نشریه علمی- پژوهشی علوم و فنون نقشه برداری، دوره پنجم، شماره ۳، بهمن ماه ۱۳۹۴

یادگیری تقویتی براساس معماری عملگر- نقاد در سیستم های چند عامله برای کنترل ترافیک

محمد اصلانی*[۱]، محمد سعدی مسگری[۲]، حمید مطیعیان[۱]

[۱]دانشجوی دکتری سیستم‌های اطلاعات مکانی- دانشکده مهندسی نقشه‌برداری- دانشگاه صنعتی خواجه نصیرالدین طوسی
maslani@mail.kntu.ac.ir

[۲]دانشیار گروه سیستم‌های اطلاعات مکانی- دانشکده مهندسی نقشه‌برداری- دانشگاه صنعتی خواجه نصیرالدین طوسی
(عضو قطب علمی مهندسی فناوری اطلاعات مکانی)
mesgari@kntu.ac.ir

(تاریخ دریافت تیر ۱۳۹۴، تاریخ تصویب دی ۱۳۹۴)

چکیده

در نیمه دوم قرن گذشته اغلب جوامع شاهد شروع پدیده ای بنام ترافیک شهری در خود بوده اند که علت رخداد چنین پدیده ای عبور تعداد زیادی خودرو در زمان یکسان از یک زیر ساخت حمل و نقلی یکسان می باشد. پدیده ترافیک شهری دارای پیامدهای اقتصادی و محیط زیستی کاملاً شناخته شده ای از جمله آلودگی هوا، کاهش در سرعت، افزایش زمان سفر، افزایش مصرف سوخت و حتی افزایش تصادفات می باشد. یکی از راه های اقتصادی برای مدیریت کردن افزایش تقاضای سفر و جلوگیری از ترافیک شهری، افزایش کارایی زیر ساخت های موجود از طریق سیستم های هوشمند کنترل ترافیک می باشد.

از سوی دیگر کنترل ترافیک به دلیل طبیعت توزیع یافته و خودمختار آن توسط سیستم های چند عامله به خوبی قابل مدلسازی می باشد. رانندگان و چراغ های راهنمایی را می توان به عنوان عامل هایی که رفتارهای هوشمندانه ای از خود نشان می دهند در نظر گرفت. برای ایجاد چنین رفتارهایی نیاز است که دانش لازمه از محیط اطراف در ذهن عامل قرار داده شود اما به دلیل پیچیدگی های بالای موجود در الگوهای ترافیک شهری و نایستا بودن اغلب محیط های ترافیکی قرار دادن یک دانش اولیه از محیط در ذهن عامل ها بسیار دشوار و غیر عملی می باشد. بنابراین نیاز به یک روشی که عامل در طول تعامل با محیط بتواند دانش لازمه را بدست آورد کاملاً ضروری است که در این تحقیق برای حل این چالش از یادگیری تقویتی استفاده شد. هدف مقاله حاضر بهبود استراتژی های کنترل ترافیک و به طور خاص کنترل هوشمند چراغ های راهنمایی از طریق توسعه تکنیک های یادگیری تقویتی در سیستم های چند عامله است. معماری عملگر – نقاد[۱] به عنوان یک معماری رایج در یادگیری تقویتی که دارای ساختار حافظه جداگانه ای هم برای سیاست و هم برای تابع ارزش است مورد استفاده قرار گرفت. نتایج این تحقیق نشان دادند که کنترل هوشمند چراغ های راهنمایی منجر به کاهش ۲۳٪ طول صف و ۱۶٪ زمان سفر نسبت به کنترل غیر هوشمند چراغ های راهنمایی برای یک تقاطع منفرد می شود.

واژگان کلیدی: سیستم های چند عامله، یادگیری تقویتی، معماری عملگر - نقاد و کنترل ترافیک

---

* نویسنده رابط

۱ Actor – Critic

۱- مقدمه

تئوری سیستم های پیچیده[۱] یکی از رویکردهایی است که در دو دهه اخیر توجه محققین را در کنترل ترافیک به خود جلب نموده است [۱]. در این رویکرد رفتار یک سیستم کنترل ترافیک از طریق بررسی اجزاء (پروسه های) تشکیل دهنده آن و تعاملات محلی میان آنها که منجر به بروز الگوهای تظاهری[۲] می شود تحلیل می گردد [۲]. در این تئوری، اثرات عدم قطعیت[۳] بر روی خروجی ها کاملاً لحاظ می شوند.

سیستم های پیچیده را می توان به صورت های مختلفی بررسی نمود که یکی ازکاربردی ترین آنها سیستم های چند عامله است [۳] که در این سیستم های عامل نقشی اساسی بازی را می کند. به طور ساده عامل ها موجودیت هایی هستند که در محیط قرار گرفته، آن را درک می کنند و در آن عمل می نمایند [۳]. علیرغم دیدگاه های مختلف پیرامون عامل، تقریباً همگان بر این باورند که عامل، خود مختار[۴]، پیش فعال[۵]، واکنشی[۶] و اجتماعی[۷] است. عامل ها برای رسیدن به اهداف خود، نیاز به تعامل با یکدیگر دارند. به دلیل پیچیدگی های بالا در محیط های ترافیک شهری و چند عامله بودن ذاتی آن ها استفاده از دیدگاه سیستم های چند عامله در کنترل ترافیک معقول به نظر می رسند [۴-۸].

نکته مهم دیگری که در حوزه هوش مصنوعی و یادگیری ماشین به آن توجه ویژه ای شده است توانایی تصمیم گیری خود مختار عامل ها در محیط های نسبتاً پیچیده است. عامل ها باید بتوانند براساس دانشی که در اثر تعامل با محیط کسب می کنند و بدون کنترل خارجی، رفتار عقلانی از خود بروز دهند. از دیدگاه دیگر، در بسیاری از موارد عامل ها فاقد دانش کافی اولیه از محیط می باشند و یا به دلیل ناایستا[۸] بودن محیط نیاز است که عامل عمل خود را متناسب با شرایط محیط و برای رسیدن به اهداف خود انتخاب نماید. از سوی دیگر، در اغلب مسائل از جمله کنترل ترافیک دسترسی اولیه به جواب و سیگنال کنترلی بهینه امکان پذیر نمی باشد؛ به همین دلیل باید از روش هایی که در آن سرپرستی احتیاج به دانش اولیه دقیق ندارد استفاده نمود. بنابراین مسئله مورد بررسی در این تحقیق اضافه نمودن توانایی یادگیری به عامل ها بدون دسترسی اولیه به جواب است. همچنین اهمیت مسئله این است که عامل می تواند بدون نیاز داشتن به مدل محیط دانش لازمه را از محیط بدست آورد. یادگیری تقویتی یک الگوریتم مدرن هوشمند است که به جهت دارا بودن قابلیتهایی همچون عدم نیاز به خروجی مطلوب، آموزش با استفاده از یک معیار اسکالر راندمان، امکان آموزش برخط، و درجه کاوش بالا، گزینه مناسبی جهت کنترل ترافیک می باشد [۹]. در واقع در یادگیری تقویتی به عامل گفته نمی شود که عمل صحیح در هر وضعیت از محیط چیست بلکه فقط با استفاده از یک معیار اسکالر که سیگنال تقویتی نامیده می شود میزان خوب بودن عمل به عامل نشان داده می شود. عامل با در دست داشتن این اطلاعات، سعی در پیدا نمودن عمل بهینه می نماید که این ویژگی یکی از نقاط قوت الگوریتم های یادگیری تقویتی به شمار می آید. الگویتم های یادگیری تقویتی متفاوتی در طول زمان ارائه شده اند که این الگوریتم ها را می توان به سه دسته ۱-عملگر- تنها[۹]، ۲-نقاد - تنها[۱۰] و ۳-عملگر- نقاد[۱۱] تقسیم نمود [۱۰]. از آنجائیکه معماری عملگر- نقاد دارای ویژگی های همگرایی مناسب تری در مقایسه با دو روش دیگر است و به طور همزمان از مزیت های روش های عملگر- تنها و نقاد- تنها بهره می برد [۹]، در این تحقیق از این معماری مورد استفاده قرار گرفت. این معماری دارای دو بخش عملگر و نقاد بوده که بخش نقاد برای تقریب تابع ارزش و بخش عملگر برای تولید عمل استفاده می شود. بخش نقاد مسئول پردازش پاداش های دریافتی از محیط و ارزیابی کیفیت سیاست مورد استفاده توسط عامل است و بخش عملگر با بکارگیری اطلاعاتی از نقاد پارامترهای سیاست خود را به روز رسانی می کند [۱۱].

