# کنترل ترافیک شهری با استفاده از سیستمهای چندعامله سلسله مراتبی و یادگیری تقویتی

**منیره عبدوس**[1]
دانشکده مهندسی و علوم کامپیوتر- دانشگاه شهید بهشتی

## چکیده

با افزایش تعداد خودروها در شبکه های شهری، روشهای کلاسیک در کنترل ترافیک شهری قابل استفاده نبوده و نیاز به روشهای هوشمند افزایش می‌یابد. در این مقاله، روشی برای کنترل ترافیک شهری با استفاده از سیستمهای چندعامله سلسله مراتبی دو سطحی پیشنهاد شده است. یک شبکه شامل تعداد زیادی از تقاطعها، توسط یک سیستم چندعامله در سطح اول مدل شده است. در سطح دوم تعدادی عامل که هر یک مسئول کنترل ترافیک ناحیه ای متشکل از تعدادی عامل هستند، قرار دارند. در سطح اول، یادگیری تقویتی برای زمانبندی چراغهای راهنمایی مورد استفاده قرار گرفته است و در سطح دوم از یک سیستم مبتنی بر قانون جهت کنترل تراکم خودروها در ناحیه استفاده شده است. نتایج حاصل از پیاده سازی روش بر روی شبکه گرید نشان می‌دهند که روش پیشنهادی باعث کاهش میزان تاخیر در زمان سفر شده و همچنین از اشباع شبکه جلوگیری می‌نماید.

***کلید واژه :*** ***کنترل ترافیک هوشمند، سیستمهای چندعامله، ساختار سلسله مراتبی، یادگیری تقویتی.***

[1] استادیار دانشکده مهندسی و علوم کامپیوتر، m_abdoos@sbu.ac.ir

## ۱- مقدمه

یکی از اساسی ترین زیرساختهای لازم برای توسعه صنایع و افزایش سطح رفاه اجتماعی در هر کشور، سیستم حمل و نقل می‌باشد. امروزه با افزایش روز افزون حجم ترافیک خودروها در شبکه های شهری، مدت زمان تلف شده و به همراه آن آلودگی محیط زیست و مصرف انرژی نیز افزایش یافته است. روند رشد سریع تقاضای حمل و نقل به ویژه در ساعات اوج در کلان شهرها، از جمله مشکلات اصلی در بسیاری از شهرهای دنیا محسوب می‌شود. از این رو تامین حمل و نقل ایمن و کارا یکی از مهم ترین مسائل مطرح در جوامع مختلف می‌باشد. در سالهای اخیر استفاده از روشها و تکنیکهای مختلف هوش مصنوعی در حوزه سیستم های حمل و نقل هوشمند به طور گسترده مورد توجه قرار گرفته است. از آنجا که مفهوم عاملهای هوشمند بر بخشهای مختلفی از سیستم مانند چراغ راهنمایی، خودروها و عابران پیاده انطباق دارد، لذا بسیاری از روشهای مطرح در این حوزه از تکنیکهای مبتنی بر عامل استفاده می‌کنند.

در این مقاله، روشی جهت کنترل چراغهای راهنمایی مبتنی بر سیستمهای چندعامله ارائه می‌شود. در این روش یک ساختار سلسله مراتبی دولایه برای کنترل ترافیک پیشنهاد شده که عاملها در سطح اول از یک مدل یادگیری تقویتی جهت یافتن خط مشی بهینه استفاده می‌کنند و عاملها در سطح دوم به عنوان ناظر عاملهای سطح اول به مدیریت زمانبندی در هر ناحیه می‌پردازند. بنابراین روش ارائه شده در این مقاله به حوزه کنترل ترافیک با استفاده از زمانبندی تقاطع ها مربوط می‌باشد.

تحقیقات بسیاری در زمینه حمل و نقل هوشمند با استفاده از سیستمهای چندعامله ارائه شده است که جهت بررسی آنها می‌توان به [۱] ارجاع نمود. در این بخش به طور اجمالی به بررسی برخی از کارهای انجام شده در راستای کنترل ترافیک با استفاده از سیستمهای چندعامله و مبتنی بر یادگیری تقویتی می‌پردازیم .

