار بزرگ (یا بی نهایت) می باشد و این باعث عدم کارایی روش های رایج یادگیری تقویتی می شود.

نوآوری تحقیق حاضر، ارائه روشی برپایه یادگیری تقویتی پیوسته برای حل این چالش در مسائل مکانی با فضای حالت-عمل بزرگ و پیوسته است. در یادگیری تقویتی پیوسته از مفهوم تعمیم[۱] استفاده می شود. ایده مورد استفاده بدین صورت است که عامل برای تقریب زدن ارزش تمام حالات-اعمال نیازی به تجربه اندوزی مستقیم ندارد و ارزش یک حالت-عمل را از روی شباهت سنجی با سایر حالات-اعمال تخمین می زند. بدین معنی که هرچه دو حالت از محیط به هم شبیه تر باشند ارزش آن ها نیز به هم نزدیک تر خواهد بود. به عبارت دیگر پیوستگی در حالت ها منجر به ایجاد شباهت در آن ها و نتیجتاً شباهت در ارزش شان می شود. در این تحقیق برای شباهت سنجی از ناحیه بندی فضای حالت[۲] که دارای حجم محاسباتی پایینی است استفاده شد [۸]. از دیگر مزیت های یادگیری تقویتی پیوسته می توان به توانایی بالا بردن ابعاد فضای حالت بی آنکه تعداد حالات محیط زیاد شود اشاره نمود. این ویژگی امکان در نظر گرفتن متغیرهای تصمیم گیری بیشتری را فراهم می آورد. تمرکز مقاله حاضر بر روی یادگیری عامل ها در محیط هاییبافضای حالات بزرگ می باشد.

در نوشتار حاضر، کوشش شده است روش ارائه شده در یک کاربرد نمونه ارزیابی شود. بدین منظور، مسئله کنترل ترافیک و به طور خاص مدیریت چراغ های راهنمایی که دارای خاصیت توزیع یافتگی، پویایی و پیچیدگی بالایی است در نظر گرفته شد. در این کاربرد خودروها و چراغ های راهنمایی به عنوان عامل های خودمختار و هوشمند در نظر گرفتهمی شوند.عامل های چراغ های راهنمایی دارای توانایی یادگیری تقویتی بوده و در طول شبیه سازی رفتار خود را بهبود می دهند و سعی می کنند مدت زمان بهینه سبز را برای هر فاز پیدا نمایند. اما عامل های خودرو که از نوع عامل های واکنشی بوده فاقد توانایی یادگیری می باشند و دارای رفتارهای ثابت و از قبل مشخص شده ای همانند شتاب گرفتن، ترمز کردن و سبقت گرفتن می باشند.کنترل بهینه چراغ های راهنمایی می تواند منجر به کاهش زمان سفر، کاهش میزان آلایندها و مصرف سوخت شود.

۲- پیشینه تحقیق

در زمینه بکارگیری یادگیری تقویتی در سامانه های چند عامله تحقیقات مختلفی در علوم مکانی انجام شده است. در سال ۲۰۰۳ عبدلهای[۳] و کاتان[۴] بعد از معرفی کلیاتی از یادگیری تقویتی، امکان و پتانسیل آن را برای کاربردهای مرتبط با حمل و نقل بررسی نمودند [۱۰]. در سال ۲۰۰۳ عبدلهای و همکاران از یادگیری تقویتی و به طور خاص یادگیری Q[۵] برای زمانمبندی چراغ راهنمایی یک تقاطع درون شهری استفاده کردند [۱۱]. در سال ۲۰۰۸ ونگ[۶] و همکاران برای شبیه سازی جرم از سامانه های چند عامله و یادگیری تقویتی استفاده نمودند. این تحقیق به صورت کاملاً ابتدایی از یادگیری تقویتی فقط جهت مسیریابی عامل ها استفاده کرده است[۱۲]. در سال ۲۰۰۸ تانگ[۷] به توسعه بسته نرم افزاری که شامل تلفیق سامانه های چند عامله، سامانه اطلاعات مکانی و یادگیری ماشین است پرداخت. بسته نرم افزاری ایشان دارای ۴ بخش می باشد: ۱- واحد پردازش مکانی[۸] ۲- واحد یادگیری ۳-واحد نمایش ۴- واحد شبیه سازی. واحد یادگیری شامل الگوریتم هایی نظیر شبکه عصبی، SARSA[۹]،

یادگیری Q، یادگیری Hebbian، مورچگان و تکاملی می باشد. او بسته نرم افزاری توسعه داده شده را برای دو کاربرد، مدلسازی مهاجرت گوزن و مدلسازی حرکت عابران، بکار برد [۱۳]. در سال ۲۰۰۹، بون[۱۰] و دراگیچویچ[۱۱] در مقاله ای از سامانه های چند عامله و یادگیری تقویتی برای حل یک مسئله تصمیم گیری چند هدفه در مدیریت جنگل های ایالت بریتیش کلمبیا[۱۲] کانادا استفاده نمودند. در این تحقیق سه عامل همسان که دارای اهداف یکسانی هستند بکار گرفته شدند. هر یک از عامل ها نماینده یک شرکت تولید الوار هستند که سه هدف اساسی را دنبال می کنند: ۱- افزایش سود اقتصادی ۲- کاهش سن متوسط درختان قطع شده ۳- کاهش مساحت کل ناحیه های برداشت شده. کل محدوده مطالعه به سه قسمت تقسیم شده است و هر عامل در یک قسمت (واحد مدیریت جنگل[۱۳]) به فعالیت می پردازد. هر واحد مدیریت جنگل از تعدادی واحد جنگلی[۱۴] تشکیل شده است که درختان آن از یک سن و گونه می باشند. در هر گام زمانی عامل ها در محدوده متعلق به خود یک واحد جنگلی را انتخاب می کنند و درختان آنرا قطع کرده و تبدیل به الوار می کنند. بنابراین اعمال عامل ها شامل انتخاب یک واحد جنگلی از میان کل واحد های جنگلی است. در پایان هر دوره[۱۵] (هر ۱۰ گام زمانی یک دوره تلقی شده است) واحد های جنگلی برداشت شده مورد بررسی قرار گرفته و با توجه به اهداف تعریف شده، به هر واحد جنگلی یک عدد به عنوان پاداش نسبت داده می شود. در نهایت کیفیت هر واحد جنگلی با توجه به پاداش های مثبت و منفی مورد محاسبه قرار می گیرد. کیفیت هر واحد جنگلی در ابتدا صفر می باشد که نشان دهنده عدم دانش اولیه عامل از محیط است و در طول زمان مقادیر آن ها به روز می شوند. بنابراین در اپیزودهای ابتدایی که دانش اندکی از محیط وجود دارد، عامل ها برای پیدا نمودن اعمال بهینه بیشتر به کاوش در محیط می پردازند و در اپیزودهای پایانی عامل ها بیشتر به استفاده از دانش کسب شده از مراحل قبلی می پردازند. ضمناً در این تحقیق تاثیر هجوم حشرات به جنگل نیز به عنوان یک فاکتور آشوب لحاظ شده است [۱۴]. بون و دراگیچویچ در سال ۲۰۰۹ در مقاله ای دیگر، سامانه های چند عامله و یادگیری تقویتی را برای مدیریت جنگل برای داده های شبیه سازی شده استفاده کردند. ایده اصلی این مقاله مشابه مقاله قبلی است با این تفاوت که محیط عامل بر خلاف مقاله قبلی به صورت سلولی است و عامل ها دارای دو هدف اقتصادی و اکولوژیکی هستند. در این مقاله به بررسی تاثیر تعداد سلول ها (اعمال عامل ها) و تعداد عامل ها و نرخ کنکاش[۱۶] عامل ها بر روی سرعت همگرایی پرداخته شده است. نتایج کار ایشان نشان می دهد که هرچه نرخ کنکاش آهسته تر کاهش بیابد سرعت همگرایی کمتر شده اما احتمال رسیدن به عمل بهینه بالاتر می رود [۱۵]. بون و دراگیچویچ در سال ۲۰۱۰ در مقاله ای دیگر باز هم در زمینه مدیریت جنگل، از یادگیری تقویتی و سامانه های چند عامله استفاده نمودند. تفاوت های این تحقیق با دو تحقیق فوق در این است که اولاً در هر گام زمانی هر عامل می تواند چندین واحد جنگلی را انتخاب نماید، بنابراین مقیاس زمانی با انعطاف بیشتری لحاظ شده است که البته این موضوع یکی از قابلیت های یادگیری تقویتی است. ثانیاً قیود مکانی به عنوان فاکتور محدود کننده اعمال عامل ها، لحاظ شده است. بدین معنی که هر عامل فقط می تواند واحدهای جنگلی را انتخاب نماید که فاصله آنها از محل جاری عامل از حداکثر فاصله تعریف شده بیشتر نباشد و در تابع پاداش نیز هرچه فاصله دو واحد جنگلی متوالی بیشتر باشد پاداش کمتری به عامل داده می شود. ضمناً فاکتور محیط زیست به صورت غیر مستقیم لحاظ شده است، به این صورت که عامل ها امکان قطع درختان به صورت خوشه ای را ندارند. در این تحقیق روش یادگیری تقویتی گسسته SARSA استفاده شده است [۱۶]. چوی[۱۷] و همکاران در سال ۲۰۰۳ یک ساختار چند عامله را برای کنترل ترافیک ارائه دادند که در پایین ترین سطح هر عامل کنترل یک تقاطع را بر عهده دارد. در سطح میانی، یک عامل چند کنترلر مربوط به تقاطع های درون یک منطقه را هماهنگ می کند. در نهایت، در لایه آخر یک عامل مرکزی بر فعالیت همه سامانه نظارت می کند. در تحقیق ایشان از روش فازی عصبی برای یادگیری استفاده شده است [۱۷]. مدینا[۱۸] و همکاران در سال ۲۰۱۰ از یادگیری تقویتی برای کنترل چراغ های راهنمایی استفاده کردند. ایشان برای ایجاد همکاری بین عامل ها در زمان یادگیری از تعداد خودروهای خارج شده از تقاطع به سمت تقاطع های مجاور استفاده کردند. به این ترتیب عامل علاوه بر در نظر گرفتن تعداد خودروهای موجود در مسیرهای ورودی به تقاطع، تعداد خودروهایی که در مسیرهای خروجی از تقاطع قرار دارند را نیز در فضای حالت لحاظ می کند. با این تعریف، هر عامل در طول یادگیری علاوه بر توجه به وضعیت تقاطع خود، وضعیت تقاطع های مجاور را در نظر گرفته و رویکردی جامع را در یادگیری خود لحاظ می کند [۱۸]. گسسته سازی اعمال و سرعت پایین یادگیری از نقاط ضعف این تحقیق می باشد.

جیکوب[۱۹] و عبدلهای در سال ۲۰۱۰، با استفاده از روش یادگیری Q، سیگنال های موجود در یک شبکه آزاد راه-خیابان شهر تورنتو را کنترل کردند. ادوات کنترلی در این مقاله یک کنترل کننده رمپ در ورودی خیابان به بزرگراه و یک VMS در ابتدای بزرگراه می باشد. وضعیت محیط توسط متغیرهایی نظیر میانگین سرعت آزاد راه و خیابان که به فاصله های گسسته و متساوی قسمت بندی شده اند، وضعیت پیام VMS، وجود یا عدم وجود تصادف در آزاد راه و خیابان، شدت تصادف در صورت وقوع و نیز زمان بندی کنترل کننده ورودی (که می توان عمل انجام گرفته توسط عامل در مرحله قبل است)، تعیین می گردد. همچنین برای بهتر شدن فرآیند یادگیری و تعمیم مقادیر ارزش هر وضعیت به وضعیت های مجاور، از روش CMAC[۲۰] استفاده شده است. با این کار تا حدودی ضعف روش های گسسته در عدم توانایی تعمیم دهی جبران گردیده است. در این مقاله برای ارزیابی چگونگی عملکرد کنترل کننده پیشنهادی، از شبیه ساز

Paramics برای شبیه سازی یک ساعت اوج ترافیک صبحگاهی استفاده شد [۱۹].