در تحقیق حاضر که از سیستم های چند عامله برای کنترل ترافیک استفاده شده است دو نوع عامل خودمختار متفاوت تعریف شده اند: عامل های خودرو (عامل های غیر

۱ Complex Systems Theory
۲ Emerging Patterns
۳ Uncertainty
٤ Autonomous
٥ Proactive
٦ Reactive
۷ Social
۸ Nonstationary

۹ Actor-Only
۱۰ Critic-Only
۱۱ Actor-Critic

فعال[۱]) که دارای رفتارهایی از جمله شتاب گرفتن، ترمز کردن و سبقت گرفتن هستند و عامل های چراغ راهنمایی (عامل های فعال[۲]) که دارای توانایی یادگیری تقویتی عملگر- نقاد می باشند. چالش های بکارگیری معماری عملگر- نقاد در سیستم های چند عامله در هر مسئله، شامل انتخاب عمل مناسب، تعریف حالت ها و تعریف تابع یادگیری تقویتی می باشد که در این تحقیق راهکار مناسبی برای موارد مذکور در مسئله کنترل ترافیک ارائه شده است.

روش پیشنهادی در دو سناریوی متفاوت مورد ارزیابی قرار گرفت. در سناریوی اول یک چراغ راهنمایی که دارای توانایی یادگیری تقویتی است سعی در کنترل یک تقاطع منفرد می نماید و در سناریوی دوم همزمان ۹ تقاطع توسط ۹ چراغ راهنمایی هوشمند کنترل می شوند. برای شبیه سازی ترافیکی از نرم افزار AIMSUN و قابلیت توسعه آن توسط زبان برنامه نویسی C++ استفاده شد.

روش ارائه شده در این تحقیق با روش کنترل غیر هوشمند تقاطع ها مقایسه شد و نتایج نشان دادند که کنترل هوشمند چراغ های راهنمایی منجر به کاهش ۲۳٪ طول صف و ۱۶٪ زمان سفر نسبت به کنترل غیر هوشمند چراغ های راهنمایی شده است.

مقاله حاضر در ۷ بخش ساختار دهی شده است. در بخش ۲ پیشینه تحقیق، در بخش ۳ مبانی نظری تحقیق، در بخش ۴ نحوه انجام پیاده سازی و در بخش ۵ نتایج پیاده سازی، در بخش ۶ بحث و اعتبار سنجی نتایج و در بخش ۷ نتیجه گیری آورده شده است.

۲- پیشینه تحقیق

در زمینه کنترل چراغ های راهنمایی از طریق روش های کلاسیک می توان به ابزار Transyt [۱۲]، سامانه هایSCOOT[۳] [۱۳]، SCAT[۴] [۱۴]، OPAC[۵] [۱۵] و PRODYNE[۶] [۱۶] و RHODES[۷] [۱۷] اشاره نمود. ابزار Transyt یک روش برون خط[۸] برای تعیین زمان بهینه مجموعه ای از چراغ های راهنمایی زمان ثابت بر اساس الگوریتم تپه نوردی[۹] است. ورودی های این ابزار شامل هندسه خیابان ها، جریان ترافیکی، زمان سفر در هر خیابان، نرخ گردش به جهات مختلف در هر تقاطع و مجموعه ای از زمان های سبز و قرمز اولیه برای هر چرخه است. نقطه ضعف این روش محاسبه زمان بهینه چراغ ها براساس شرایط ترافیکی کاملاً استاتیک می باشد و این در حالی است که شرایط ترافیکی در روزها و ساعات مختلف روز با یکدیگر متفاوت هستند. سامانه SCOOT عملکردش شبیه Transyt است با این تفاوت که قابلیت لحاظ نمودن شرایط ترافیکی متفاوت به صورت بر خط[۱۰] را دارا می باشد. سامانه SCAT همانند سامانه SCOOT براساس داده های بر خط ترافیکی عمل می نماید و تفاوتش با SCOOT بکار گیری ساختار سلسله مراتبی و توزیع یافته است. در این سامانه کل منطقه به چندین زیر ناحیه تقسیم می شود که هر زیر ناحیه دارای ۱ تا ۱۰ تقاطع می باشد و هر زیر ناحیه به صورت مستقل توسط یک واحد مجزا کنترل می شود. سامانه های OPAC، PRODYN و RHODES به صورت توزیع یافته عمل می نمایند و نحوه عملکرد آنها به این صورت است که در هر بازه زمانی مشخص (مثلاً ۵ ثانیه) چراغ تصمیم می گیرد که آیا فاز جاری را تغییر دهد یا خیر؟ در خیابان های منتهی به هر تقاطع تعدادی سنسور قرار داده می شود که وضعیت ترافیکی آن خیابان ها را برای چراغ راهنمایی ارسال می کنند. پیچیدگی های بالای محاسباتی از جمله نقاط ضعف این سامانه ها به حساب می آید [۱۸].

ویرینگ[۱۱] در سال ۲۰۰۰، برنامه ریزی پویا[۱۲] را برای کنترل چراغ های راهنمایی به منظور کاهش زمان انتظار استفاده نمود. در این تحقیق فرض می شود که چراغ ها و خودروها دارای توانایی ارتباط برقرار کردن با یکدیگر می باشند و همچنین چراغ های راهنمایی از مقصد خودروها اطلاع دارند. خودروها زمان متوسط انتظار خود را در طول یادگیری تخمین زده و این زمان را به چراغ تقاطع پیش‌رو ارسال می کنند و چراغ مسیری را سبز می کند که در آن مجموع زمان انتظار خودروها بیش از سایر مسیرها باشد. نتایج این تحقیق نشان می دهد که روش پیشنهادی زمان

۱ Passive Agents
۲ Active Agents
۳ Split Cycle Offset Optimization Technique
۴ Sydney Coordinated Adaptive Traffic System
۵ Optimized Policies for Adaptive Control
۶ ProgrammationDynamique
۷ Real-Time, Hierarchical, Optimized, Distributed, and Effective System
۸ Offline
۹ Hill-Climbing
۱۰ Online
۱۱ Wiering
۱۲ Dynamic programming

انتظار را ۲۲٪ نسبت به حالتی که از چراغ های زمان ثابت استفاده شود کاهش می دهد [۱۹]. فرضیات بکار رفته در این مقاله با توجه به زیر ساخت های موجود در خیابان ها و چراغ ها غیر عملی می باشد. از طرف دیگر استفاده از برنامه ریزی پویا که یک روش مدل مبنا[۱] در یادگیری تقویتی است پیچیدگی های غیر ضروری را در مقایسه با روش های مستقل از مدل[۲] وارد می کند.