در [۲] سیستم کنترل ترافیک مبتنی بر عامل ارائه شده است که با به کارگیری قوانین داخلی قادر است در برابر تغییرات محیطی خود را تطبیق دهد. سیستم ارائه شده دارای چندین عامل جهت کنترل تقاطع، قطعه های یک خیابان و همچنین تعدادی عامل مسئول می‌باشد. عاملهای تقاطع با کمک عاملهای قطعه های خیابانها به مدیریت زمانبندی در تقاطع ها می‌پردازند. عاملهای مسئول کار هماهنگی و کنترل عاملهای ذکر شده را در جهت رسیدن به بهینه سراسری را بر عهده دارند. در [۳] سیستمی متشکل از چهار نوع عامل، تقاطع، خیابان، قطعه ها خیابانها، کنترل مرکزی ارائه شده است. نشان داده شده است که سیستم ارائه شده با مکانیزم اشتراک داده ها، در عین حال که باعث کاهش حجم ترافیک و مصرف سوخت می‌شود، به بهینه سراسری در هر ناحیه نیز دست می‌یابد. عامل کنترل مرکزی که در بالای سلسله مراتب قرار دارد، تصمیمات عاملها را تحلیل کرده و سیاستهای مدیریتی را اعمال می‌نماید. این روش از این جهت که از ساختار سلسله مراتبی جهت کاهش پیچیدگی سیستم و رسیدن به بهینه سراسری استفاده کرده است اهمیت زیادی دارد، اما چون روش کنترل مرکزی در راس سلسله مراتب قرار دارد، معایب کنترل متمرکز از جمله پیچیدگی طراحی و عدم تحمل پذیری خطا را دارا می‌باشد.

تجزیه سلسله مراتبی سیستم به زیر سیستمها در مقالات دیگری نیز مورد استفاده قرار گرفته است. به عنوان مثال در [۴] و [۵] سیستم چندعامله سلسله مراتبی ارائه شده که متشکل از سه لایه می‌باشد. در پائین ترین سطح، عاملهای کنترل کننده تقاطع قرار دارند، عاملهای کنترل کننده یک ناحیه در سطح میانی و عامل کنترل کننده نواحی در بالاترین سطح قرار دارد. سیستم ارائه شده مبتنی بر شبکه های عصبی و تئوری فازی می‌باشد، همچنین دارای بخشهای یادگیری بلادرنگ شامل یادگیری تقویتی، نرخ یادگیری و تنظیم وزن یالهای شبکه، به روز رسانی پویای روابط فازی با استفاده از الگوریتمهای تکاملی می‌باشد و امکان تطبیق با تغییرات محیطی را برای عامل فراهم می‌آورد. بدیهی است به کار بردن روشهای ترکیبی مزایای روشهای به کارگرفته شده را در راستای رسیدن به هدف ترکیب می‌نمایند، اما طراحی و پیاده سازی بخشهای مختلف و در کنار هم قرار دادن آنها پیچیده است و هزینه محاسباتی بالایی دارد .

روشهای یادگیری از دیگر روشهایی است که در سیستمهای چندعامله مطرح شده و در کنترل و مدیریت ترافیک قابل استفاده است. یادگیری تقویتی مزایای قابل توجهی در محیطهای پویا دارد و به همین دلیل به عنوان یکی از راهکارهای موفق در کنترل یک تقاطع و یا شبکه ای از آنها که دانش قبلی از محیط موجود نیست مورد استفاده قرار می‌گیرد. استفاده از یادگیری تقویتی جهت کنترل ترافیک، این امکان را فراهم می‌سازد که راهکار پیشنهادی تا حد امکان وابسته به دانش خبره نباشد و با توجه به تاریخچه آماری اجزای شبکه، زمان‌بندی مناسب برای چراغ‌های راهنمایی و رانندگی تنظیم شوند .