هولی[۲۱] و همکاران در سال ۲۰۱۰ از یادگیری تقویتی چند هدفه برای کنترل چندین چراغ راهنمایی استفاده کردند. اهداف بهینه سازی شامل تعداد توقف های اتومبیل ها، متوسط زمان توقف و حداکثر طول صف اتومبیل ها در هر تقاطع است [۲۰]. سرعت پایین یادگیری، بکارگیری روش های بسیار ابتدایی یادگیری تقویتی از نکات قابل تعمق در تحقیقات فوق می باشد. همچنین هدف این مقالات صرفاً نشان دادن عملیاتی بودن روش یادگیری تقویتی در مسائل مکانی است. در این تحقیق برای حل مشکل سرعت همگرایی در مسائل با فضای حالت بزرگ از یادگیری تقویتی پیوسته استفاده شده است.

۳- بیان مسئله

پیچیدگی های بالا در محیط های مکانی که نشات گرفته از تعاملات بالا میان موجودیت های مکانی است، انجام تصمیم گیری های بهینه و خودمختار توسط عامل ها را دشوار می نماید. زیرا تصمیم گیری بهینه نیاز به دانش کافی از محیط اطراف دارد. قرار دادن دانش لازم و رفتار بهینه در ذهن عامل ها از ابتدا توسط طراح به دلیل پیچیدگی های بالا و ناایستا بودن محیط امری دشوار می باشد.یادگیری تقویتی به عنوان رویکردی که در آن عامل از طریق تعامل با محیط اطرافش در طول زمان رفتار بهینه را می آموزد می تواند برای حل این چالش مفید واقع شود. در روش های رایج یادگیری تقویتی عامل ها نیاز به تجربه اندوزی در تمامی حالات محیط را دارند. در صورتی که تعداد حالات محیط بالا و یا پیوسته باشند عامل ها نیاز به زمان بیشتری برای تجربه اندوزی و یادگیری دارند که این موضوع می تواند منجر به کاهش سرعت یادگیری و عملکرد عامل ها شود. مسائل مکانی نظیر کنترل ترافیک، مدیریت بحران و برنامه ریزی شهری دارای فضای حالت بزرگ و پیوسته هستند که عملکرد روش های رایج یادگیری تقویتی را با مشکل مواجه می کنند. مسئله اساسی در این تحقیق حل چالش بزرگ بودن و پیوسته بودن فضای حالت در مسائل مکانی است.

۴- یادگیری تقویتی پیوسته

نگرش های موجود در یادگیری تقویتی به دو دسته گسسته و پیوسته تقسیم می شوند. در نگرش های گسسته فضای حالات و اعمال گسسته هستند و تعداد تلاش های لازم برای یادگیری تابعی از تعداد حالات و اعمال می باشد (رابطه ۱).

$$\text{تعداد تلاش های لازم برای یادگیری} = f(|S|, |A|) \qquad (۱)$$

در رابطه ۱، $|S|$ بیانگر تعداد حالات و $|A|$ بیانگر تعداد اعمال است. افزایش تعداد حالات و اعمال، تعداد تلاش های مورد نیاز یادگیری را افزایش می دهد که این افزایش در روش های مونت کارلو به صورت نمایی و در روش های $TD(\lambda)$ به صورت پلی نومیال می باشد[۲۱]. در کل، روش های یادگیری تقویتی گسسته برای مسائل با تعداد حالات بالا دارای سرعت همگرایی پایینی می باشند[۲۲]. در مواجه با مسائلی که فضای حالات آن ها پیوسته و یا بسیار بزرگ باشد استفاده از نگرش یادگیری تقویتی پیوسته معقول تر به نظر می رسد. یادگیری تقویتی پیوسته از مفهوم شباهت سنجی استفاده می کند. بدین معنی که هرچه دو حالت از محیط به هم شبیه تر باشند ارزش آن ها نیز به یکدیگر نزدیک تر خواهد بود. به عبارت دیگر پیوستگی در حالت ها منجر به ایجاد شباهت در آن ها و نتیجتاً شباهت در ارزش شان خواهد شد که از این ویژگی مثبت در یادگیری تقویتی گسسته نمی توان استفاده نمود. در یادگیری تقویتی پیوسته عامل برای بدست آوردن ارزش حالت های مختلف محیط نیازی به تجربه اندوزی مستقیم در آن ها ندارد و می تواند ارزش یک حالت را از ارزش حالت های مشابه آن بدست آورد که این موضوع باعث افزایش سرعت همگرایی می شود. باید اشاره شود که اصول یادگیری تقویتی پیوسته درمقایسه با یادگیری تقویتی گسسته به نحوه یادگیری انسان ها نیز نزدیک تر است.

۴-۱- تابع هزینه

در یادگیری تقویتی پیوسته از مفهوم تعمیم استفاده می شودو تابعی مانند f که دارای پارامترهای مختلف تنظیمی θ است $(\hat{V}^{\pi}(S) =$ $f(S,\theta))$ بر زوج داده های $(S,\ V^{\pi}(S))$ برازش داده می شود. میزان نزدیکی $\hat{V}^{\pi}(S)$ به $V^{\pi}(S)$ نشان دهنده میزان دقت برازش می باشد. به عبارت بهتر هرچه تابع f دقیق تر انتخاب شود و پارامترهای θ به مقدار بهینه θ^* نزدیک تر باشند مقدار تابع هزینه (رابطه ۲) کاهش می یابد[۲۳].

$$C = E\left[\left(V^{\pi}(S) - \hat{V}^{\pi}(S)\right)^2\right] = E\left[\left(V^{\pi}(S) - f(S,\theta)\right)^2\right], \quad \theta^* = \operatorname{argmin}_{\theta} C \qquad (۲)$$

در رابطه ۲، E امید ریاضی، $V^{\pi}(S)$ ارزش واقعی حالت S تحت سیاست π، θ پارامتر تنظیمی، $\hat{V}^{\pi}(S)$ ارزش تخمین زده شده حالت S تحت سیاست πو C نیز تابع هزینه می باشد که باید بهینه شود. به دلیل اینکه در یادگیری تقویتی با یک مسئله یادگیری آنلاین مواجه هستیم نیاز است که در زمان محدودی مقدار θ^* بدست آید. در روش های گرادیان مبنا در صورتی که C نسبت به θ درجه دوم باشد جواب بهینه کلی θ^* بدون قرار گرفتن در بهینه های محلی قابل محاسبه است. بر طبق رابطه ۲، C نسبت به θ درجه دوم است اگر $f(S,\theta)$ نسبت به θ خطی باشد. اما در عمل لزومی ندارد که ارتباط S و $V^{\pi}(S)$ از نوع خطی باشد، بنابراین بجای استفاده مستقیم از S از مقدار کد شده آن، $\varphi(S)$ استفاده می شود (رابطه ۳)[۲۱].

$$f(S,\theta) = \theta^T.\varphi(S) \qquad (۳)$$

در رابطه ۳، $\varphi(S)$ کد کننده ارتباط شباهت در حالت ها با شباهت در ارزش آنها است و θ نیز پارامتر تنظیمی می باشد. مزیت اینکار انتقال پیچیدگی های موجود در تابع f به $\varphi(S)$ است. همچنین تعیین φ بهینه نسبت به f نیاز به دانش کمتری از مسئله دارد. این کدگذاری

بیان کننده معیار شباهت میان حالت های مختلف محیط می باشد. معیار شباهت باید بگونه ای تعریف شود که شباهت میان دو حالت از محیط منجر به شباهت میان ارزش حالت ها شود. تعریف معیار شباهت مستلزم داشتن یک دانش سطح بالا از مسئله است که دانستن این دانش سطح بالا از محیط عامل مستلزم شناخت جامع از مسئله است. در این تحقیق فرض شده است که عامل هیچ گونه دانش سطح بالایی از محیط پیرامون خود ندارد بلکه شباهت سنجی فقط در فضای حالت-عمل براساس فاصله اقلیدسی انجام می پذیرد. شباهت سنجی و کدگذاری فضای حالات به روش های مختلفی از جمله ناحیه بندی فضا[۲۲]، توابع پایه شعاعی[۲۳] و کانروا[۲۴] انجام می شود، که در این تحقیق از ناحیه بندی فضا که دارای حجم محاسبات پایینی است استفاده شده است [۲۱].

از آنجائیکه عامل در طول زندگی خود نمی تواند $V^{\pi}(S)$ را مشاهده نماید و فقط قادر به مشاهده پاداش های لحظه ای می باشد، تابع هزینه(C) به فرم رابطه ۴ بازنویسی می شود.

$$C = E\left[\left(r + \gamma\, \theta^{T}.\varphi(S') - \theta^{T}.\varphi(S)\right)^{2}\right], 0 < \gamma < 1 \qquad (۴)$$

در رابطه ۴، r پاداش لحظه ای، $\varphi(S)$ کد کننده ارتباط شباهت در حالت ها با شباهت در ارزش آنها، θ پارامتر تنظیمی، S و S' به ترتیب حالت قبلی و فعلی محیط می باشند. همچنین γ نرخ تخفیف است که هرچه میزان آن به ۱ نزدیک تر باشد عامل آینده نگرتر و هرچه به صفر نزدیک تر باشد عامل نزدیک بین تر است.

۴-۲- ناحیه‌بندی فضا

ناحیه بندی فضا یکی از روش های کدگذاری فضای حالت می باشد. در این روش فضای حالت توسط مجموعه ای از شبکه ها به ناحیه های کوچکتری تقسیم می شود که به هر شبکه یک Tiling و به هر ناحیه داخل شبکه یک Tile (Partition) گفته می شود (شکل ۱). هر ناحیه دارای یک مقدار باینری است که نشان می دهد که آیا حالت فعلی محیط در آن ناحیه قرار گرفته است یا خیر. در صورتی که از یک Tiling استفاده شود، در هر حالت محیط فقط ارزش یک ناحیه به روز می شود اما در صورت استفاده از چندین شبکه (همانند شکل ۱)، در هر حالت محیط ارزش چندین ناحیه به روز می شود و همین موضوع سرعت همگرایی را افزایش می دهد. لازم به ذکر است که شکل و ابعاد هر ناحیه و تعداد Tiling ها در روش ناحیه بندی فضا بر روی قدرت تعمیم یادگیری تاثیر گذار هستند. در این تحقیق تعداد Tiling ها در سناریوهای مختلف برابر ۱، ۳، ۶ و ۹ در نظر گرفته شدند.