عبدلهای[۳] و همکاران در سال ۲۰۰۳ روش یادگیری تقویتی را برای یک تقاطع منفرد دارای ۲ فاز بکار گرفتند. ایشان طول صف خودروهای منتظر در پشت چراغ راهنمایی را به عنوان حالت محیط که توسط عامل قابل اندازه گیری می باشد در نظر گرفتند. عامل می تواند زمان سبز چراغ را تمدید و یا آنرا به فاز بعدی تغییر دهد به گونه ای که تعداد ماشین های منتظر در پشت تقاطع مینیمم شوند. ایشان از سه جریان ورودی ترافیکی یکپارچه، نسبت ثابت و متغیر برای تست کردن عملکرد روش پیشنهادی تحت شرایط ترافیکی متفاوت استفاده نمودند [۲۰].

کامپونوگارا[۴] و کراس[۵] در سال ۲۰۰۳ از الگوریتم یادگیری Q[۶] برای کنترل دو تقاطع مجاور به هم به صورت مستقل استفاده نمودند. آنها در مقاله خود نشان دادند که کنترل هوشمند چراغ ها با استفاده از یادگیری Q باعث بهبود شگرف عملکرد سیستم در مقایسه با غیر هوشمند بودن چراغ ها خواهد شد [۲۱].

چوی[۷] 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 و 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 سیستم نظارت می کند. در تحقیق ایشان از روش فازی عصبی برای یادگیری استفاده شده است [۲۲].

بول[۸] و همکاران در سال ۲۰۰۴ از سیستم های طبقه بندی کننده یادگیر برای کنترل شبکه ترافیکی متشکل از ۴ تقاطع استفاده نمودند. در تحقیق ایشان چراغ های راهنمایی که توسط یک سیستم طبقه بندی کننده یادگیر کنترل می شوند در هر تقاطع دارای ۲ فاز هستند که یک فاز برای حرکت از شمال به جنوب و فاز دیگر برای حرکت از شرق به غرب است. سیستم کنترل کننده در هر تقاطع زمان فاز بهینه را از طریق استخراج تعدادی قانون اگر-آنگاه بدست می آورد. نتایج کار ایشان نشان دادند که عملکرد چراغ راهنمایی با بکارگیری سیستم طبقه بندی کننده یادگیر بهبود قابل ملاحظه ای در مقایسه با عملکرد چراغ راهنمایی زمان ثابت داشته است [۲۳].

درسنر[۹] و استون[۱۰] در سال ۲۰۰۵ از روشی بر پایه اختصاص دادن فضا در یک تقاطع استفاده نمودند. در روش ایشان خودروها تقاطع پیش روی خود را از سرعت، شتاب، جهت و زمانی که به آن خواهند رسید مطلع می سازند. تقاطع با استفاده از اطلاعات دریافتی از خودروها تعیین می کند که فضای لازم برای عبور کدام خودروها وجود خواهد داشت. خودروهایی که فضای لازم برای عبور را داشته باشند اجازه عبور خواهند داشت اما خودروهای دیگر باید سرعت خود را کاهش دهند تا فضای لازم برای عبور آنها فراهم شود [۲۴]. فرضیات بکار رفته در این مقاله با توجه به زیر ساخت های موجود غیر عملی می باشد.

مدینا[۱۱] 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 مورد کنترل به سمت تقاطع های مجاور استفاده کردند. به این ترتیب عامل ها علاوه بر در نظر گرفتن تعداد خودروهای منتظر در مسیرهای ورودی خود، تعداد خودروهایی که در تقاطع های مجاور متوقف هستند را به عنوان وضعیت عامل در نظر می گیرند. 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 [۲۵]. گسسته سازی اعمال و سرعت پایین یادگیری از نقاط ضعف این تحقیق می باشد.

هولی[۱۲] و همکاران در سال ۲۰۱۰ از یادگیری تقویتی چند هدفه برای کنترل چندین چراغ راهنمایی استفاده

۱ Model Based
۲ Free Model
۳ Abdulhai
۴ Camponogara
۵ Kraus
۶ Q-Learning
۷ Choy
۸ Bull
۹ Dresner
۱۰ Stone
۱۱ Medina
۱۲ Houli

کردند. اهداف بهینه سازی شامل تعداد توقف های خودروها، متوسط زمان توقف و حداکثر طول صف خودروها در هر تقاطع است [۲۶].

تمام منابع آورده شده در فوق جزء بهترین تحقیقات در زمینه هوش مصنوعی و کنترل ترافیک بوده اند. اما در تمام این منابع محیط ترافیکی شبیه سازی شده کاملاً ساده و دور از واقعیت های موجود می باشد. در این تحقیق سعی شده است که شبیه سازی ترافیکی و رفتار رانندگان تا حد امکان به واقعیت نزدیک باشند. همچنین در اغلب تحقیقات انجام گرفته از الگوریتم های ساده و ابتدایی یادگیری تقویتی نظیر یادگیری Q و سارسا[۱] استفاده شده است.

۳- مبانی نظری تحقیق

۳-۱- مدل تصمیم گیری مارکوف و یادگیری تقویتی

یادگیری تقویتی به معنای آموزش آنچه باید انجام شود –چگونگی نگاشت وضعیت ها به عمل- برای ماکزیمم نمودن یک معیار اسکالر راندمان می باشد. در یادگیری تقویتی، تصمیم گیرنده را عامل هوشمند و چیزی که عامل با آن تعامل دارد شامل همه چیز غیر از خود عامل، محیط نامیده می شود.

در مسائل یادگیری تقویتی محیط باید از دید عامل دارای خاصیت مارکوف باشد. خاصیت مارکوف بدین معنی است که حالت بعدی محیط و پاداش دریافتی تنها به عمل و حالت قبلی عامل در محیط بستگی دارد. یک چارچوب ریاضی مرسوم برای مسئله یادگیری تقویتی که دارای خاصیت مارکوف است، مدل تصمیم گیری مارکوف[۲] (MDP) می باشد [۲۷]. مدل تصمیم گیری مارکوف از یک چهارتایی $\{S, A, R^{a}_{ss'}, P^{a}_{ss'}\}$ تشکیل شده است که S مجموعه حالت های محیط، A مجموعه اعمال عامل، $P^{a}_{ss'}$ احتمال انتقال از حالت s به s' تحت انجام عمل a و $R^{a}_{ss'}$ متوسط پاداش بدست آمده در صورت انتقال از حالت s به s' تحت عمل a است.

در یادگیری تقویتی، در هر گام زمانی t، عامل حالت فعلی s_t از فضای حالت S را مشاهده نموده و عملی را از فضای عمل متناهی A_s، براساس سیاست فعلی اش انتخاب و به محیط اعمال می کند و در پی آن، محیط با احتمال $P(s_t,a_t,s_{t+1})$ به حالت جدید s_{t+1} از فضای S انتقال یافته و عامل سیگنال تقویتی $r(s_t,a_t)$ را که با r_{t+1} نشان داده می شود دریافت می کند [۹].