روشهای یادگیری تقویتی به دو دسته مبتنی بر مدل و مستقل از مدل تقسیم می‌شوند. از روشهای مبتنی بر مدل در کنترل ترافیک می‌توان به روشی که در [۶] آورده شده است، اشاره نمود که در آن عاملها از تعامل مستقیم با محیط سعی می‌کنند جریان ترافیک را بهینه نمایند. اگرچه نشان داده شده است که این روش بهتر از روش زمانبندی ثابت عمل کرده است، اما ارزیابی روش تنها برای یک تقاطع صورت گرفته است و چنانچه تعداد تقاطع بیشتر شود، تغییرات محیطی بیشتر شده و تضمینی بر عملکرد این روش یادگیری وجود ندارد. در [۷] نیز یادگیری تقویتی برای کنترل ترافیک در یک شبکه گرید مورد استفاده قرار گرفته است. در این روش از نوعی یادگیری همکار استفاده شده است که به طور همزمان خط مشی جهت کنترل سیگنالهای ترافیکی و مسیریابی بهینه صورت می‌گیرد. یکی از مشکلات این روش مربوط به هزینه بالای ارتباطات و دانش می‌باشد، به خصوص زمانی که تعداد تقاطع های شبکه افزایش می‌یابد. روش یادگیری مشابهی که مبتنی بر مدل می‌باشد نیز در [۸] معرفی شده است. هدف یادگیری طوری مدل شده است که نمایش حالت، مبتنی بر مجموع زمان انتظار خودروها در اطراف تقاطع می‌باشد. بدیهی است هرچه اطلاعات بیشتری از خودروها دریافت شود، مدل پیچیده تر و فضای حالت بزرگتر خواهد بود، این مساله برای شبکه های بزرگ یک مشکل اساسی محسوب می‌شود.

از آنجا که روشهای مبتنی بر مدل در محیطهای ایستا قابل استفاده هستند، قابلیت تطبیق با تغییرات در محیطهای پویا را ندارند. در محیطهای پویا و غیر ایستا، روشهای یادگیری تقویتی مستقل از محیط قابل استفاده می‌باشد. در

[۹] روشی برای دسته مشخصی از محیطهای غیرایستا بیان شده است که قادر است مدل ناقصی از محیط را تخمین بزند. این روش در شبکه ای متشکل از ۹ تقاطع مورد استفاده قرار گرفته است. نتایج گزارش شده نشان می‌دهد که با فرض برخی خصوصیات، این روش کارایی بهتری نسبت به روشهای مستقل از مدل و مبتنی بر مدل دارد، اما وجود این مفروضات باعث می‌شود این روش قابل تعمیم و استفاده در شبکه های بزرگتر نباشد.

یادگیری تقویتی تطبیقی در مقالات [۱۰] و [۱۱] و یادگیری تقویتی موازی [۱۲] جهت کنترل محیط مستقل از مدل ارائه شده است. کنترل تطبیقی در [۱۳] نیز معرفی شده است که از تقریب تابع به عنوان نگاشتی از حالات و زمان بندی استفاده می‌نماید.

اگرچه روشهای یادگیری تقویتی مستقل از مدل در محیطهای غیر ایستا که مدلی از محیط در دسترس نیست قابل استفاده می‌باشد، اما برخی روشها مانند [۱۴] در یک تقاطع و برخی دیگر مانند روشهای [۱۲] و [۱۳] شبکه های گرید منظم و یا خطی را در نظر گرفته اند. این روشها به دلیل افزایش نمایی حجم محاسبات در شبکه های بزرگتر و غیر منظم قابل پیاده سازی نیستند.

از میان روشهای موجود، برخی محققان روشی ارائه کرده اند که قابل اجرا بر روی شبکه های بزرگ می‌باشد. یادگیری تقویتی همکار، سعی به استخراج دانش از عاملهای همسایه در راستای یادگیری زمانبندی دارند، این روش که در [۱۵] معرفی شده است در قسمتی از شهر دوبلین شامل ۶۴ تقاطع پیاده سازی شده است. در مقاله دیگری نیز ۲۹ تقاطع با استفاده از یادگیری تقویتی و اطلاعاتی که میان عاملها رد و بدل می‌شود، کنترل می‌گردند [۱۴].