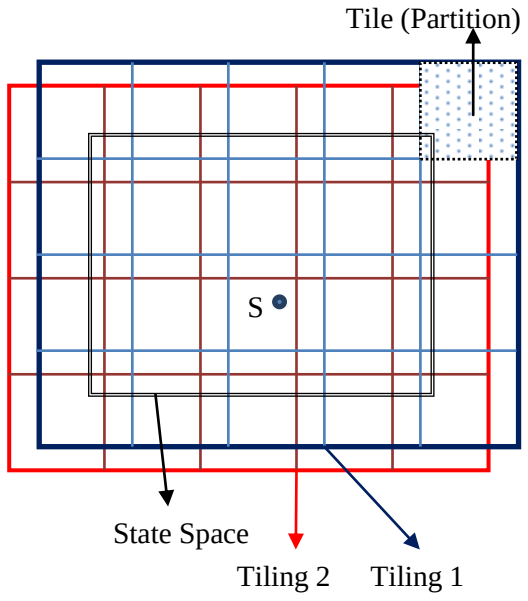
۴-۳- تخمین تابع ارزش حالت-عمل

ارزش هر زوج حالت-عمل تحت سیاست π توسط رابطه ۵ تخمین زده می شود که گسترش یافته رابطه ۳ است.

$$Q^{\pi}(S, a_i) = \theta^{T}_{a_i \epsilon A}.\varphi(S) \qquad (۵)$$

در رابطه ۵، $\theta_{a_i \epsilon A}$ بردار پارامترها برای عمل a_i و $\varphi(S): X \to R^{d}$ تصویر فضای حالت به یک بردار d بعدی است (کد گذاری فضای حالت). با تلفیق روابط ۲ و ۵ تابع هزینهبه صورت رابطه ۶ باز نویسی می شود.

$$C = E\left[\left(r + \gamma\, \max_{a'} \theta^{T}_{a'}.\varphi(S') - \theta^{T}_{a_i \epsilon A}.\varphi(S)\right)^{2}\right] \qquad (۶)$$



شکل (۱): ناحیه بندی فضای حالت دو بعدی با استفاده از دو شبکه

در رابطه ۶، θ_a بردار پارامترها برای عمل a_i، r پاداش لحظه ای، S و S' به ترتیب حالت قبلی و فعلی محیط هستند. همچنین $\varphi(S)$ تصویر فضای حالت به یک بردار d بعدی است.با اعمال روش گرادیان نزولی بر رابطه ۶ جهت بهینه سازی رابطه ۷ بدست می آید. درواق رابطه ۷ نحوه به روزرسانی $\theta_{a_i \epsilon A}$ با استفاده از روش گرادیان نزولی را نشان می دهد. در رابطه ۷ α نرخ یادگیری است.

$$\theta^{t+1}_{a_i \epsilon A} = \theta^{t}_{a_i \epsilon A} - \frac{\alpha}{2}\nabla_{\theta_{a_i}} C \qquad (۷)$$

به دلیل وجود امید ریاضی در تابع هزینه، برای محاسبه $\nabla_{\theta_{a_i}} C$ گرادیان احتمالاتی[۲۵] مورد استفاده قرار می گیرد. در گرادیان احتمالاتی بجای گرادیان گرفتن از امید ریاضی یک عبارت، از نمونه های آن گرادیان گرفته می شود و در صورتی که محیط MDP و ergodic و $0 < \alpha < 1$، $\sum \alpha = \infty$ و $\sum \alpha^2 < \infty$ باشد روش گرادیان احتمالاتی به همگرایی خواهد رسید. نحوه محاسبه گرادیان احتمالاتی در رابطه ۸ آورده شده است.

$$\frac{1}{2}\nabla_{\theta_{a_i}} C = -[r + \gamma\, \max_{a'} \theta^{T}_{a'}.\varphi(S') - \theta^{T}_{a_i \epsilon A}.\varphi(S)]\varphi(S) \qquad (۸)$$

از تلفیق روابط ۸ و ۷، رابطه ۹ بدست خواهد آمد.

$$\theta_{a_i \epsilon A}^{t+1} = \theta_{a_i \epsilon A}^{t} - \frac{\alpha}{2}\nabla_{\theta_{a_i}} C =$$
$$\theta_{a_i \epsilon A}^{t} + \alpha[r + \gamma \max_{a'} \theta_{a'}^{T}.\varphi(S') - \theta_{a_i \epsilon A}^{T}.\varphi(S)]\varphi(S) \quad (۹)$$

برای افزایش سرعت همگرایی در این تحقیق از روش اثر شایستگی برای به روز رسانی مقادیر $\theta_{a_i \epsilon A}$ استفاده شد (رابطه ۱۰) [۲۴].

$$Z^0 = 0$$
$$Z^{t+1} = \gamma\lambda Z^t + \varphi(S) \quad (۱۰)$$

کد کننده ارتباط شباهت در $\varphi(S)$ در رابطه ۱۰، γ نرخ تخفیف، حالت ها با شباهت در ارزش آنهاو λ میزان تاثیر پذیری ارزش حالات-اعمال ابتدایی اپیزود از ارزش حالات-اعمال و سیگنال های انتهایی با تلفیق روابط ۹ و ۷ به روز می شوند (رابطه $\theta_{a_i \epsilon A}$اپیزود است. مقادیر ۱۱).

$$\delta^{t+1} = r^t + \gamma \max_{a'} \theta_{a'}^{T}.\varphi(S') - \theta_{a_i \epsilon A}^{T}.\varphi(S)$$
$$\theta_{a_i \epsilon A}^{t+1} = \theta_{a_i \epsilon A}^{t} - \alpha\delta^{t+1}Z^{t+1} \quad (۱۱)$$

در رابطه ۱۱، γ نرخ تخفیف،r^tپاداش دریافتی از محیط در زمان t،α نرخ یادگیری،λ میزان تاثیر پذیری ارزش حالات-اعمال ابتدایی اپیزود از ارزش حالات-اعمال و سیگنال های انتهایی اپیزود است،S و S' به ترتیب حالت قبلی و فعلی محیط هستند و همچنین $\varphi(S)$ کد کننده ارتباط شباهت در حالت ها با شباهت در ارزش آنها است. در این تحقیق از روابط ۱۰ و ۱۱ برای یادگیری چراغ های راهنمایی استفاده شد.

۵-ترافیک و هوشمند سازی چراغ ها

افزایش شهر نشینی و در پی آن گسترش شهرها باعث افزایش تقاضای سفرهای درون شهری در دهه های اخیر شده است. از سوی دیگر در دنیای امروز که دنیای سرعت، پیشرفت و تکنولوژی است روز به روز بر تعداد خودروها افزوده می شود. این افزایش تقاضای سفر و افزایش تعداد خودروها منجر به ظهور پدیده ترافیک در شهرها شده است. حجم ترافیک در سال های اخیر افزایش چشمگیری داشته است و پیش بینی ها حاکی از آن است که این حجم در سال های ۲۰۰۰ تا ۲۰۲۰، ۵۰٪ افزایش خواهد داشت [۲۵]. بنابراین بررسی دلایل بروز ترافیک و اثرات آن ضروری به نظر می رسد. در ساده ترین حالت، ترافیک زمانی ایجاد می شود که حجم ترافیک ورودی بیشتر از ظرفیت معابر شود. مواردی نظیر تصادفات، انسداد راه ها و شرایط جوی سبب کاهش ظرفیت راه ها می شود.

از پیامدهای ترافیک می توان به افزایش تولید و انتشار گازهای آلاینده، از بین رفتن زمین های کشاورزی، دگرگون شدن آب و هوای کره زمین، باران های اسیدی، آلودگی هوای مناطق شهری و تهدید سلامت شهرنشینان اشاره نمود. طبق آمار منتشر شده از ترازنامه انرژی کشور در سال ۱۳۸۸، ۹۷ درصد منواکسید کربن و ۷۹ درصد هیدروکربن های منتشر شده در کشور مربوط به بخش حمل و نقل است [۲۶]. به طور خاص شهرهای ایران به دلیل رشد سریع جمعیت، ناوگان فرسوده خودروها، تعداد زیاد واحد های صنعتی، عوامل جغرافیایی و هواشناسی منطقه با کاهش شدید کیفیت هوا رو به رو شده اند که گاهاً سطح بالای آلاینده ها مسئولین را مجبور به تعطیلی مدارس و تحمیل محدودیت های ترافیکی کرده است [۲۷]. یکی دیگر از پیامدهای ترافیک، افزایش مصرف انرژی می باشد که منجر به مشکلات اقتصادی فراوانی برای دولت ها جهت تامین سوخت خودروها می شود.

احداث راه های جدید و افزایش ظرفیت معابر، یکی از روش های کاهش معضل ترافیک شهری است. اما باید در نظر داشت که این راه حل خود هزینه های بسیاری را درپی 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 پیاده سازی چنین راه حلی در مقیاس وسیع نیازمند هزینه و زمان زیادی است[۲۸]. یکی دیگر از شیوه های کاهش ترافیک، استفاده از چراغ های راهنمایی هوشمند در تقاطع ها می باشد. چراغ های راهنمایی رایج به صورت پیش زمانبندی شده عمل می نمایند. بدین معنی کهآنها زمانبندی از پیش تعیین شده و معلومی را بدون توجه به تغییرات شرایط واقعی ترافیکی لحاظ می کنند و این خود باعث افزایش ترافیک در زمان های اوج عبور و مرور می شود. در چراغ های راهنمایی هوشمند پارامترهای زمانبندی چراغ ها با توجه به شرایط ترافیکی موجود (به عنوان مثال، تعداد ماشین های منتظر در تقاطع) برای رسیدن به یک سری اهداف خاص (به عنوان مثال، بیشترین تعداد ماشین های عبوری از تقاطع) بهینه می شوند.

۶- شبیه سازی میکروسکوپیک ترافیک

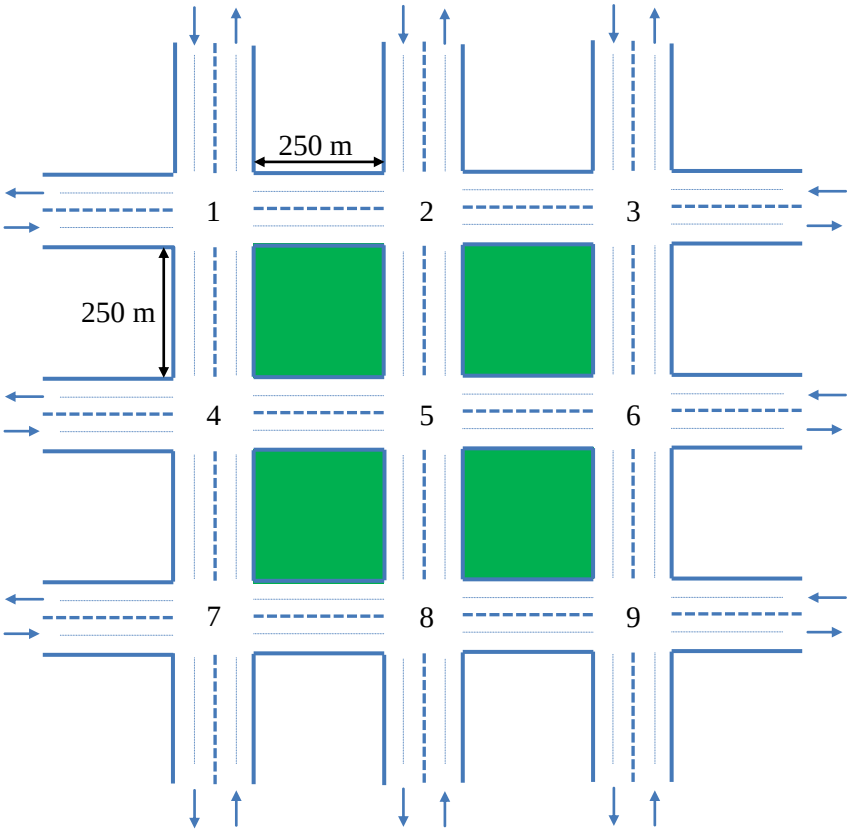
شبیه سازی عبارت است از فرآیند طراحی یک مدل از دید یک سامانه واقعی و کاربرد آن به منظور درک سامانه و یا ارزیابی استراتژی های مختلف برای عملکرد سامانه [۲۹]. در حقیقت، کار شبیه سازی تقلید رفتار یک سامانه واقعی است. امروزه با پیشرفت علم و فناوری در حوزه کامپیوتر، شبیه سازی های ترافیکی به عنوان راه حلی کم هزینه و مطمئن به کمک مهندسین و کارشناسان آمده است. بالا بودن تعداد زیاد داده ها، پیچیدگی بالای شبکه های مورد تحلیل و تعداد بالای رویدادهایی که باید هم زمان در نظر گرفته شوند باعث گسترش استفاده از شبیه سازی های ترافیکی شده است. شبیه سازی های ماکروسکوپیک، میکروسکوپیک و مزوسکوپیک سه نوع از معروفترین رویکردها در شبیه سازی های ترافیکی هستند [۳۰]. در این تحقیق از شبیه سازی میکروسکوپیک که در آن وضعیت تک تک وسایل نقلیه به صورت خرد و جزئی لحاظ می شود استفاده شده است. در این نوع شبیه سازی، رفتار هر یک از رانندگان به تنهایی مدل می شود و در نتیجه عکس العمل و جزئیات رفتاری رانندگان و تعامل آنها با یکدیگر کاملاً در مدلسازی در نظر گرفته می شود. پیش شرط انجام شبیه سازی ترافیکی مشخص نمودن شبکه ترافیک، رفتار راننده ها و رفتار چراغ های راهنمایی است که در ادامه مختصراً تشریح می شوند.