قانونی که عامل با توجه به آن در هر حالت، عملی را برای اجرا انتخاب می کند، سیاست می نامند و معمولاً با $\pi(s,a)$ که نشان دهنده احتمال انتخاب عمل a، در حالت s است نمایش داده می شود. مبنای کار در یادگیری تقویتی بر اساس پاداش و جریمه است و هدف پیدا نمودن سیاستی است که منجر به حداکثر کردن مجموع پاداش های دریافتی در طول یادگیری شود. بر این اساس عامل یاد می گیرد عملی را انتخاب کند که او را به حالتی با بیشترین ارزش برساند. ارزش حالت s تحت سیاست π توسط رابطه ۱ تعریف می شود. به عبارت دیگر ارزش یک حالت، کل مقدار پاداشی است که عامل می تواند بعد از شروع از آن نقطه، انتظار دریافت آنرا داشته باشد.

$$V^{\pi}(s) = E_{\pi}\left\{\sum_{k=0}^{\infty}\gamma^{k} r_{t+k+1} \mid s_t = s\right\}, 0 \le \gamma \le 1 \qquad (۱)$$

به طور مشابه ارزش زوج وضعیت - عمل (s,a) تحت سیاست π که با نماد $Q^{\pi}(s,a)$ نشان داده می شود، برابر با امید ریاضی کل پاداش های است که اگر عامل در وضعیت s عمل a را انجام دهد و سپس تا پایان، انتخاب های خود را با سیاست π ادامه دهد بدست خواهد آورد (رابطه ۲).

$$Q^{\pi}(s,a) = E_{\pi}\{\textstyle\sum_{k=0}^{\infty}\gamma^{k} r_{t+k+1} \mid s_t = s\,, a_t = a\} \qquad (۲)$$

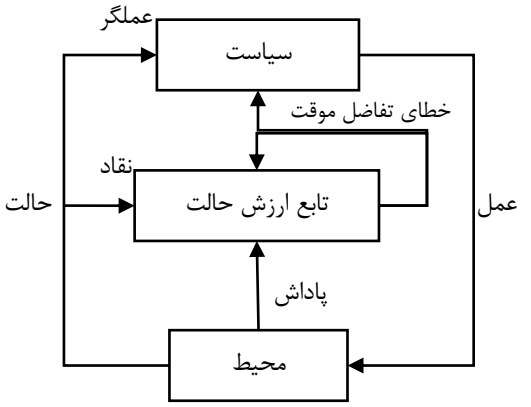
هرچه تعداد تجربه ها و تعامل های عامل با محیط بیشتر شود، تخمین بهتری از توابع ارزش می توان بدست آورد. حل یک مسئله یادگیری تقویتی به معنی پیدا نمودن سیاستی است که ارزش تمام حالت های محیط تحت آن سیاست بیشینه شود. در این تحقیق از معماری عملگر-نقاد برای پیدا نمودن سیاست بهینه استفاده شده است.

۳-۲- معماری عملگر- نقاد لاندا

در این روش یادگیری ساختار حافظه جداگانه ای هم برای سیاست و هم برای تابع ارزش در نظر گرفته می شود. از آنجائیکه معماری عملگر- نقاد از اصول یادگیری تقویتی تفاضل موقتی استفاده می نماید، قابلیت پیاده سازی به صورت زمان حقیقی در طی مسیر سیستم را دارا

[۱] SARSA
[۲] Markov Decision Process (MDP)

می باشد. در این معماری ساختار سیاست به عنوان عملگر شناخته می شود زیرا از آن برای تولید عمل استفاده می شود و ساختار تابع ارزش به عنوان نقاد شناخته می شود زیرا آن برای نقد اعمال انجام گرفته توسط عملگر بکار گرفته می شود. یادگیری در معماری عملگر- نقاد به صورت On-Policy است به این معنی که نقاد باید درباره سیاستی که توسط عملگر دنبال می شود یادگیری را به طور همزمان انجام دهد. در طول یادگیری در هر گام زمانی نقاد یک خطای تفاضل موقت را تولید و براساس آن، یادگیری در عملگر و نقاد انجام می شود (شکل ۱). بعد از اجرای هر عمل، حالت جدید محیط توسط نقاد (رابطه ۳) ارزیابی شده و تعیین می شود که آیا حالت محیط بهتر شده است یا خیر؟



شکل ۱- معماری عملگر- نقاد

$$\delta_t = r_{t+1} + \gamma V(s_{t+1}) - V(s_t) \qquad (۳)$$

در صورت مثبت بودن δ_t تمایل برای انتخاب عمل انجام شده باید تقویت شود و در صورت منفی بودن δ_t تمایل برای انتخاب عمل انجام شده باید کاهش یابد. در این تحقیق برای افزایش سرعت یادگیری روش اثر شایستگی[۱] برای به روز رسانی ارزش حالت های مختلف محیط بکار گرفته شد. در روش اثر شایستگی ارزش حالت های مختلف محیط توسط روابط ۴ و ۵ به روز می شوند:

$$V(s_t) \leftarrow V(s_t) + \alpha\delta_t e_t(s_t), \quad 0 < \alpha < 1 \qquad (۴)$$

$$e_t(s_t) = \begin{cases} \gamma\lambda e_{t-1}(s) & ,if\ s \neq s_t \\ \gamma\lambda e_{t-1}(s) + 1 & ,if\ s = s_t \end{cases} \quad 0 \leq \gamma, \lambda \leq 1 \qquad (۵)$$

در روابط ۴ و ۵، α نرخ یادگیری، γ نرخ تخفیف، λ میزان تاثیر پذیری ارزش حالت های ابتدایی اپیزود از ارزش حالت ها و سیگنال های انتهایی محیط است. برای λ=0، فقط یک حالت از محیط در گام زمانی t مقدار غیر صفر شایستگی دارد و بنابراین فقط ارزش آن حالت به روز می شود. برای λ مثبت، عامل می بایست در هر گام زمانی پیش بینی ها و آثار شایستگی را برای تمام حالات بروز نماید و به همین دلیل پیاده سازی با استفاده از λ>0 از نظر محاسباتی سنگین تر از زمانی است که λ=0 مورد استفاده قرار گیرد، مخصوصاً در مواقعی که فضای حالت بزرگ باشد. به هر حال، استفاده از λ مثبت سرعت یادگیری را به طور قابل ملاحظه ای افزایش می دهد. مقدار α در این تحقیق برابر ۰,۲، مقدار γ برابر ۰,۹۰ و مقدار λ برابر ۰,۸۵ انتخاب شدند.

احتمال انجام اعمال مختلف توسط سیاست ε-greedy محاسبه می شود (رابطه ۶) که ε نشان دهنده میزان تمایل عامل برای کنکاش ارزش اعمال مختلف در حالت های مختلف محیط است. هرچه میزان ε به یک نزدیک تر باشد، سیاست عامل به تصادفی نزدیک تر و تمایل عامل به کنکاش اعمال مختلف افزایش می یابد. هرچه میزان ε به صفر نزدیک تر باشد، سیاست عامل به حریصانه نزدیک تر و تمایل عامل به کنکاش کاهش می یابد.

$$\begin{aligned} &\pi_t(s,a) \\ &= \Pr\{a_t = a | s_t = s\} \\ &= \begin{cases} 1 - \varepsilon + \frac{\varepsilon}{|A_s|} & ,\ if\ a = argmax_{a' \in A_s} P(s,a') \\ \frac{\varepsilon}{|A_s|} & ,\ else \end{cases} \\ &\qquad 0 \leq \varepsilon \leq 1 \end{aligned} \qquad (۶)$$

در رابطه ۶، P(s,a) مقادیر پارامترهای سیاست در عملگر هستند که در طول یادگیری تغییر می کنند و نشان دهنده تمایل برای انتخاب هر عمل a در حالت محیط s است. تقویت کردن و یا ضعیف کردن تمایل برای انتخاب هر عمل می تواند توسط افزایش یا کاهش $P(s_t,a_t)$ در زمان های مختلف انجام شود (رابطه ۷).

$$P(s_t, a_t) \leftarrow P(s_t, a_t) + \beta\delta_t \qquad (۷)$$

در رابطه ۷، β پارامتر طول گام[۲] بوده و دارای یک مقدار مثبت می باشد. مقدار β در این تحقیق برابر ۱۰۰ انتخاب شد.