در [۱۶] نیز مساله کنترل چراغ‌های راهنمایی و رانندگی با هدف استفاده بهینه از شبکه ترافیک شهری مورد بررسی قرار گرفته است. رویکرد ما این مقاله تنها از نظر استفاده از ساختار سلسله مراتبی شبیه به رویکرد این تحقیق می‌باشد. ساختاری که به عنوان راه‌حل در مقاله مذکور ارائه شده است، شامل سه سطح از عامل‌ها بوده که هر سطح وظیفه متفاوتی را بر عهده دارد. در پایین‌ترین سطح این سلسله مراتب، هر عامل سعی در بهینه کردن ترافیک در یک تقاطع دارد. این بهینه‌سازی محلی بوده و ممکن است در یک بهینه محلی به دام بیافتد. از این رو، در یک سطح بالاتر، عامل‌هایی سعی در برقراری همکاری میان عامل‌های لایه پایین‌تر می‌نمایند تا از این طریق از دام بهینه‌های محلی خارج شوند. در این سطح، هر عامل ناظر بر تعدادی از عامل‌های لایه‌ی پایین تر است. در شرایطی هم که شبکه به اندازه‌ی کافی بزرگ باشد، یک عامل در سطح سوم، به عنوان ناظر کلی در نظر گرفته می‌شود که بر عامل‌های لایه دوم نظارت می‌کند. علاوه بر این مجموعه از عامل‌ها، عامل یکتای دیگری نیز نقش به اشتراک گذاشتن اطلاعات ترافیکی را بین عامل‌های مختلف محیط ایفا می‌کند. در این مقاله با وجود استفاده از یک ساختار سلسله مراتبی برای حل مسئله، محیط شبیه‌سازی ترافیک به صورت کلان پیاده‌سازی شده و همین امر مقایسه نتایج آن را با این مقاله بی‌معنی می‌سازد.

اکثر روش‌های مبتنی بر سیستمهای چند عامله که برای کنترل ترافیک با استفاده از زمان‌بندی چراغ‌های راهنمایی پیشنهاد شده‌اند ممکن است در یک شبکه ترافیکی کوچک متشکل از تعداد محدود و انگشت شماری از تقاطع‌ها،

جواب‌های قابل قبول و گاهی نزدیک به بهینه‌ای را تولید کنند، اما با بزرگتر شدن شبکه ترافیکی کارآیی خود را به شدت از دست می‌دهند وکیفیت جواب‌هایی که تولید می‌کنند از راه‌حل‌های ساده نیز ممکن است پایین‌تر شود. این امر بدان علت است که در یک شبکه بزرگ و در حالت‌های واقعی ما تعداد بسیار زیادی از تقاطع ها و اجزا ترافیکی را داریم که به صورت پویا با یکدیگر در تعامل هستند. این تعداد از اجزا و یا به عبارت بهتر عامل‌ها نیازمند این هستند که برای ارائه یک راه‌حل قابل قبول با یکدیگر ارتباط داشته باشند. حال آن‌که کنترل یک چنین شبکه ارتباطی بسیار پیچیده بوده و سربارهای محاسباتی فراوانی را به همراه می‌آورد و باعث می‌شود در دنیای واقعی و در محیط‌های بلادرنگ نتوان از آن‌ها استفاده کرد. از سویی دیگر، ارتباطات بین عامل‌ها در شبکه ترافیکی را نیز نمی‌توان نادیده گرفت و چنین ارتباط ضمنی و در عین حال واقعی، در تصمیماتی که جهت کنترل آن گرفته می‌شود مهم بوده و روی نتیجه به شدت تاثیر می‌گذارند. در چنین محیط‌هایی که ما با تعداد زیادی از عامل‌ها سر و کار داریم و حجم وسیعی از ارتباطات و تعاملات لازم بین آنها را شاهد هستیم، ساختارهای سلسله مراتبی به طور قابل توجهی می‌توانند در ساده‌سازی مساله، کاهش حجم محاسبات و حذف ارتباطات غیر ضروری و در نتیجه افزایش کارایی و مقاومت سیستم یاری رسانند.