در این تحقیق از سامانه های چند عامله ناهمگن دارای ارتباط غیر مستقیم برای شبیه سازی و کنترل ترافیک استفاده شده است. بدین معنی که عامل ها (خودروها و چراغ های راهنمایی) دارای اهداف و اعمال غیر یکسان هستند. به طور خاص خودروها، عامل هایی هستند که فاقد توان یادگیری می باشند و دارای اعمالی از جمله تغییر مسیر، تغییر سرعت و تغییر خط عبور می باشند. هدف نهایی خودروها رسیدن به مقصد در کوتاهترین زمان ممکنه است. چراغ های راهنمایی، عامل های با قابلیت یادگیری و بهبود رفتار در طول زمان هستند. هدف نهایی چراغ های راهنمایی کاهش حجم ترافیک و عبور هرچه بیشتر خودروها از تقاطع ها می باشد. اعمال چراغ های راهنمایی انتخاب بهترین زمان بندی برای فازهای مختلف با توجه به شرایط ترافیکی است.

معماری تمام عامل ها در این تحقیق از نوع واکنشی است. معماری واکنشی فرآیند تصمیم گیری را به صورت یک نگاشت مستقیم از وضعیت موجود به عملیات، پیاده سازی می کند و بر اساس یک مکانیزم تحریک-پاسخ که عمل تحریک بر اساس داده های حس شده انجام می شود پایه گذاری می گردد. در این معماری هر رفتار نتیجه یک تابع عمل منحصر به فرد است. این تابع ورودی های قابل درک را به عمل تبدیل می کند. به عنوان مثال عامل های خودرو با درک وضعیت ترافیکی موجود مسیر خود را تغییر می دهند و از مسیری عبور می کنند که دارای ترافیک کمتری است. لازم به ذکر است که در این تحقیق محیط شبیه سازی ترافیکی،AIMSUNاست و از زبان برنامه نویسی C++جهت توسعه آن استفاده شده است[۳۱].

۶-۱- شبکه ترافیک

شبکه ترافیکی ۳×۳ مورد بررسی که از ۹ تقاطع و ۲۴ خیابان تشکیل شده است در شکل ۲ نشان داده شده است. در هر تقاطع یک عامل یادگیر چراغ راهنمایی قرار دارد که توسط یادگیری تقویتی در طول زمان و در اثر تعامل با محیط اطراف خود به یادگیری می پردازد. تمام خیابان ها دوطرفه (دو جهته) و دارای دو خط (مسیر) در هر طرف می باشند که ظرفیت هر خط ۴۰ خودرو است. همچنین طول هر خیابان ۲۵۰ متر با حداکثر سرعت مجاز ۵۰ کیلومتر در ساعت است. در طول شبیه سازی فرض می شود ۳۳٪ از ماشین ها مسیر مستقیم، ۳۳٪ گردش به چپ و ۳۳٪ گردش به راست را انجام می دهند. به منظور لحاظ نمودن ناایستا بودن محیط، در زمان های مختلف ۳ جریان ترافیکی مختلف ۲۰۰، ۴۰۰ و ۶۰۰ خودرو در هر ساعت وارد خیابان های ورودی می شوند.



شکل (۲): شبکه ترافیکی متشکل از ۹ تقاطع

۶-۲- شبیه سازی رفتار رانندگان

رانندگی شامل مجموعه ای از مانورها و رفتارهایی است که هر کدام از آنها به راننده و شرایط رانندگی وابسته است. رفتار راننده را می توان در یک ساختار سلسله مراتبی به سه سطح استراتژیک[۲۶]، سطح تاکتیک[۲۷] و سطح کنترل تقسیم بندی نمود [۳۲].

در سطح استراتژیک، اهداف کلی هر راننده و در نتیجه آن یک مسیر برای هر راننده مشخص می شود. در این تحقیق هر راننده از الگوریتم دایجسترا برای انتخاب کوتاه ترین مسیر [۲۴] استفاده می کند.

در پایین ترین سطح که همان سطح کنترل است، اعمال واقعی راننده همانند چرخاندن فرمان، فشردن پدال ها و تعویض دنده ها انجام می شود. راننده دستورات کنترلی را دائماً و در تعامل با دینامیک قسمت های دیگر سامانه همانند خودروی پیشین و ترافیک مسیر پیش رو، وارد می کند. سطح کنترل در شبیه سازی مقاله حاضر نادیده گرفته شده است. در سطح تاکتیکی، برخی مانورها و حرکات همانند تغییر خط حرکت در یک مسیر، سبقت گرفتن، انتخاب سرعت مناسب و گردش به راست یا چپ برپایه دستیابی به اهداف کوتاه مدت انتخاب و اجرا می شود. رفتار تعقیب خودرو[۲۸] و تغییر خط دو مدل حائز اهمیت در سطح تاکتیکی هستند که در این تحقیق بکار رفته اند.

همچنین در نرم افزار شبیه ساز ترافیکی مورد استفاده(AIMSUN) ویژگی های اخلاقی هر راننده توسط تعدادی پارامتر مشخص می شود. مهمترین پارامترها و مقادیر آن ها به شرح ذیل هستند [۳۳، ۳۴]:

- سرعت مطلوب راننده: سرعتی است که راننده تمایل دارد در طول سفر خود با آن حرکت نماید. مقدار این سرعت از یک تابع توزیع گوسین با میانگین $۱۱۰\frac{Km}{h}$ و انحراف از معیار $۱۰\frac{Km}{h}$ انتخاب می شود.
- میزان تبعیت از حداکثر سرعت مجاز خیابان ها: مقدار آن به تصادف برای هر خودرو از یک تابع توزیع گوسین با میانگین ۱,۱ و انحراف از معیار ۰,۱ انتخاب می شود.
- زمان واکنش راننده: عبارت است از مدت زمانی که طول می کشد تا راننده به تغییرات سرعت خودروی جلویی واکنش نشان دهد. این زمان برابر ۱ ثانیه در نظر گرفته شد.
- زمان واکنش در حالت توقف: عبارت است از مدت زمانی که طول می کشد تا خودروی متوقف شده به شتاب گرفتن خودروی جلویی یا تغییرات فاز چراغ راهنمایی مقابلش واکنش نشان دهد. این زمان واکنش فقط برای خودروهایی که متوقف هستند بکارگرفته می شود که مقدار آن برابر ۱,۳۵ ثانیه انتخاب شد.

۶-۲-۱- رفتار تعقیب خودرو

رفتار تعقیب خودرو از اصلی ترین و معمول ترین رفتارهای رانندگی است و تاکنون مدل های زیادی برای این رفتار ارائه شده است. مدل تعقیب خودرو برای نخستین بار توسط Reuschel[۳۵] و سپس Pipe[۳۶] ارائه شد. هم چنین Herman و همکارانش بررسی های زیادی را در این موضوع انجام داده اند که نتایج این بررسی ها اساس مدل های تعقیب خودرو را تشکیل داده اند [۳۷]. بر طبق این رفتار، سرعت خودرو که ناشی از تعامل آن با جاده و کاربران آن است تعیین می شود. در این رفتار فرض می شود که بین خودروهایی که فاصله شان با یکدیگر بین صفر تا حدود ۱۰۰ الی ۱۲۵ متر باشد رابطه ای وجود دارد و این رابطه به شکل روابط صریح ریاضی معرفی می شود. این رفتار حرکت طولی یک خودرو را زمانی که اتومبیل دیگری را دنبال و تلاش می کند فاصله ایمنی خود را با خودروی راهنمای[۲۹] خود نگه دارد، توصیف می کند. ایده اصلی این روش، بر مبنای وابستگی حرکت هر خودرو نسبت به حرکت خودروی پیش رو در ترافیک است. سامانه های مبتنی بر رفتار تعقیب خودرو با کنترل حرکت طولی خودرو و حفظ فاصله مطلوب با خودروی راهنما، از بروز تصادف از جلو جلوگیری می کنند. برای این منظور مدل گیپس استفاده شده است [۳۸]. مدل گیپس از دسته مدل های اجتناب از برخورد است که در سال ۱۹۸۱ معرفی شد. در مدل گیپس، خودروها به دو دسته خودروهای آزاد (خودروهایی که حرکتشان را خودروی جلویی شان محدود نمی کند) و خودروهای مقید (خودروهایی که حرکتشان را خودروهای جلوی شان محدود می کند)، تقسیم می شوند. این مدل چنان عمل می کند که خودروی آزاد همواره سعی در رسیدن به سرعت مطلوب خود دارد. سرعت مطلوب سرعتی است که یک خودرو می تواند با توجه به خصوصیات فنی خود و مقررات حاکم بر قطعه ای از راه که در آن حرکت می کند، در غیاب سایر خودروها به دست آورد. مدل تعقیب خودرو با توجه به پارامترهای محلی از جمله نوع راننده (اینکه راننده تا چه حدی به محدودیت سرعت توجه کند)، هندسه مقطع (محدودیت سرعت مقطع، محدودیت سرعت گردش و غیره)، اثر خودروهای خط مجاور و غیره تعریف می شود. این مدل شامل دو مولفه افزایش سرعت و کاهش سرعت است. افزایش سرعت (رابطه ۱۲) بیانگر قصد خودرو برای رسیدن به سرعت مطلوب است در حالیکه کاهش سرعت (شتاب کاهشی) ناشی از محدودیت هایی است که توسط خودروی جلویی تحمیل می شود (رابطه ۱۳)[۳۸].

$$V_a(n,t+T) = V(n,t) + 2.5a(n)T\left[1-\frac{V(n,t)}{V^*(n)}\right]\sqrt{0.025+\frac{V(n,t)}{V^*(n)}} \quad (۱۲)$$

در رابطه ۱۲، V(n,t) سرعت خودروی n در زمان t، a(n) شتاب خودروی n، V*(n) نیز سرعت مطلوب راننده خودروی n و T نیز زمان واکنش راننده است.

$$V_b(n,t+T) = d(n)T + (d(n)^2T^2 - d(n)[2\{x(n-1,t) - S(n-1) - X(n,t)\} - V(n,t)T - \frac{V(n-1,t)^2}{d'(n-1)}])^{0.5} \quad (۱۳)$$

در رابطه ۱۳، d(n)<0 شتاب کاهشی خودروی n، X(n-1,t) موقعیت خودروی جلویی(n-1) در لحظه t، X(n,t) موقعیت خودروی n در

لحظه t، S(n-1) طول خودروی جلویی، V(n-1,t) سرعتخودروی جلویی در زمان t و d'(n-1) نیز تخمینی از شتاب کاهشی خودروی جلویی(n-1) توسط خودروی عقبی (n) است که توسط d'(n-1)=d(n-1)×α که α، فاکتور حساسیت است تخمین زده می شود. به عبارت بهتر فاکتور حساسیت میزان درستی تخمین شتاب خودروی جلویی توسط خودروی تعقیب کننده را نشان می دهد. توضیحات جامعدر رابطه با رفتار تعقیب خودرو در[۳۱]ذکر شده است.