۱ Eligibility Trace

۲ Step-Size

۴- پیاده سازی

دو سناریو در این تحقیق برای کنترل ترافیک در نظر گرفته شده اند. در سناریوی اول یک چهار راه منفرد که دارای ۴ مسیر ورودی است توسط یک چراغ راهنمایی هوشمند چهار فازه با توانایی یادگیری تقویتی کنترل می شود. این چهار فاز به ترتیب از چپ به راست در شکل ۲ آورده شده اند. در این سناریوی زمان کل هر چرخه چراغ متغیر بوده اما ترتیب فازها ثابت می باشند. در ابتدای هر فاز و براساس تعداد ماشین های منتظر در هر مسیر ورودی، برای فاز جاری مدت زمانی به عنوان زمان سبز انتخاب می شود و بعد از اتمام زمان سبز در هر فاز، ۵ ثانیه به عنوان زمان زرد قبل از شروع فاز بعدی در نظر گرفته می شود. تمام خیابان های متصل به چهار راه سه خطه بوده و به طول ۳۰۰ متر می باشند. برای ارزیابی کامل عملکرد روش پیشنهادی در این تحقیق، بجای استفاده از جریان ترافیکی ثابت از یک جریان ترافیکی متغیر با زمان استفاده شد. به این ترتیب که سه نوع نرخ جریان ترافیکی سبک $\frac{Veh}{h}$ ۵۰۰، نیمه سنگین $\frac{Veh}{h}$ ۷۵۰ و سنگین $\frac{Veh}{h}$ ۱۰۰۰ در مسیرهای ورودی به تقاطع وارد می شوند. همچنین در طول شبیه سازی فرض می شود که ۶۰٪ از ماشین ها مسیر مستقیم و ۲۰٪ گردش به چپ و ۲۰٪ گردش به راست را انجام می دهند. همچنین مدت زمان شبیه سازی ۸۰۰ ساعت در نظر گرفته شد.

در سناریوی دوم یک شبکه ترافیکی متشکل از ۹ تقاطع که هر تقاطع توسط یک چراغ راهنمایی هوشمند چهار فازه با توانایی یادگیری تقویتی که فاز بندی آن همانند سناریوی اول است مورد بررسی قرار می گیرد (شکل ۳). همچنین در این سناریو از سه نوع نرخ جریان ترافیکی $\frac{Veh}{h}$ ۲۰۰، $\frac{Veh}{h}$ ۴۰۰ و $\frac{Veh}{h}$ ۶۰۰ برای مسیرهای ورودی استفاده شد. در طول شبیه سازی فرض شده است که ۳۳٪ از ماشین ها مسیر مستقیم، ۳۳٪ گردش به چپ و ۳۳٪ گردش به راست را انجام می دهند. شبیه سازی ترافیکی برای ۸۰۰ ساعت انجام گرفته و طول تمام خیابان ها ۲۵۰ متر، دو خطه با ماکزیمم سرعت ۵۰ کیلومتر بر ساعت می باشد. محیط شبیه سازی ترافیکی AIMSUN و از زبان برنامه نویسی C++ برای توسعه آن استفاده شد.

۴-۱- رفتارها و خصیصه های اجزاء مختلف کنترل ترافیک

۴-۱-۱- خودروها

رفتار و ویژگی خودروها توسط پارامترهای حداکثر سرعت، حداکثر شتاب افزایشی و حداکثر شتاب کاهشی قابل توصیف هستند . در این تحقیق حداکثر سرعت، حداکثر شتاب افزایشی و حداکثر شتاب کاهشی هر خودرو به ترتیب از توابع توزیع گوسین با میانگین های $۱۱۰\frac{\mathrm{Km}}{\mathrm{h}}$، $۳\frac{\mathrm{m}}{\mathrm{s}^2}$ و $۶\frac{\mathrm{m}}{\mathrm{s}^2}$ و انحراف از معیارهای $۱۰\frac{\mathrm{Km}}{\mathrm{h}}$ ، $۰,۲\frac{\mathrm{m}}{\mathrm{s}^2}$ و $۰,۵\frac{\mathrm{m}}{\mathrm{s}^2}$ انتخاب می شوند. موقعیت، سرعت و شتاب خودروها در هر ثانیه در طول شبیه سازی به روز می شوند.

۴-۱-۲- رانندگان به عنوان عامل های هوشمند

تصمیمات رانندگان در طول سفر به دو دسته تصمیمات کلان[۱] و تصمیمات خرد[۲] تقسیم می شود [۲۸]. تصمیمات کلان شامل انتخاب مقصد و انتخاب مسیر مناسب برای رسیدن به آن می باشد. تصمیمات خرد شامل تغییر خط حرکت در یک مسیر، سبقت گرفتن، انتخاب سرعت مناسب و گردش به راست یا چپ می باشد. هر راننده شیوه رانندگی مخصوص به خود را دارا می باشد. سبک رانندگی را می توان از طریق یک سری پارامترها که به نوعی مشخص کننده ویژگی های اخلاقی رانندگان هستند تقریب زد. در این تحقیق از پارامترهای زیر برای تقریب ویژگی های اخلاقی رانندگان استفاده شده است [۲۹, ۳۰]:

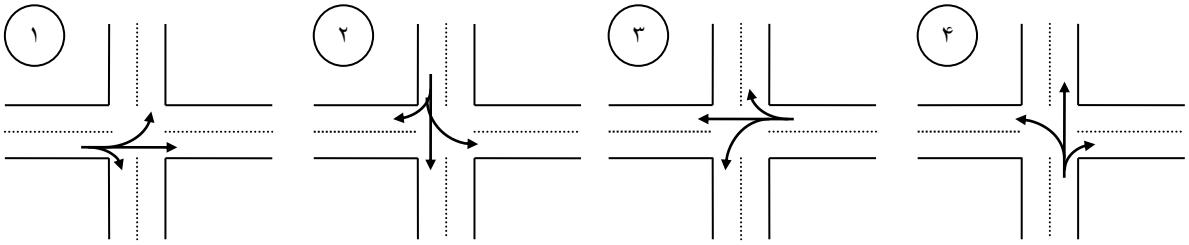
- سرعت مطلوب راننده: سرعتی است که راننده تمایل دارد در طول سفر خود با آن حرکت نماید. مقدار این سرعت از یک تابع توزیع گوسین با میانگین $۱۱۰\frac{Km}{h}$ و انحراف از معیار $۱۰\frac{Km}{h}$ انتخاب می شود.
- میزان تبعیت از حداکثر سرعت مجاز خیابان ها: مقدار آن به تصادف برای هر خودرو از یک تابع توزیع گوسین با میانگین ۱,۱ و انحراف از معیار ۰,۱ انتخاب می شود.
- آستانه تحمل راننده: هنگامی که خودرو در موقعیتی قرار دارد که حق تقدم عبور با خودروهای دیگر است، حداکثر

---

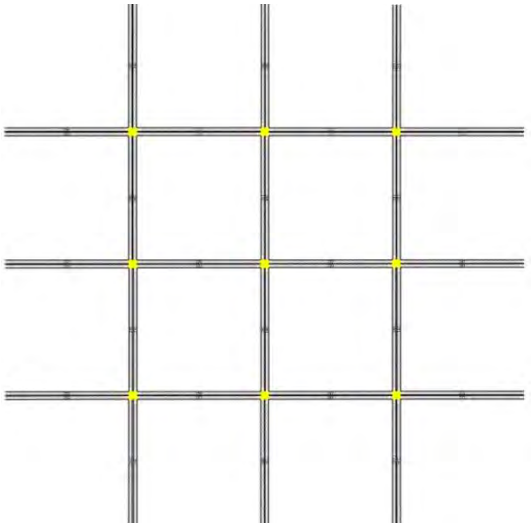
۱ Macro
۲ Micro

به اندازه یک بازه زمانی مشخص منتظر می ماند و بعد از آن بازه زمانی در صورتی که فضای مناسب برای عبور پیدا نکند اقدام به عبور از فضاهای کوچک و غیر ایمن میان خودروهای دیگر می کند. مقدار آستانه تحمل برای هر راننده به تصادف از یک تابع توزیع گوسین با میانگین ۱۰ ثانیه و انحراف از معیار ۲,۵ ثانیه انتخاب می شود.