در این مقاله، روشی برای کنترل چراغهای راهنمایی با استفاده از یادگیری تقویتی ارائه می‌شود. شبکه ای بزرگ متشکل از ۳۶ تقاطع که در یک شبکه گرید در کنار یکدیگر قرار گرفته اند، با یک سیستم سلسله مراتبی دو لایه مدل می‌شود. به این صورت که هر یک از عاملهای لایه اول مسئول زمانبندی یک چراغ راهنمایی می‌باشند. شبکه به تعدادی نواحی تقسیم شده، که هر ناحیه توسط یک عامل که در سطح دوم قرار گرفته است کنترل می‌شود. در لایه اول عاملها از یادگیری تقویتی برای زمانبندی استفاده می‌کنند در حالی که عاملهای لایه دوم از یک سیستم مبتنی بر قانون جهت کاهش حجم ترافیک در ناحیه استفاده می‌نمایند .

در بخش بعدی، مقدمه ای کوتاه در مورد یادگیری تقویتی مورد استفاده در این مقاله که یادگیری Q می باشد شامل می شود. قسمتهای مختلف روش پیشنهادی در بخش ۳ معرفی می‌شوند. در بخش ۴، نتایج تجربی حاصل از اجرای الگوریتم بر روی شبکه ترافیک شهری گزارش می‌شود. در بخش ۵، نتیجه گیری و کارهای آینده بیان می‌گردند.

## ۲- یادگیری Q-

تاکنون روشهای بسیاری جهت شبیه سازی رفتار انسانها ارائه شده است. در میان این روشها، سیستمهای چندعامله مناسب ترین ابزار جهت مدل سازی سیستمهای مشابه انسان به حساب می‌آید. سیستمهای مبتنی بر عامل امکان در نظر گرفتن جنبه فردی و اجتماعی انسان را فراهم می‌سازند. در مدلهای مبتنی بر عامل، رفتار هر عامل بر اساس ویژگیهای مربوط به آن عامل تعریف می‌شود و به این ترتیب می‌توان تفاوتهای موجود در رفتارهای فردی عاملها را مدل سازی نمود. همچنین از آنجا که طراحی آن می‌تواند به صورت غیرمتمرکز انجام گیرد، امکان مدل سازی

عاملهایی با رفتار مبتنی بر هدف را نیز فراهم می‌آورند. سیستم رفتاری انسان بر مبنای یادگیری است، بنابراین می‌توان گفت یادگیری، مهمترین جز در سیستمهای چندعامله می‌باشد.

با توجه به نقش ناظر خارجی می‌توان روشهای یادگیری را به دو نوع روش یادگیری با ناظر و یادگیری بدون ناظر تقسیم بندی نمود. یادگیری با ناظر، یک روش عمومی در یادگیری ماشین است که در آن به یک سیستم، مجموعه نمونه های ورودی – خروجی ارائه شده و سیستم تلاش می‌کند تا تابعی از ورودی به خروجی را فرا گیرد. در یادگیری بدون ناظر، فرآیند یادگیری با استفاده از مجموعه نمونه های ورودی انجام می‌گیرد.

در یادگیری ماشین، مسائلی وجود دارند که در آنها منابع قابل استفاده برای حل مساله آنقدر کم و ناقص می‌باشند که امکان استفاده از الگوریتم های یادگیری با ناظر وجود ندارد. در برخی موارد حتی اطلاع دقیقی از آنچه باید یاد گرفته شود نیز در اختیار نیست و مجموعه داده آماده ای وجود ندارد تا یادگیری بدون ناظر روی آن صورت گیرد. در چنین شرایطی یادگیری تقویتی می‌تواند مفید واقع شود که یادگیری در آن به کمک تجارب صورت می‌گیرد.

در یادگیری تقویتی، سیستم تلاش می‌کند تا تعامل خود با یک محیط پویا را از طریق خطا و آزمایش بهینه نماید. یادگیری تقویتی راهی برای یادگیری رفتارها به کمک تعامل با محیط بدون داشتن هیچ ناظری می‌باشد. یادگیری تقویتی در واقع چگونگی نگاشت موقعیت های مختلف به اعمال برای حصول بهترین نتیجه یا بیشترین پاداش می‌باشد. در بسیاری موارد، اعمال نه تنها روی پاداش همان مرحله بلکه مراحل بعد هم تاثیرگذار است. این دو خصوصیت، یعنی «سعی و خطا» و «پاداش با تاخیر» مهمترین خصوصیات یادگیری تقویتی می‌باشند. یادگیری تقویتی از این رو مورد توجه است که راهی برای آموزش عاملها برای انجام یک عمل از طریق دادن پاداش و تنبیه است بدون اینکه لازم باشد نحوه انجام عمل را برای عامل مشخص نمائیم.