۶-۲-۲- رفتار تغییر خط

در رفتار تغییر خط، یک خودرو موقعیت خود را از خط مسیر فعلی به مسیر مجاورش تغییر می دهد. برای انجام رفتار تغییر خط، خودرو در عرض جاده حرکت می کند، بنابراین در این رفتار لازم است علاوه بر حرکت طولی، حرکت جانبی خودرو نیز بررسی شود. در سال ۱۹۹۹،Das و همکارانش رفتار تغییر خط را به دو گروه تغییر خط اجباری[۳۰] و تغییر خط اختیاری[۳۱] تقسیم بندی نمودند. رفتار تغییر خط اجباری وقتی رخ می هد که راننده مجبور است خط فعلی خود را ترک کند. به عنوان مثال زمانی که راننده می خواهد از ورودی مسیر[۳۲] وارد بزرگراه شود یا از خروجی بزرگراه[۳۳] خارج گردد، لازم است خط حرکت خود را تغییر دهد. معمولاً رانندگان برای عبور از موانع مسیر، پیروی از قوانین خطوط و پاسخ به علامت ها و پیام های خطوط، تغییر خط اجباری انجام می دهند. در تغییر خط اختیاری، عموماً راننده از شرایط رانندگی در خطوط کنونی خود راضی نیست و به همین علت مسیر حرکت خود را تغییر می دهد. مثلاً رانندگان در مواجه با خودروی آهسته از آن سبقت می گیرند یا برای افزایش سرعت، به خطوط با سرعت بالاتر می روند[۳۴]. تغییر خط اختیاری عموماً زمانی رخ می دهد که سرعت خودروی پیش رو از سرعت مطلوب خودروی اصلی کمتر باشد. در این صورت خودروی اصلی با تغییر خط، شرایط رانندگی خود را بهبود می بخشد.

رفتار تغییر خط شامل چند مرحله است، ابتدا در یک سطح استراتژیک، راننده برای انتخاب مسیر مناسب، خط فعلی خود را شناسایی می کند. این بررسی شامل موانع موجود در خط، ورودی و خروجی به خط و سایر ادغام شدن های اجباری است. در مرحله استراتژیک، راننده با افزایش یا کاهش شتاب برای رفتارهای تغییر خط آماده می شود. در آخر نیز در یک سطح عملیاتی، راننده تعیین می کند که رفتار تغییر خط ایمن و قابل قبول هست یا خیر. این تصمیم با استفاده از مدل های پذیرش فاصله[۳۴] صورت می گیرد که در آنها راننده فاصله موجود را با فاصله مناسب یا فاصله بحرانی[۳۵]مقایسه می کند و در صورتی که این فاصله کمتر از میزان ایمن باشد، راننده سعی می کند فاصله مناسب را ایجاد نماید [۳۳].

اولین سوالی که راننده شبیه سازی شده در فرایند تصمیم گیری جهت اجرای مانور تغییر خط در نظر می گیرد، امکان پذیر بودن و یا غیر ممکن بودن اجرای مانور تغییر خط به خط عبوری هدف است. جهت بررسی امکان پذیر بودن چنین مانوری در گام اول به خودروی تغییر خط دهنده شتابی مساوی با دو برابر شتاب کاهشی که راننده در شرایط نرمال مورد استفاده قرار می دهد، اختصاص داده می شود. اگر خودروی جلویی در خط هدف به عنوان خودروی پیشرو و خودروی تغییر خط دهنده به عنوان خودروپیرو در نظر گرفته شود، براساس روابط۱۲و ۱۳ سرعت خودروی تغییر خط دهنده در گام زمانی بعدی محاسبه می شود. سپس بر اساس سرعت محاسبه شده از طریق مدل تعقیب خودرو، شتاب کاهشی جهت اجتناب از برخورد بر اساس رابطه ۱۴ محاسبه می گردد[۳۱].

$$\widehat{b_n} = \frac{(v_n(t+\tau)-v_n(t))}{\tau} \qquad (۱۴)$$

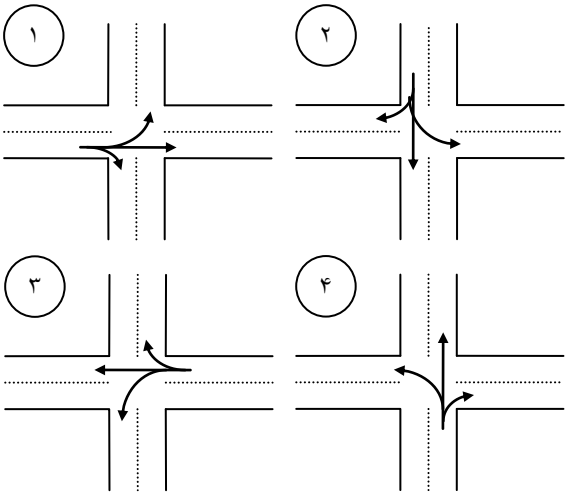
در رابطه ۱۴، $v_n(t+\tau)$ سرعت خودروی تغییر خط دهنده در گام زمانی $t+\tau$ ، $v_n(t)$ سرعت خودروی تغییر خط دهنده در گام زمانی t هستند.τنیز گام زمانی شبیه سازی بر حسب ثانیه می باشد. در ادامه شتاب تخمین زده شده با شتاب کاهشی اختصاص داده شده به خودروی تغییر خط دهنده مقایسه شده و شتاب کاهشی خودروی تغییر خط دهنده به شتاب نرمال از طریق روابط ۱۲ و ۱۳ بازگردانده می شود. سپس شتاب کاهشی محاسبه شده برای خودروی عقبی با شتاب نرمال این خودرو مقایسه و در نهایت در صورت قابل قبول بودن هر دو شتاب، اجرای مانور تغییر خط امکان پذیر می شود. خوانندگان برای توضیحات بیشتر می توانند به [۳۱] مراجعه کنند.

۶-۳- چراغ های راهنمایی هوشمند مبتنی بر یادگیری تقویتی پیوسته

چراغ های راهنمایی مستقر در هر تقاطع به عنوان عامل های هوشمند خودمختار با توانایی یادگیری تقویتی پیوسته در نظر گرفته می شوند. ساختار چراغ های راهنمایی به صورت مرکزی نمی باشد بلکه به صورت توزیع یافته است. بدین معنی که تمامی چراغ های راهنمایی، ترافیک موجود در محیط اطراف خود را حس کرده و براساس آن سعی در کاهش وضعیت ترافیکی محلی خود می نمایند. همچنین معماری چراغ های راهنمایی (عامل ها) واکنشی است و به تغییرات موجود در محیط ترافیکی واکنش نشان می دهند. نقش چراغ های راهنمایی هوشمند، انتخاب بهترین زمان بندی با توجه به شرایط ترافیکی، برای عبور هرچه بیشتر خودروها است.

هر چراغ راهنمایی در ابتدای هر فاز (شکل ۳) تعداد خودروهای منتظر در تمام مسیرهای ورودی به تقاطع را مشاهده نموده (حالت محیط) و براساس دانش خود زمانی را به عنوان زمان سبز برای فاز جاری تقاطع انتخاب می کند. حالت محیط یک بردار متشکل از ۵ المان است که ۴ المان اول آن تعداد خودروهای منتظر در هر مسیر ورودی را نشان می دهد که دارای مقادیر پیوسته از۰ تا ۸۰ خودرو می باشند. المان آخر نیز فاز جاری چراغ راهنمایی است که دارای مقادیر ۱ تا ۴ است. اعمال چراغ راهنمایی مدت زمان سبز بودن هر فاز است که از میان اعداد {۹۰ و ۸۰ و ۷۰ و ۶۰ و ۵۰ و ۴۰ و ۳۰ و ۲۰} ثانیه انتخاب می شوند.

همانطور که مشخص است حالات محیط پیوسته و اعمال چراغ های راهنمایی گسسته می باشند. بعد از اتمام زمان سبز، عامل تعداد ماشین های عبوری از تقاطع را به عنوان سیگنال پاداش (جریمه) از سوی محیط دریافت می کند. ارزش عمل انجام شده با توجه به سیگنال پاداش دریافتی به روز می شود(Q(S,a)). به عبارت بهتر هرچه عمل انجام شده منجر به عبور تعداد ماشین های بیشتری از تقاطع شود ارزش آن عمل نیز بالاتر خواهد رفت. در تکرار های بعدی در صورتی که عامل در حالت های مشابه آن حالت قرار گیرد آن عمل را با احتمال بیشتری انجام خواهد داد. شکل ۴ اجزا مختلف چراغ راهنمایی را به صورت کلی نشان می دهد. همانطور که در شکل ۴ مشخص است تعداد خودروهای موجود در هر مسیر منتهی به تقاطع توسط سنسورهای چراغ راهنمایی مشاهده و حالت محیط و پاداش دریافتی تخمین زده می شود. بر اساس پاداش دریافتی از محیط ارزش اعمال و نهایتاً سیاست رفتاری چراغ راهنمایی به روز رسانی می شود. در انتها نیز چراغ راهنمایی با توجه به سیاست خود یک عمل که نشان دهنده زمان سبز فاز فعلی است انتخاب می نماید. در این تحقیق به منظور ارزیابی کامل از عملکرد چراغ های راهنمایی سه جریان ترافیکی ۲۰۰، ۴۰۰ و ۶۰۰ خودرو در هر ساعت برای تغییر ترافیک اعمال می شود.

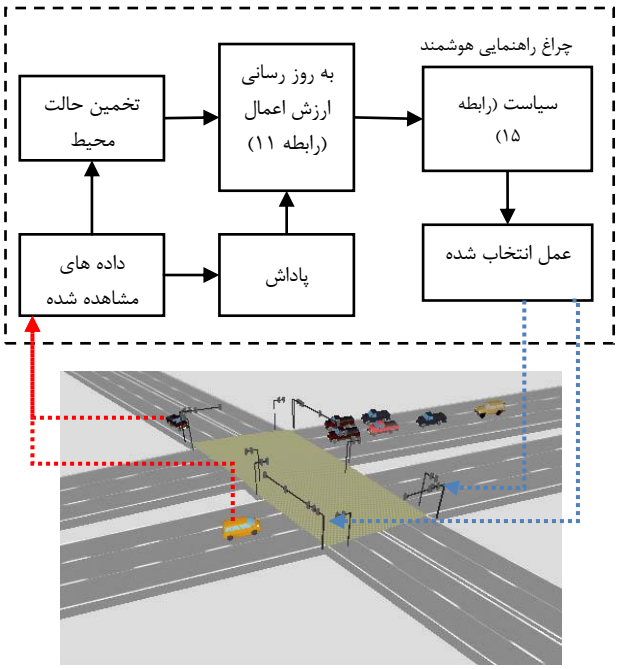


شکل (۳): ترتیب فازها

احتمال انجام اعمال مختلف (سیاست عامل) در هر حالت محیط بر طبق رابطه ۱۵ محاسبه می شود. این رابطه بیان کننده سیاست ε-greedy است [۲۱].

$$P(S,a_i)=\begin{cases}1-\varepsilon+\frac{\varepsilon}{|A_s|}, & \text{if } a_i=\text{argmax}_{a'\in A_s}Q(S,a')\\ \frac{\varepsilon}{|A_s|}, & \text{else}\end{cases} \quad (۱۵)$$

در رابطه ۱۵، $P(S,a_i)$ احتمال انتخاب عمل a_i در حالت محیط S، $|A_s|$ تعداد اعمال ممکن، $Q(S,a')$ ارزش عمل a' در حالت محیط S و ε احتمال انتخاب اعمال غیر بهینه بر اساس دانش فعلی چراغ راهنمایی هستند. در صورتی که مقدار ε برابر صفر انتخاب شود چراغ راهنمایی به صورت حریصانه و در صورتی که مقدار ε برابر یک انتخاب شود چراغ راهنمایی به صورت تصادفی عمل می کند. لازم به ذکر است که نحوه عملکرد چراغ راهنمایی هوشمند به زبان C++ نوشته و به صورت یک فایل DLL به نرم افزار AIMSUN اضافه شده است.



شکل (۴): بخش های مختلف چراغ راهنمایی هوشمند

۷- نتایج

پیاده سازی ها در چهار سناریوی مختلف با تعداد Tile ها و Tiling های مختلف انجام شده است. همانطور که در بخش ۴-۲ نیز اشاره شد در روش ناحیه بندی فضا، به هر شبکه یک Tiling و به هر ناحیه داخل شبکه یک Tile (Partition) گفته می شود. لازم به ذکر است که چراغ های راهنمایی برای به روز رسانی ارزش اعمال از رابطه ۱۱ و انتخاب اعمال از رابطه ۱۵استفاده می کنند. جدول ۱ تعداد Tiling ها و Tile ها را برای سناریوهای مختلف نشان می دهد.