- زمان واکنش راننده[۱]: 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[۲]: 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.
- فاکتور حساسیت: هنگامی که خودرو میخواهد سرعت خود را بخاطر محدودیت اعمال شده توسط خودروی جلویی کاهش دهد نیاز دارد که شتاب کاهشی خودرو جلویی را تخمین بزند. میزان درستی تخمین شتاب خودروی جلویی توسط خودروی تعقیب کننده فاکتور حساسیت می گویند که مقدار آن برابر ۱ در نظر گرفته شد که بیان کننده این است که خودروی تعقیب کننده شتاب خودروی جلویی را به درستی تخمین می زند [۳۱].



شکل ۲- ترتیب فازها



شکل ۳- شبکه ترافیکی متشکل از ۹ تقاطع

سرعت حرکت یک خودرو براساس چهار فاکتور سرعت مطلوب راننده، حداکثر سرعت مجاز خیابان ها، میزان تبعیت از حداکثر سرعت مجاز خیابان ها و سرعت حرکت خودروی جلویی تعیین می شود. به عنوان مثال فرض نمایید که سرعت مطلوب یک راننده ۱۰۰ کیلومتر بر ساعت و حداکثر سرعت مجاز خیابان نیز ۵۰ کیلومتر بر ساعت و میزان تبعیت از حداکثر سرعت مجاز ۱,۳ باشد. حداکثر سرعت حرکت خودروی فرضی برابر V=min(100,1.3*50)=65km/h خواهد شد. اما باید توجه نمود که سرعت ۶۵ کیلومتر بر ساعت حداکثر سرعتی است که یک خودروی فرضی با مشخصات داده شده می تواند برود اما اگر خودروی جلویی آن دارای سرعت کمتری باشد ناچار به کاهش سرعت و یا سبقت گرفتن است.

۴-۱-۳- چراغ های راهنمایی به عنوان عامل های یادگیر

چراغ های راهنمایی در هر تقاطع در ابتدای هر فاز وضعیت ترافیکی تقاطع (حالت محیط) را بررسی نموده و بر اساس دانش کسب شده از محیط مدت زمان سبز بودن آن فاز را مشخص می کنند. در انتهای هر فاز بر اساس تعداد ماشین های عبوری از هر تقاطع مشخص می شود که آیا مدت زمان سبز مناسب بوده یا خیر؟ چراغ های

۱ Driver Reaction Time
۲ Reaction Time of Stop

عملگر- نقاد برای یادگیری و تجربه اندوزی خود از محیط استفاده می کنند.

وضعیت ترافیکی هر تقاطع شامل یک بردار است که هر المان آن تعداد خودروها در مسیرهای ورودی به تقاطع را نشان می دهد. در سناریوی اول (تقاطع منفرد) تعداد خودروها در هر مسیر ورودی به تقاطع در بازه [۱۶۰ و ۰] قرار دارند اما در سناریوی دوم (۹ تقاطع) به دلیل کوتاه بودن طول خیابان ها تعداد خوردوها در هر مسیر ورودی در بازه [۸۰ و ۰] قرار دارند. مهمترین مزیت این تعریف این است که بار ترافیکی به نوعی در تعریف وضعیت محیط کد می شود. دیگر مزیت این تعریف، مدیریت کردن خودروهای پرسرنشین است. به عنوان مثال می توان به اتوبوس ها یا خودروهای پرسرنشین ضریب بالاتری را نسبت داد و با این روش اهمیت بیشتری به وسایل نقلیه عمومی داده می شود. همچنین شماره فاز جاری چراغ راهنمایی به عنوان یک المان دیگر در حالت محیط گنجانده می شود. در معماری عملگر – نقاد حالت های محیط باید به صورت گسسته لحاظ شوند بنابراین در سناریوی اول تعداد خودروها در هر مسیر ورودی به دسته های خطی با طول ۱۰ یعنی [۱۵۰-۱۴۰-۱۳۰-۱۲۰-۱۱۰-۱۰۰-۹۰-۸۰-۷۰-۶۰-۵۰-۴۰-۳۰] و در سناریوی دوم تعداد خودروها در هر مسیر ورودی به دسته های خطی با طول ۱۵ یعنی [۶۵-۵۰-۳۵-۲۰-۵] افراز شدند. علت تفاوت طول دسته ها در سناریوی اول با سناریوی دوم صرفاً اشغال حافظه[۱] کمتر می باشد.

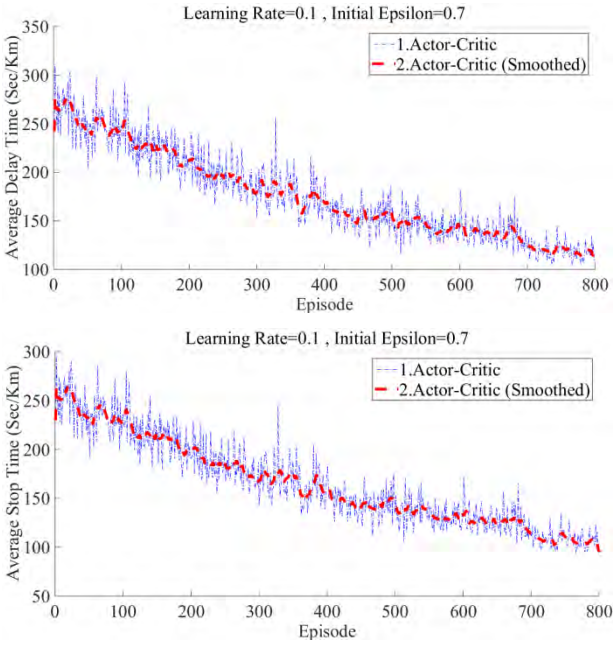
مقادیر {۹۰ و۸۰ و۷۰ و۶۰ و۵۰ و۴۰ و۳۰ و۲۰} ثانیه به عنوان اعمال عامل (مدت زمان سبز بودن هر فاز) در نظر گرفته شدند. جدول Q در سناریوی اول دارای $۱۴^۴×۴$ سطر و ۹ ستون و در سناریوی دوم برای هر عامل دارای $۶^۴×۴$ سطر و ۹ ستون می باشد.