روش یادگیری-Q [۱۷]، یک روش مستقل از مدل است که در آن، عامل هیچ نوع دسترسی به مدل انتقال ندارد. این روش، یکی از بهترین و پرکاربردترین روشهای یادگیری تقویتی در حل مسائل یادگیری می‌باشد. در این روش، عامل ارزش انتخاب عمل a در حالت s که با $Q^*(s,a)$ نشان داده می شود را با استفاده از تعامل پیوسته با محیط و با سعی وخطا تخمین می‌زند. در این روش، عامل با مقادیر تصادفی از تخمین‌ها شروع کرده و بعد از هر عمل، چندتایی به صورت $<s,a,r,s'>$ را دریافت می‌کند که در آن s حالت فعلی؛ a عمل انجام شده در حالت s، r پاداش فعلی و $s'$ حالت بعد از اجرای a می‌باشد. عامل برای هر چندتایی می‌تواند ارزش حالت-عمل مربوطه را به صورت زیر محاسبه کند:

$$Q(s,a) = (1-\alpha) \times Q(s,a) + \alpha \times [r + \max_{a'} Q(s',a')] \qquad (۱)$$

که در آن، $\alpha \in [0,1]$نرخ یادگیری عامل است و مشخص می‌کند که تا چه حدی اطلاعات جدید بدست آمده، جایگزین اطلاعات قدیمی شوند. مقدار ۱ برای این نرخ سبب می‌شود که عامل فقط جدیدترین اطلاعات را در نظر گیرد و مقدار صفر باعث می‌گردد عامل یادگیری نداشته باشد؛ و $\gamma \in [0,1]$ که فاکتور کاهش نامیده می شود، برای مشخص کردن اهمیت پاداش‌های آینده است. مقدار صفر برای این فاکتور، عامل را فرصت طلب می‌کند-یعنی عامل

فقط پاداش فعلی را در نظر می‌گیرد. از سوی دیگر، نزدیکی به مقدار ۱ سبب می‌شود که عامل برای یک پاداش بالا در طولانی مدت منتظر بماند.

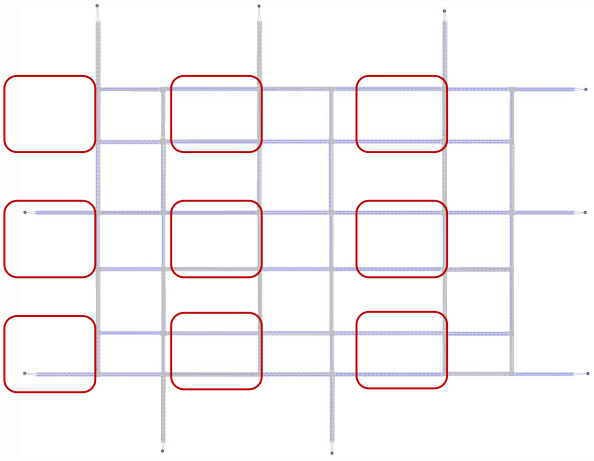
اثبات می‌شود که اگر تمامی زوج‌های حالت-عمل به صورت مکرر تجربه شوند و نرخ یادگیری در طول زمان کاهش یابد، یادگیری- Q با احتمال ۱ به مقدار بهینه‌ی $Q^*(s,a)$ همگرا می‌شود. با توجه به این امر، سیاست بهینه به صورت زیر تعریف می‌شود:

$$\pi^*(s) = \max_a Q^*(s,a) \qquad (۲)$$

در این مقاله از روش یادگیری-Q برای زمان‌بندی چراغ‌های راهنمایی و رانندگی توسط عامل‌های تعریف شده در هر تقاطع، استفاده می‌کنیم. جزئیات بیشتر در رابطه با نحوه تعریف فضای حالت و عمل برای کنترل ترافیک، در بخش بعدی به تفصیل شرح داده خواهد شد.