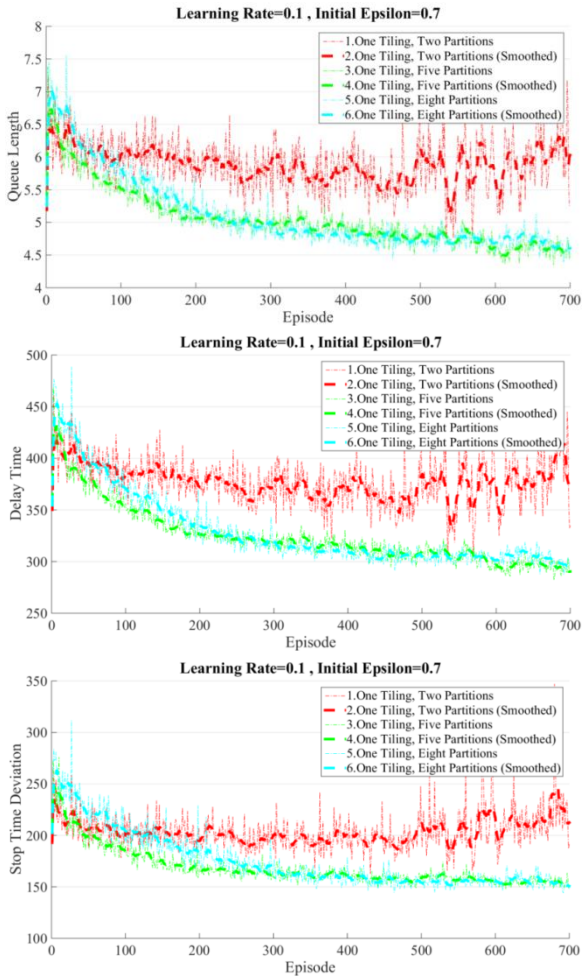
جدول (۱): ویژگی های سناریوهای مختلف

شماره سناریو	Tiling #	Tile (Partition) #
۱	۱	۲و ۵ و ۸
۲	۳	۲و ۵ و ۸
۳	۶	۲و ۵ و ۸
۴	۹	۲و ۵ و ۸

همانطور که از جدول ۱ مشخص است هر سناریوبه ازای تعداد Tile های ۲ و ۵ و ۸ پیاده سازی شده است. ارزیابی عملکرد سناریوهای مختلف براساس شاخص های متوسط طول صف، متوسط زمان تاخیر و انحراف از معیار زمان توقف می باشد.شاخص متوسط طول صف عبارت است از متوسط تعداد ماشین هایی که در هر خط در هر مسیر ورودی

در زمان مشخصی قرار می گیرند. شاخص متوسط زمان تـاخیر، نشـان دهنده زمان تاخیر برای هر ماشین در هر کیلومتر می باشد. این مقدار از اختلاف بین زمان سفر در شـرایط بـدون ترافیـک و زمـان سـفر در شرایط ترافیکی موجود محاسبه می شود. انحراف از معیار زمان توقـف نیز مشخص کننده عدالت و مساوات الگوریتم بین تمام خودروها است. تمامی این شاخص ها توسط نرم افـزار شـبیه سـاز ترافیکـی در طـول شبیه سازی به طور خودکار محاسبه می شوند.لازم به ذکر است که در این تحقیق از متوسط طول صـف فقـط بـه عنـوان یـک معیـار جهـت ارزیابی رفتار سیستم پیشنهادی استفاده شده است و بررسـی مفـاهیم موجود در تئوری صف هدف مقاله حاضر نمی باشد.

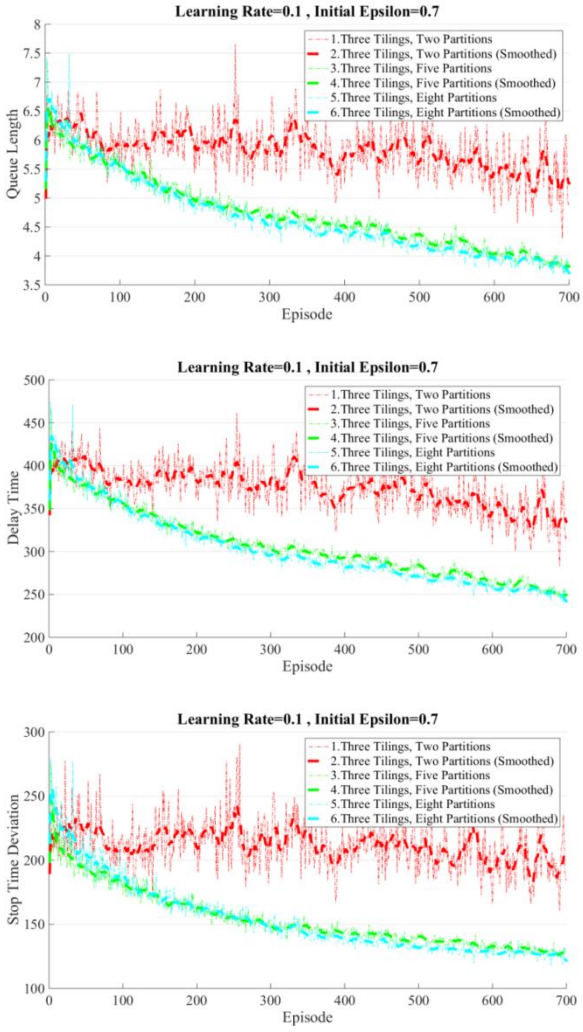
در شکل ۵، متوسط طول صف، زمان تـاخیر و انحـراف از معیـار زمـان توقف برای سناریوی اول نشان داده شده است. در ایـن شـکل خطـوط نازک از میانگین گیـری ۴ بـار اجـرای الگـوریتم و خطـوط ضـخیم از میانگین گیری خطوط نازک در بازه های ۱۱ سـاعته (اپیـزود) بدسـت آمده اند.



شکل (۵):مقایسه عملکرد الگوریتم یادگیری تقویتی در سناریوی اول به ازای تعداد Tile های مختلف

همانطور که مشخص است با۲Tile هیچ گونـه یـادگیری انجـام نشـده است و علت این امر کمبود بیش از حد تعداد تقسیم بندی هـا در هـر بعد است. برای ۵Tiles و ۸ یادگیری انجام گرفته است و تقریباً به یک مقدار یکسانی همگرا شده انـد. امـا سـرعت همگرایـی الگـوریتم بـرای ۵Tiles بیشتر است که علت این امر کم بودن تعـداد پارامترهـا بـرای ۵Tileدر مقایسه با۸Tile است. همچنین یادگیری با ۵Tile منجـر بـه میانگین طول صف، زمان تاخیر و انحراف از معیار زمان توقف کمتـری شده است.

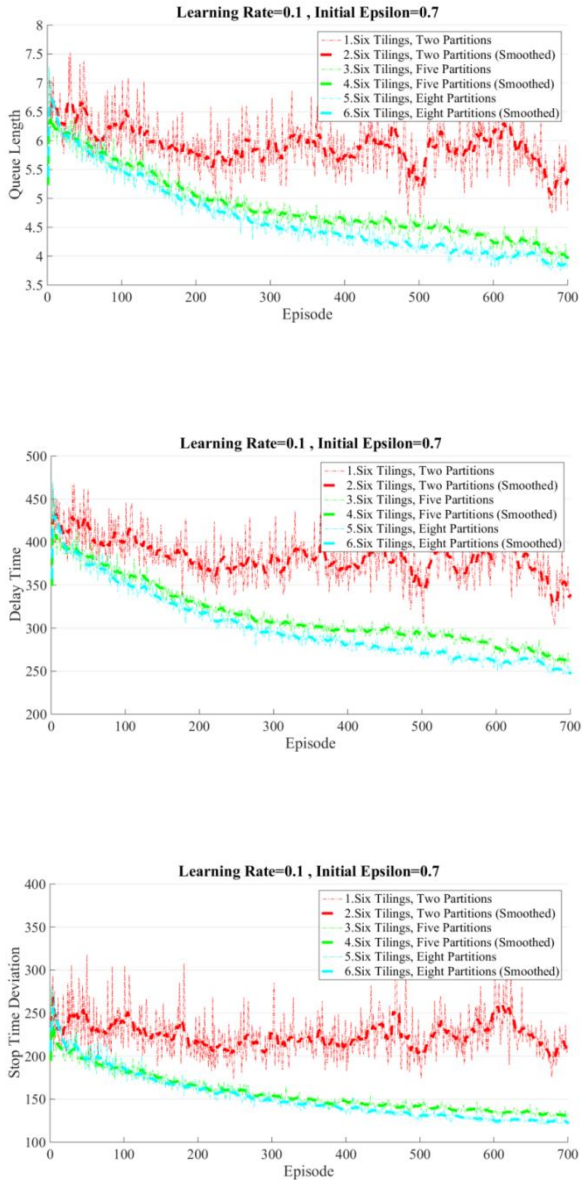
در شکل ۶، عملکرد الگوریتم یادگیری تقویتی در سناریوی دوم به ازای شاخص های مختلف آمده است. همانطور کـه مشـخص اسـت هماننـد سناریوی اول به ازای Tile ۲ هیچ گونه یادگیری انجام نگرفته اسـت و یادگیری به ازای ۵Tileو ۸ به نتیجه تقریباً مشابه ای رسیده است.



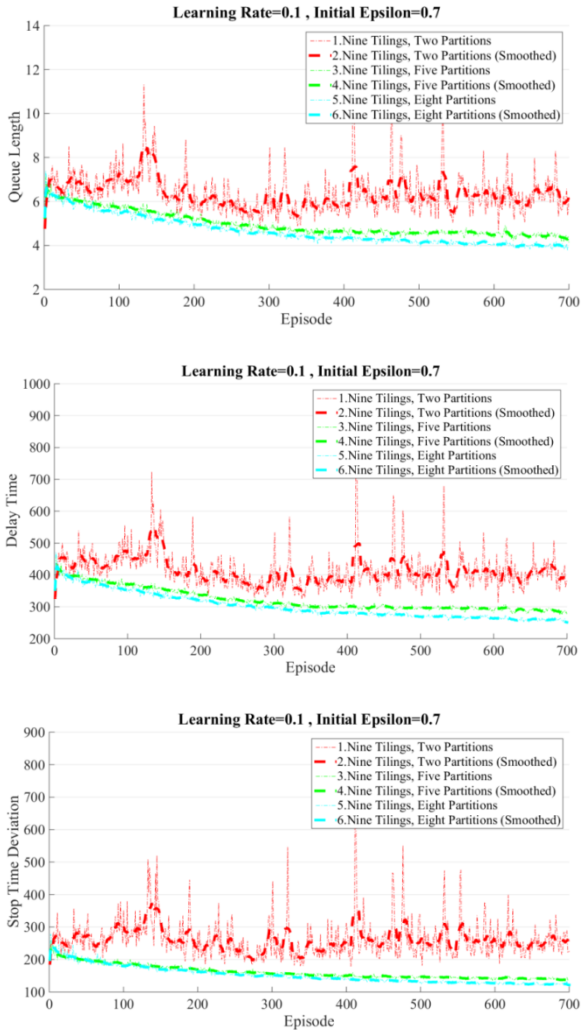
شکل (۶): مقایسه عملکرد الگوریتم یادگیری تقویتی در سناریوی دوم به ازای تعداد Tile های مختلف

نتایج عملکرد الگوریتم یادگیری تقویتی در سناریوی سوم و چهـارم در شـکل هـای۷ و ۸ آمـده اسـت. هماننـد دو سـناریوی اول بـه ازای

Tile۲نمودار به همگرایی نرسیده است. عملکرد یادگیری تقویتی به ازای Tile۸ به مراتب بهتر از Tile۵بوده استکه علت این موضوعاین است که با فرض بالا بودن تعداد Tile ها در صورتی الگوریتم عملکرد مناسبی خواهد داشت که تعداد Tiling ها نیز بالا باشد. به عنوان مثال در سناریوی اول به دلیل اینکه فقط از یک Tiling استفاده شده بود الگوریتم به ازای Tile۸ عملکرد مناسبی نداشت. زیرا در هر حالت محیط فقط ارزش یک Tile به روز می شود اما در سناریوی سوم و چهارم در هر حالت محیط ارزش چندین Tile به روز رسانی می شود. در شکل ۹، عملکرد بهترین Tile ها برای سناریوی های مختلف با یکدیگر مقایسه شده اند.