اختلاف تعداد خودروها قبل و بعد از هر فاز در همه مسیرهای ورودی به عنوان مکانیزم پاداش دهی در نظر گرفته شده است. نرخ یادگیری در نقاد برابر ۰,۲ و در عملگر برابر با ۱۰۰ انتخاب شد. سیاست مورد استفاده در هر دو سناریو ε-greedy می باشد و مقدار ε مطابق رابطه ۸ در طول یادگیری تغییر می کند. عامل در ابتدای یادگیری تمایل بیشتری برای عملکرد تصادفی دارد و در طول یادگیری سیاست آن به سمت حریصانه میل می‌کند.

$$\varepsilon_{t+1} = \frac{\varepsilon_t}{1.0036}, \quad \varepsilon_0 = 0.7 \qquad (۸)$$

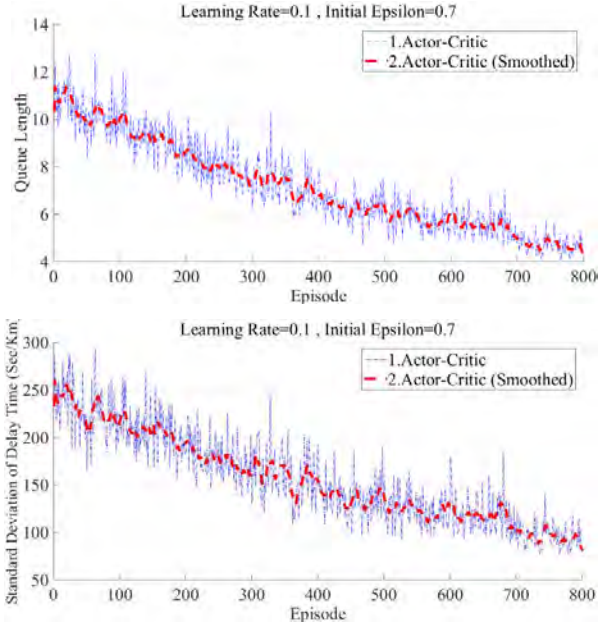
۵- نتایج

در این تحقیق معیارهای متوسط زمان تاخیر[۲]، متوسط زمان توقف[۳]، متوسط طول صف[۴] و انحراف از معیار زمان تاخیر برای بررسی عملکرد معماری عملگر – نقاد لاندا بکار گرفته شدند. معیار اول نشان دهنده زمان تاخیر برای هر ماشین در هر کیلومتر می باشد که این مقدار از اختلاف بین زمان سفر در شرایط ایده آل و بدون ترافیک و زمان سفر در شرایط ترافیکی محاسبه می شود، معیار دوم نشان دهنده متوسط زمان توقف هر ماشین در هر کیلومتر است، معیار سوم نشان دهنده متوسط تعداد ماشین هایی که در هر خط در هر مسیر ورودی قرار گرفته اند و معیار چهارم مشخص کننده عدالت و مساوات الگوریتم بین خودروها است. شکل ۴ عملکرد چراغ راهنمایی هوشمند را برای چهار معیار فوق در طول ۸۰۰ ساعت شبیه سازی نشان می دهد. خطوط نازک از میانگین گیری نتایج ۵ بار پیاده سازی و خطوط ضخیم از میانگین گیری خطوط نازک در بازه های ۱۱ ساعته بدست آمده اند.



۱ RAM

۲ Delay Time
۳ Stop Time
۴ Queue Length



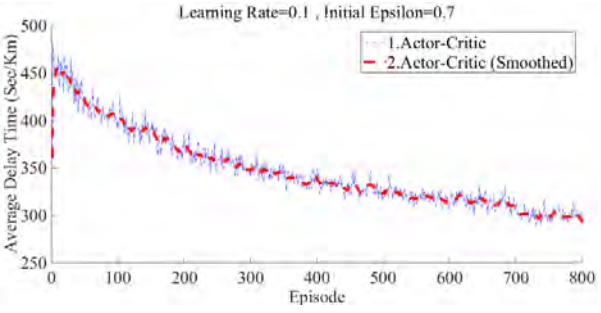
شکل ۴- عملکرد چراغ راهنمایی در طول روند یادگیری در سناریوی اول

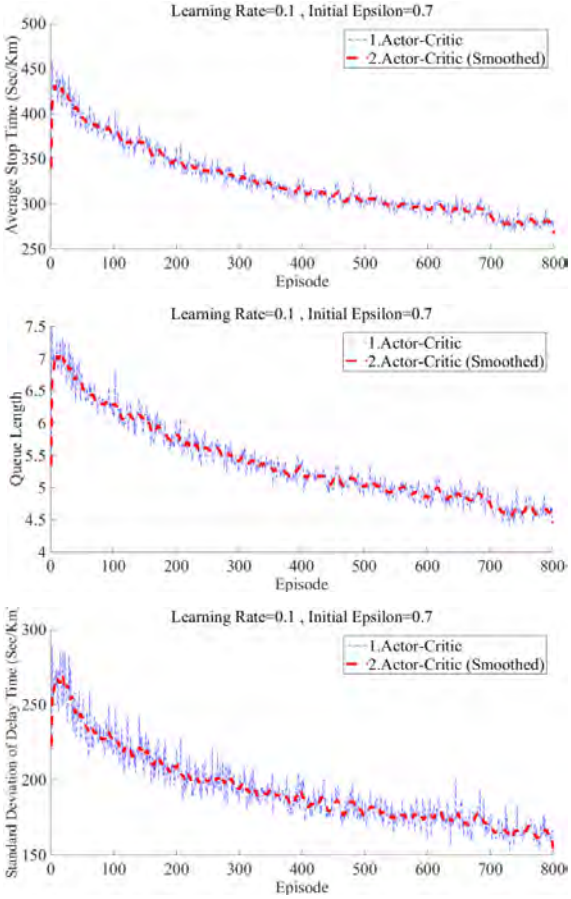
برای ارائه بهتر نتایج در جدول ۱ متوسط مقادیر زمان تاخیر، زمان توقف، طول صف و انحراف از معیار زمان تاخیر برای ۱۰۰ اپیزود انتهایی (۷۰۰ تا ۸۰۰) که در آنها عامل به صورت حریصانه عمل می نماید آورده شده است.

جدول ۱- عملکرد چراغ راهنمایی در اپیزودهای ۷۰۰ تا ۸۰۰ در سناریوی اول

متوسط زمان تاخیر (ثانیه/کیلومتر)	۱۱۸,۶۳۶
متوسط زمان توقف (ثانیه/کیلومتر)	۱۰۸,۱۱۶
انحراف از معیارزمان تاخیر (ثانیه/کیلومتر)	۱۱۸,۶۳۶
متوسط طول صف برای مسیر ورودی غرب به شرق	۴,۹۱۷
متوسط طول صف برای مسیر ورودی جنوب به شمال	۴,۶۴۶
متوسط طول صف برای مسیر ورودی شرق به غرب	۴,۷۸۰
متوسط طول صف برای مسیر ورودی شمال به جنوب	۴,۵۷۹

شکل ۵ عملکرد کلی ۹ چراغ راهنمایی هوشمند را برای ۸۰۰ ساعت شبیه سازی نشان می دهد (سناریوی دوم). خطوط نازک از میانگین گیری نتایج ۵ بار پیاده سازی و خطوط ضخیم از میانگین گیری خطوط نازک در بازه های ۱۱ساعته بدست آمده اند.





شکل ۵- عملکرد ۹ چراغ راهنمایی در طول روند یادگیری در سناریوی دوم

برای ارائه بهتر نتایج متوسط مقادیر زمان تاخیر، زمان توقف، طول صف و انحراف از معیار زمان تاخیر برای ۱۰۰ اپیزود انتهایی (۷۰۰ تا ۸۰۰) که عامل ها در آن بازه به صورت حریصانه عمل می کنند در جدول ۲ آورده شده است.

جدول ۲- عملکرد ۹ چراغ راهنمایی در اپیزودهای ۷۰۰ تا ۸۰۰ در سناریوی دوم

متوسط زمان تاخیر (ثانیه/کیلومتر)	۳۰۰,۴۹۲
متوسط زمان توقف (ثانیه/کیلومتر)	۲۷۹,۴۵۵
متوسط طول صف	۴,۶۳۸
انحراف از معیارزمان تاخیر (ثانیه/کیلومتر)	۱۶۵,۱۸۱

۶- بحث و اعتبار سنجی نتایج

به منظور اعتبار سنجی روش ارائه شده، نتایج آن با روش رایج زمان بندی در مهندسی ترافیک: چراغهای پیش زمانبندی شده مقایسه شدند. چراغ های پیش زمان بندی شده، چراغ هایی هستند که زمانبندی از پیش تعیین شده و معلومی را بدون توجه به تغییرات شرایط واقعی ترافیک لحاظ می کنند. شکل ۶ عملکرد روش زمان

ثابت، را با یادگیری تقویتی عملگر- نقاد بر اساس شاخص های مختلف مقایسه می کند.