## ۳- کنترل ترافیک با استفاده از یادگیری-Q

همان‌گونه که پیش از این نیز اشاره شد، در راه حل پیشنهادی برای مسئله ترافیک، به هر یک از تقاطع‌ها یک عامل نسبت داده می‌شود و مجموعه‌ی تمامی این عامل‌ها، یک محیط چند عامله می‌سازند. این عامل‌ها با استفاده از حسگرهایی که در سطح خیابان‌های اطراف خود نصب گردیده‌اند، دانشی را از محیط دریافت کرده و اقدام به زمان‌بندی چراغ‌های راهنمایی و رانندگی نصب شده در تقاطع مربوط به خود می‌کنند. شبکه ترافیکی مورد نظر، شبکه ای متشکل از ۳۶ تقاطع و ۷۰ خیابان دوطرفه می‌باشد که در شکل ۱ نشان داده شده است. در این مقاله از یک ساختار سلسله مراتبی دو لایه استفاده شده است که لایه اول متشکل از ۳۶ عامل است که هر کدام مسئول کنترل یک تقاطع می‌باشند. شبکه به ۹ ناحیه تقسیم بندی شده است که هر یک دارای چهار تقاطع می‌باشد و به هر ناحیه عاملی در سطح دوم اختصاص داده شده است. تقاطع ها به صورت چهار راه و سه راه می‌باشند که در یک شبکه گرید قرار گرفته اند.

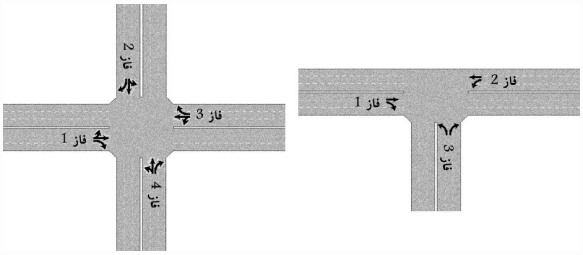


*شکل ۱- شبکه شهری مورد استفاده در شبیه سازی*

### ۳-۱- یادگیری تقویتی در سطح اول

در این شبکه عاملها همگن بوده و هرکدام کنترل سیگنال ترافیکی با تعدادی خیابان ورودی را بر عهده دارند. در این روش، به هر خیابان ورودی یک فاز اختصاص داده شده است. عاملها برای تخمین حالت از میانگین طول صف در خیابانهای ورودی استفاده می‌کنند، به این ترتیب که خیابانهای ورودی بر اساس این مقدار مرتب می‌شوند و در نتیجه تعداد حالتهای ممکن برابر با تعداد جایگشتهای ممکن خیابانهای ورودی می‌باشد.

فرض کنید عامل $\alpha^i$ دارای $k$ خیابان ورودی است که با $l_j$، $j=1,...,k$ نشان می‌دهیم. در این صورت، $l_1 \geq l_2 \geq ... \geq l_k$حالتی را نشان می‌دهد که در آن طول صف $l_1$ بیشترین مقدار و طول صف $l_k$کمترین مقدار می‌باشد. در حالتی که دو خیابان با طول صف یکسان وجود داشته باشد، بر اساس ترتیب فاز آن در طراحی سیگنال مرتب می‌شوند. شکل۲ نشان دهنده ترتیب خیابانها و فازها در یک تقاطع متشکل از سه و چهار فاز می‌باشد. تعداد حالتها بستگی به تعداد فازها دارد، با توجه به اینکه تعداد جایگشتهای ممکن برابر با $k!$ می‌باشد، در تقاطعهایی که دارای سه فاز هستند، ۶ حالت و برای تقاطعهای چهار راه، ۲۴ حالت خواهیم داشت.