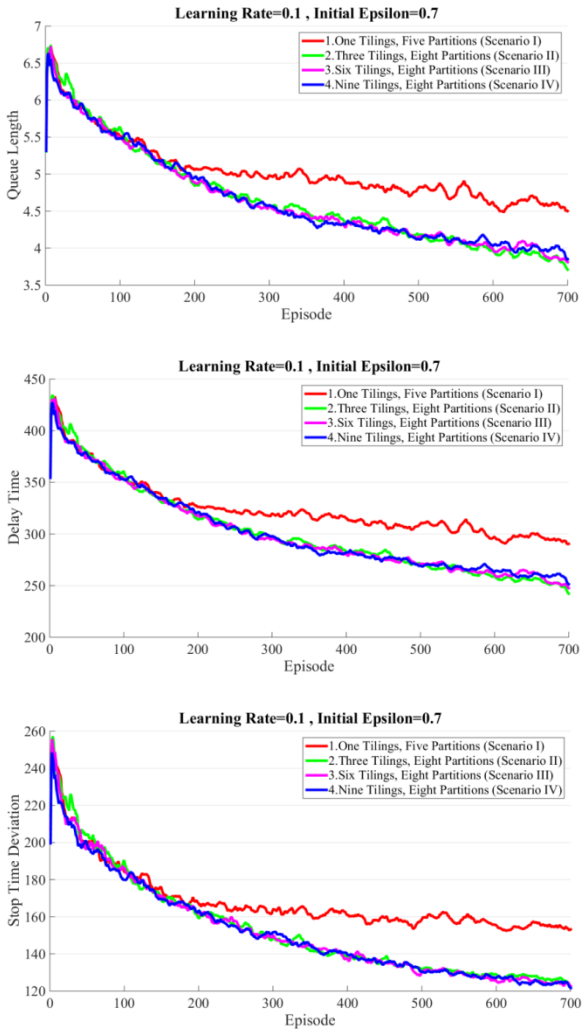
شکل (۷):مقایسه عملکرد الگوریتم یادگیری تقویتی در سناریوی سوم به ازای تعداد Tile های مختلف



شکل (۸): مقایسه عملکرد الگوریتم یادگیری تقویتی در سناریوی چهارم به ازای تعداد Tile های مختلف

همانطور که مشخص است اختلاف فاحشی بین عملکرد سناریوی اول با سه سناریوی بعدی وجود دارد که نشان می دهد انعطاف پذیری عامل و سرعت یادگیری آن به علت استفاده کردن از یک Tiling بسیار پایین تر از سایر سناریوی ها می باشد. همچنین با افزایش تعداد Tling ها به ۳ سرعت یادگیری عامل بهبود قابل ملاحظه ای یافته است اما با افزایش تعداد Tiling ها به ۶ و ۹ تغییر قابل ملاحظه ای در سرعت یادگیری ایجاد نشده است و فقط تعداد پارامترهای مجهول تنظیمی θ افزایش یافته است. عملکرد الگوریتم یادگیری تقویتی برای سناریوی دوم به ازای 3 Tilings و 8 Tiles برای شاخص های طول صف و زمان تاخیر از سایر سناریوها بهتر بوده اما در مورد شاخص سوم عملکرد سناریوهای سوم و چهارم بهتر از سایر سناریوها بوده است. جدول ۲، متوسط مقادیر طول صف و متوسط زمان تاخیر و انحراف از معیار زمان توقف را برای ۱۰ اپیزود انتهایی نشان می دهد. این جدول نشان می دهد که سناریوی دوم به ازای تمام Tile های ۲، ۵ و ۸ دارای عملکرد بهتری نسبت به سایر سناریوها است. همچنین از جدول

۲ و شکل ۹ می تواند فهمید که افزایش تعداد Tiling ها به معنی بهبود عملکرد یادگیری تقویتی و افزایش سرعت آن نمی باشد.



شکل (۹): مقایسه عملکرد بهترین Tile ها در سناریوی های مختلف

۸- ارزیابی و اعتبار سنجی نتایج

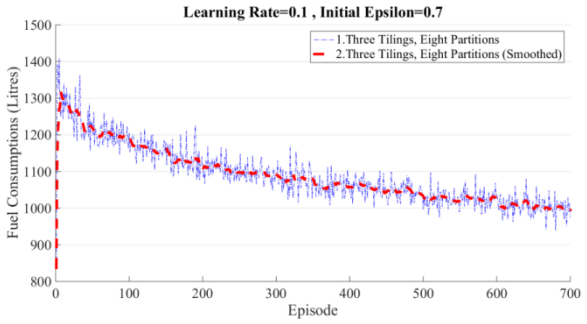
یکی از تبعات منفی عملکرد غیر بهینه چراغ راهنمایی افزایش میزان آلودگی هوا و مصرف سوخت است. آلاینده های هوا تاثیر زیانباری بر محیط زیست باقی می گذارند. در مقیاس وسیع تماس انسان با هوای آلوده منشاء بیماری های متعدد و گاه کشنده است. بیماری های قلبی و تنفسی که به خصوص بر اثر وجود ذرات معلق ایجاد و تشدید می شوند از جمله شایع ترین پیامدهای هوای آلوده در شهرهای بزرگ است. باران های اسیدی که خود ناشی از وجود آلاینده ها در هوا هستند، موجب از میان رفتن جنگل ها، پوششگیاهی و بناهای تاریخی می شوند.

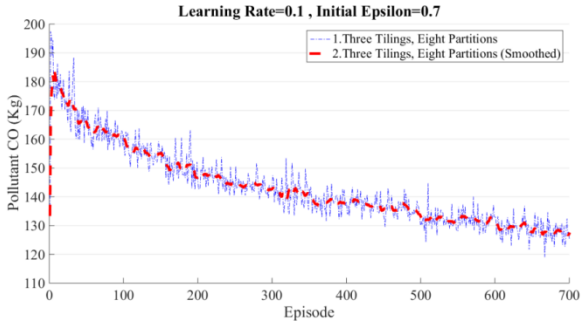
جدول (۲): ارزیابی سناریوی ها بر اساس شاخص های مختلف

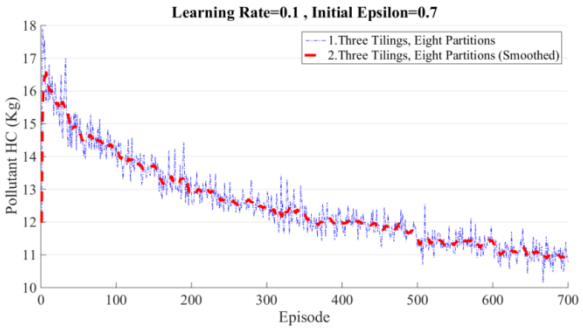
شاخص	سناریو	Tiles# ۲	۵	۸
متوسط طول صف	سناریوی اول (۱ Tiling)	۶,۰	۴,۶	۴,۶
	سناریوی دوم (۳ Tilings)	۵,۳	۳,۹	۳,۸
	سناریوی سوم (۶ Tilings)	۵,۴	۴,۰	۳,۸
	سناریوی چهارم (۹ Tilings)	۵,۹	۴,۳	۳,۹
متوسط زمان تاخیر	سناریوی اول (۱ Tiling)	۳۸۹,۸	۲۹۳,۲	۲۹۶,۳
	سناریوی دوم (۳ Tilings)	۳۳۶,۰	۲۵۱,۸	۲۴۸,۰
	سناریوی سوم (۶ Tilings)	۳۴۶,۷	۲۶۳,۱	۲۵۱,۱
	سناریوی چهارم (۹ Tilings)	۳۸۹,۸	۲۸۱,۴	۲۵۷,۷
انحراف از معیار زمان توقف	سناریوی اول (۱ Tilings)	۲۱۷,۳	۱۵۳,۸	۱۵۱,۱
	سناریوی دوم (۳ Tilings)	۱۹۷,۲	۱۲۷,۷	۱۲۴,۹
	سناریوی سوم (۶ Tilings)	۲۱۵,۶	۱۳۱,۴	۱۲۴,۴
	سناریوی چهارم (۹ Tilings)	۲۶۵,۳	۱۵۸,۱	۱۴۷,۷

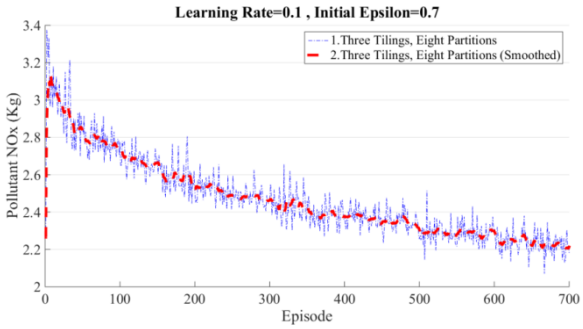
آلاینده های مربوط به موتورهای بنزینی، گازوئیلی و سوخت های جایگزین، عمدتاً کربن مونواکسید، اکسیدهای نیتروژن، هیدروکربن های نسوخته هستند که سه آلاینده گاز هستند. با توجه به اینکه سناریوی دوم به ازای تعداد ۸Tile دارای بهترین عملکرد می باشد میزان مصرف سوخت و انتشار آلاینده های CO، NOx و HC برای این سناریو بررسی می شود. نر م افزار شبیه ساز ترافیکی مورد استفاده، توانایی شبیه سازی میزان آلایندگی و مصرف سوخت را دارد. در این نرم افزار جهت محاسبه میزان مصرف سوخت و آلایندگی، برای هر خودرو چهار حالت توقف کامل، کاهش شتاب، افزایش شتاب و حرکت با سرعت ثابت در نظر گرفته شده است و برای هر حالت میزان مصرف سوخت و میزان تولید آلاینده ها (CO، NOx و HC) به صورت مجزا معرفی می شود.در طول شبیه سازی براساس سرعت هر خودرو و مسافتی که در شبکه طی می نماید میزان آلاینده ها و مصرف سوخت محاسبه می شود. شکل ۱۰ میزان مصرف سوخت و آلایندگی را برای سناریوی دوم نشان می دهد. همانطور که مشخص است مجموع میزان سوخت مصرفی و آلاینده های مختلف در طول یادگیری دارای روند کاهشی بوده که این موضوع نشان دهنده یادگیری چراغ های راهنمایی و نزدیک شدن آن ها به سیاست بهینه است.

همچنین به منظور اعتبار سنجی روش ارائه شده، نتایج آن با روش رایج زمان بندی در مهندسی ترافیک؛چراغهای پیش زمانبندی شده (زمان ثابت)؛مقایسه شده است. چراغ های پیش زمان بندی شده، چراغهایی هستند که زمانبندی از پیش تعیین شده و معلومی را بدون توجه به تغییرات شرایط ترافیک (نوسانات جریان ورودی به شبکه) لحاظ می کنند. جدول ۳ عملکرد روش زمان ثابت، را با روش پیشنهادی بر اساس شاخص های مختلف برای ۱۰ اپیزود انتهای مقایسه می کند. همانطور که مشخص است عملکرد روش پیشنهادی نسبت به کنترلر زمان ثابت به طور متوسط ۱۴٫۲٪ برای کل شاخص ها بهبود داشته است.