شکل ۶- مقایسه عملکرد روش زمان ثابت و یادگیری تقویتی عملگر- نقاد برای سناریوی های اول و دوم

در شکل ۶، زمان سفر عبارت است از میانگین زمانی که یک خودرو نیاز دارد تا یک کیلومتر را در شبکه طی نماید. متوسط زمان سفر در هر ساعت از متوسط گیری زمان سفر خودروها برای ساعت های ۷۰۰ تا ۸۰۰ بدست می آید.

متوسط سرعت نشان دهنده متوسط سرعت کلیه خودروها در طول مسیر حرکتشان می باشد. همانطور که مشخص است تمامی شاخص ها بهبود یافته اند.

۷- نتیجه گیری

افزایش تقاضا برای جابه جایی در جوامع بشری باعث ایجاد چالش های متعددی در مهندسی ترافیک شده است. در اغلب مواقع اضافه نمودن زیر ساخت های جدید (به عنوان مثال احداث خیابان های جدید) همواره ممکن نبوده و استفاده بهینه تر از زیر ساخت های حمل و نقل

موجود احساس می شود. کنترل و مدیریت ترافیک به دلیل ذات توزیع یافتگی آن ارتباط نزدیکی با مفاهیم و اصول سیستم های چند عامله دارد زیرا به عنوان مثال خودروها، عابرین پیاده و چراغ های راهنمایی را می توان به عنوان عامل های خودمختار در نظر گرفت. استفاده از سیستم های چند عامله در کنترل ترافیک همواره با چالش های فراوانی روبه رو هستند از جمله اینکه عامل ها به تغییرات در محیط در محدوده دیدشان واکنش نشان می دهند که همین امر منجر به الگوهای ترافیکی متفاوت می شود. بنابراین استفاده از روش های رایج و اولیه در سیستم های چند عامله لزوماً به نتایج مطلوبی منجر نخواهد شد. در این تحقیق سعی شده است که از یادگیری تقویتی برای حل چالش فوق به گونه ای استفاده شود که عامل های یادگیر (چراغ های راهنمایی) به تغییرات ترافیکی در محدوده عملکردشان واکنش مناسب را که از تجربیات قبلی بدست آورده اند نشان دهند. نتایج این تحقیق نشان داد که کنترل هوشمند چراغ های راهنمایی منجر به کاهش متوسط طول صف، زمان سفر در مقایسه با روش های غیر هوشمند شده است.

مراجع

[1] J. H. Holland, (1992). "Complex Adaptive Systems." Daedalus. Vol. 121, No. 1, PP. 17-30.

[2] S. M. Manson, (2001). "Simplifying complexity: a review of complexity theory." Geoforum. Vol. 32, No. 3, PP. 405-414.

[3] M. Wooldridge, (2009)."An Introduction to MultiAgent Systems - Second Edition." London: John Wiley & Sons.

[4] R. Itami, R. Raulings, G. MacLaren, K. Hirst, R. Gimblett, D. Zanon, and P. Chladek, (2003). "RBSim 2: simulating the complex interactions between human movement and the outdoor recreation environment." Journal for Nature Conservation. Vol. 11, No. 4, PP. 278–286.

[5] D. A. Bennett and W. Tang, (2006). "Modelling adaptive, spatially aware, and mobile agents: Elk migration in Yellowstone." International Journal of Geographical Information Science. Vol. 20, No. 9, PP. 1039-1066.

[6] R. Shad, M. S. Mesgari, H. Ebadi, A. Alimohammadi, A. Abkar, and A. Vafaeenezhad, (2009). "An Intelligent Fuzzy Agent for Spatial Reasoning in GIS." Advances in Artificial Intelligence. Vol. 5803, No. PP. 639-647.

[7] S. Behzadi and A. A. Alesheikh, (2013). "Introducing a novel model of belief-desire-intention agent for urban land use planning." Engineering Applications of Artificial Intelligence. Vol. 26, No. 9, PP. 2028-2044.

[8] S. Behzadi and A. A. Alesheikh, (2013). "Hospital Site Selection Using a BDI Agent Model." International Journal of Geography and Geology. Vol. 2, No. 4, PP. 36-51.

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

[10] V. R. Konda and J. N. Tsitsiklis, (2003). "On Actor-Critic Algorithms." SIAM Journal on Control and Optimization. Vol. 42, No. 4, PP. 1143–1166.

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

[12] D. I. Robertson, "TRANSYT: A traffic network study tool," Road Research Laboratory Report London1969.

[13] 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.

[14] 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

[15] 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.

[16] J. J. Henry, J. L. Farges, and J. Tufal, (1983). "The PRODYN real-time traffic algorithm." in Proceedings of the 5th IFAC/IFIP/IFORS Symposium on Control in Transportation Systems, Baden-Baden, Germany.

[17] 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.

[18] 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.

[19] M. Wiering, (2000). "Multi-agent reinforcement learning for traffic light control." presented at the 17th International Conference on Machine Learning, Stanford,CA.

[20] B. Abdulhai, R. Pringle, and G. J. Karakoulas, (2003). "Reinforcement learning for true adaptive traffic signal control." Journal of Transportation Engineering. Vol. 129, No. 3, PP. 278–285.

[21] E. Camponogara and W. J. Kraus, (2003). "Distributed Learning Agents in Urban Traffic Control." in Proceedings of the 11th Portuguese Conference on Artificial Intelligence Beja, Portugal, pp. 324-335.

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

[23] L. Bull, J. Sha'Aban, A. Tomlinson, J. D. Addison, and B. G. Heydecker, (2004). "Towards Distributed Adaptive Control for Road Traffic Junction Signals using Learning Classifier Systems." in Applications of Learning Classifier Systems. vol. 150, L. Bull, Ed., ed Berlin Heidelberg: Springer PP. 276-299.

[24] K. Dresner and P. Stone, (2004). "Multiagent traffic management: A reservation-based intersection control mechanism." in Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems, New York, pp. 530–537.

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

[26] D. Houli, L. Zhiheng, and Z. Yi, (2010). "Multiobjective Reinforcement Learning for Traffic Signal Control Using Vehicular Ad Hoc Network." EURASIP Journal on Advances in Signal Processing. Vol. 2010, No. PP. 7-17.

[27] M. v. Otterlo and M. Wiering, (2012). "Reinforcement Learning and Markov Decision Processes." in Reinforcement Learning State-of-the-Art, M. Wiering and M. v. Otterlo, Eds., ed Berlin Heidelberg: Springer Berlin Heidelberg, PP. 3-42.

[28] A. Reuschel, (1950). "Vehicle movements in a platoon with uniform acceleration or deceleration of the lead vehicle." Zeitschrift des Oesterreichischen Ingenieur-und Architekten-Vereines. Vol. 95, No. PP. 59-62 and 73-77.

[29] R. Tao, H. Wei, Y. Wang, and V. Sisiopiku, (2005). "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, No. PP. 83-93.

[30] S. Moridpour, M. Sarvi, and G. Rose, (2010). "Modeling the lane changing execution of multi class vehicles under heavy traffic conditions. " Transportation Research Record. No. 2161, PP. 11-19.

[31] P. G. Gipps, (1981). "A behavioural car-following model for computer simulation." Transportation Research Part B: Methodological. Vol. 15, No. 2, PP. 105–111.