شکل (۱۰):بررسی عملکرد سناریوی دوم به ازای ۸ Tiles

جدول (۳): مقایسه عملکرد چراغ های پیش زمان بندی شده و چراغ های هوشمند مبتنی بر یادگیری تقویتی پیوسته

	زمان ثابت	یادگیری تقویتی پیوسته	میزان بهبود
متوسط طول صف در هر ساعت	۴٫۷۱۵	۳٫۷۹	٪۱۹٫۶
متوسط زمان سفردر هر ساعت (ثانیه/کیلومتر)	۳۷۴٫۲۳۳	۳۱۳٫۹۸	٪۱۶٫۱
انحراف از معیار زمان سفر در هر ساعت (ثانیه/کیلومتر)	۱۵۱٫۶۴۱	۱۲۸٫۶۸۹۳	٪۱۵٫۱
متوسط زمان تاخیر در هر ساعت (ثانیه/کیلومتر)	۳۰۸٫۳۱۳	۲۴۸٫۰۶۲	٪۱۹٫۵
انحراف از معیار زمان تاخیر در هر ساعت (ثانیه/کیلومتر)	۱۵۱٫۵۳۳	۱۲۸٫۵۹۱۵	٪۱۵٫۱
متوسط زمان توقف در هر ساعت (ثانیه/کیلومتر)	۲۸۸٫۲۳۱	۲۲۷٫۶۲۹۸	٪۲۱٫۰
انحراف از معیار زمان توقف در هر ساعت (ثانیه/کیلومتر)	۱۴۸٫۷۲۹	۱۲۴٫۹۳۴۹	٪۱۶٫۰
متوسط سوخت مصرف شده در هر ساعت (لیتر)	۱۰۵۷٫۷۰	۹۹۳٫۹۸۱۸	٪۶٫۰۲
متوسط آلاینده CO تولید شده در هر ساعت (کیلوگرم)	۱۴۰٫۱۰	۱۲۷٫۱۱۶۴	٪۹٫۲۶
متوسط آلاینده HC تولید شده در هر ساعت (کیلوگرم)	۱۲٫۴۴	۱۰٫۹۰۸۶	٪۱۲٫۳۱
متوسط آلاینده NOx تولید شده در هر ساعت (کیلوگرم)	۲٫۳۷	۲٫۲۱۰۵	٪۶٫۷۵

۹- نتیجه‌گیری

طراحی چراغ های راهنمایی بهینه و وفق پذیر همواره یکی از دغدغه های مهم مهندسان ترافیک بوده است. این دغدغه هنگامی بیشتر می

گردد که با محیط ترافیکی پیچیده، با تعداد ابعاد بالا و غیر قطعی سروکار داشته باشیم. از طرف دیگر باید در نظر داشت که در مسائل کنترل ترافیک، دسترسی اولیه به جواب و سیگنال کنترلی بهینه امکان پذیر نمی باشد، بنابراین نمی توان از روش های با سرپرست (همانند شبکه عصبی) که به طور مسقیم عامل را از عمل بهینه مطلع می سازد؛ استفاده نمود. به همین دلیل باید از روش هایی که در آن سرپرستی و احتیاج به دانش اولیه وجود ندارد استفاده کرد.یادگیری تقویتی یکی از این روش ها است که امروزه به دلیل ویژگی های منحصر به فرد آن بسیار مورد استفاده قرار می گیرد. در یادگیری تقویتی، تنها مسیر اطلاع رسانی به عامل سیگنال پاداش است که با استفاده از آن به عامل گفته می شود عملی که انجام داده به چه میزان مناسب و یا نامناسب بوده است. در صورتی که فضای حالت-عمل بسیار بزرگ باشد، روش های رایج یادگیری تقویتی (گسسته) که نیاز به تجربه اندوزی در تمامی حالت های محیط دارند کارایی خود را از دست می دهند. استفاده از روش های یادگیری تقویتی پیوسته که عامل برای تقریب زدن ارزش تمام حالات-اعمال نیازی به تجربه اندوزی مستقیم ندارد می تواند برای حل این چالش مفید واقع شود. نتایج این تحقیق نشان دادند که یادگیری تقویتی پیوسته دارای سرعت همگرایی بالایی در مسائل با فضای حالت-عمل بزرگ همانند ترافیکاست.

مراجع

[1] By, R. A. d., Georgiadou, P. Y., Knippers, R. A., Kraak, M. J., Sun, Y., Weir, M. J. C., and Westen, C. J. v., Principles of Geographic Information Systems, Enschede, The Netherlands: The International Institute for Geo-Information Science and Earth Observation (ITC), 2011.

[2] Ligtenberg, A., Wachowicz, M., Bregt, A. K., Beulens, A., and Kettenis, D. L., "A design and application of a multi-agent system for simulation of multi-actor spatial planning ", Journal of environmental management, Vol. 72, No.1-2 , 2004.

[3] Bazzan, A. L. C., "A Distributed Approach for Coordination of Traffic Signal Agents", Autonomous Agents and Multi-Agent Systems, Vol. 10, No. 1, 2005.

[4] Crooks, A., Castle, C., and Batty, M., "Key challenges in agent-based modelling forgeo-spatial simulation", Computers, Environment and Urban Systems, Vol. 32, No. 6, 2008.

[5] Tian, G., Ouyang, Y., Quan, Q., and Wu, J., "Simulating spatiotemporal dynamics of urbanization with multi-agent systems—A case study of the Phoenix metropolitanregion, USA", Ecological Modelling, Vol. 222, No. 5, 2011.

[6] Shoham, Y., and Leyton-Brown, K., Multiagent Systems: Algorithmic, Game Theoretic and Logical Foundations, Cambridge: Cambridge University Press, 2008.

[7] Weiss, G., Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence, London: The MIT Press, 2000.

[8] Sutton, R. S., and Barto, A. G., Introduction to Reinforcement Learning, Cambridge, MA: MIT Press, 1998.

[9] Schwartz, H. M., Multi-Agent Machine Learning A Reinforcement Approach, New Jersey: John Wiley & Sons, 2014.

[10] Abdulhai, B., and Kattan, L., "Reinforcement learning: Introduction to theory and potential for transport applications", Canadian Journal of Civil Engineering, Vol. 30, No. 6, 2003.

[11] Beigy, H., and Meybodi, M. R., "Utilizing Distributed Learning Automata to Solve Stochastic Shortest Path Problems", International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, Vol. 14, No. 5, 2006.

[12] Wang, X., Liu, L., and Eck, J., "Crime Simulation Using GIS and Artificial Intelligent Agents," Artificial Crime Analysis Systems: Using Computer Simulations and Geographic Information Systems, L. Liu and J. Eck, eds.: IGI Global, 2008.

[13] Tang, W., "Simulating Complex Adaptive Geographic Systems: AGeographically Aware Intelligent Agent Approach", Cartography and Geographic Information Science, Vol. 35, No. 4, 2008.

[14] Bone, C., and Dragicevic, S., "GIS and Intelligent Agents for Multiobjective Natural Resource Allocation: A Reinforcement Learning Approach.", Transactions in GIS, Vol. 13, No. 3, 2009.

[15] Bone, C., and Dragicevic, S., "Defining Transition Rules with Reinforcement Learning for Modeling Land Cover Change", Simulation, Vol. 85, No. 5, 2009.

[16] Bone, C., and Dragićević, S., "Incorporating spatio-temporal knowledge in an Intelligent Agent Model for natural resource management", Landscape and Urban Planning, Vol. 96, No. 2, 2010.

[17] Choy, M. C., Srinivasan, D., and Cheu, R. L., "Cooperative, hybrid agent architecture for real-timetraffic signal control", IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans, Vol. 33, No. 5, 2003.

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

[19] Jacob, C., and Abdulhai, B., "Machine learning for multi-jurisdictional optimaltraffic corridor control", Transportation Research Part A: Policy and Practice, Vol. 44, No. 2, 2010.

[20] Houli, D., Zhiheng, L., and Yi, Z., "Multiobjective Reinforcement Learning for Traffic Signal Control Using Vehicular Ad Hoc Network", EURASIP Journal on Advances in Signal Processing, Vol. 2010, 2010.

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

[22] Szepesvári, C., Algorithms for Reinforcement Learning, Canada: Morgan & Claypool Publishers, 2009.

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

[24] Russell, S., and Norvig, P., Artificial Intelligence: A Modern Approach (Third Edition), USA: Prentice Hall International, 2009.
[25] Slinn, M ,.Matthews, P., and Guest, P., Traffic Engineering Design Principles and Practice, Jordan Hill, Oxford, UK: Elsevier Butterworth-Heinemann Linacre House, 2005.
[26] وزارت‌نیرو, ترازنامه انرژی ۱۳۸۸.
[27] Morabia, A., Amstislavski, P. N., Mirer, F. E., Amstislavski, T. M., Eisl, H., Wolff, M. S., and Markowitz, S. B., "Air Pollution and Activity During Transportation by Car, Subway, and Walking ", American Journal of Preventive Medicine Vol. 37 No. 1, 2009.
[28] Timotheou, S., Panayiotou, C. G., and Polycarpou, M. M., "Transportation Systems: Monitoring, Control, and Security," Intelligent Monitoring, Control, and Security of Critical Infrastructure Systems, Studies in Computational Intelligence 565, E. Kyriakides and M. Polycarpou, eds., Berlin Heidelberg :Springer, 2015.
[29] Sun, J., Ma, Z., Li, T., and Niu, D., "Development and application of an integrated traffic simulation and multi-driving simulators", Simulation Modelling Practice and Theory, Vol. 59, 2015.
[30] Burghout, W., "Hybrid microscopic mesoscopic traffic simulation," Department of Infrastructure, Division of Transportation & Logistics, Centre for Traffic Simulation, Royal Institute of Technology, Stockholm, Sweden, 2004.
[31] Casas, J., Ferrer, J. L., Garcia, D., Perarnau, J., and Torday ,A., "Traffic Simulation with Aimsun," Fundamentals of Traffic Simulation, J. Barceló, ed., New York, NY: Springer New York, 2010.
[32] Peters, B., and Nilsson, L., "Modelling the Driver in Control," Modelling Driver Behaviour in Automotive Environments, P .C. Cacciabue, ed., Springer London, 2007.
[33] Tao, R., Wei, H., Wang, Y., and Sisiopiku, V., "Modeling Speed Disturbance Absorption Following Current State-Control Action-Expected State Chains: Integrated Car-Following and Lane-Changing Scenarios ", Transportation Research Record: Journal of the Transportation Research Board, Vol. 1934, 2005.
[34] Moridpour, S., Sarvi, M., and Rose, G., "Modeling the lane changing execution of multi class vehicles under heavy traffic conditions ", Transportation Research Record, No. 2161, 2010.
[35] Reuschel, A., "Vehicle movements in a platoon with uniform acceleration or deceleration of the lead vehicle", Zeitschrift des Oesterreichischen Ingenieur-und Architekten-Vereines, Vol. 95, 1950.
[36] Pipes, L. A., "An Operational Analysis of Traffic Dynamics", Journal of Applied Physics, Vol. 24, No. 3, 1953.
[37] Herman, R., and Potts, R. B., "Single Lane Traffic Flow Theory and Experiment".
[38] Gipps, P. G., "A behavioural car-following model for computer simulation", Transportation Research Part B: Methodological, Vol. 15, No. 2, 1981.

زیرنویسها

[1]Generalization
[2]Tile Coding
[3] Abdulhai
[4]Kattan
[5]Q Learning
[6]Wang
[7]Tang
[8]Geoprocessing unit
[9]State-Action-Reward-State-Action
[10]Bone
[11]Dragicevic
[12] British Columbia
[13]Forest Management Unit
[14]Forest Stand
[15]Episode
[16]Explore
[17] Choy
[18]Medina
[19]Jacob
[20] Cerebeller Model Articulate Controller
[21]Houli
[22]Tile Coding
[23]Radial Basis Functions
[24]Kanerva
[25]Stochastic Gradient
[26]Strategic Level
[27] Tactical Level
[28]Car following
[29]Leader Vehicle
[30]Mandatory Lane Change
[31]Discretionary Lane changes
[32]On-Ramp
[33]Off-Ramp
[34]Gap Acceptance
[35]Critical Gap