*شکل ۲- ترتیب فازها در تقاطعهای سه فاز و چهار فاز*

اعمالی که برای عامل تعریف شده است، نحوه تقسیم زمان سبز مربوط به هر فاز را مشخص می‌کند. در این روش یک طول ثابت زمانی را در نظر گرفته و آن را طوری میان فازهای مختلف تقسیم می‌کنیم که به هر فاز، زمانی به عنوان حداقل زمان سبز اختصاص یابد و علاوه بر آن تعدادی بازه سبز برای تمدید در نظر گرفته شده است که هر عمل عامل، نحوه تخصیص تمدیدهای زمانی به فازهای مختلف را مشخص می‌نماید. فضای عمل با چهارتایی $< n_{ph}, t_{\min}, n_{ex}, h_{ex} >$ مشخص می‌گردد که در آن:

$n_{ph}$: تعداد فازها

$t_{\min}$: حداقل زمان سبزی است که باید به هر سیگنال تخصیص یابد

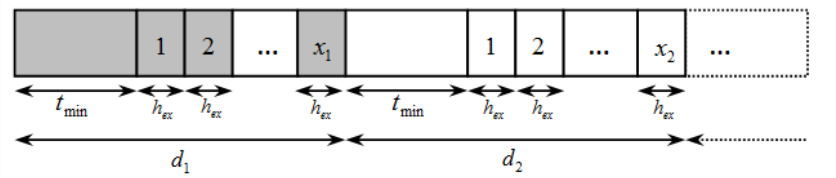
$n_{ex}$: تعداد تمدیدهای زمانی

$h_{ex}$: طول هر تمدید زمانی بر حسب ثانیه

زمان سبز تخصیص داده شده به فازهای مختلف که با $\delta$ نشان داده می‌شود به صورت زیر قابل محاسبه است.

$$\delta = h_{ex} \times n_{ex} + n_{ph} \times t_{\min} \quad (۳)$$

این رابطه به صورت گرافیکی در شکل ۳ آورده شده است. $di$ زمان سبز اختصاص داده شده به فاز iام را مشخص می‌نماید.




*شکل ۳- نمایش گرافیکی زمان اختصاص داده شده به فازها*

تخصیص تمدیدهای زمانی از نظر تئوری مشابه با حل مسئله زیر می‌باشد:

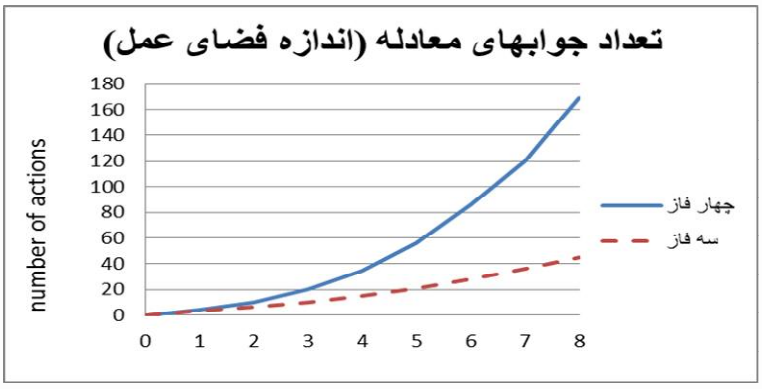
$$\sum_{i=1}^{n_{ph}} x_i = n_{ex}, x_i \in \mathrm{N} \qquad (۴)$$

که در آن $n_{ex}$، تعداد تمدیدهای زمانی و $n_{ph}$ تعداد فازها می‌باشد. واضح است که $x_i$ باید اعداد حسابی باشند. از آنجا که تعداد فازها در شبکه مورد بررسی برای تقاطعهای سه راه و چهارراه به ترتیب ۳ و ۴ می‌باشد، لذا رابطه (۴) به صورت زیر نوشته می‌شود:

$$x_1 + x_2 + x_3 = n_{ex} \qquad (۵)$$

$$x_1 + x_2 + x_3 + x_4 = n_{ex} \qquad (۶)$$

تعداد جوابهای معادله فوق، تعداد اعضای مجموعه اعمال عامل را مشخص می‌نماید و این تعداد به $n_{ex}$، تعداد تمدیدها، بستگی دارد. شکل ۴ تعداد اعضای مجموعه اعمال را بر حسب $n_{ex}$ نشان می‌دهد.




*شکل ۴- نمودار تعداد عمل بر حسب تعداد تمدیدهای زمانی در تقاطعهای سه فاز و چهار فاز*

همانطور که در شکل ۴ دیده می‌شود، روند افزایش تعداد اعمال نسبت به $n_{ex}$ نمایی می‌باشد. در یادگیری تقویتی چنانچه تعداد اعمال ممکن برای عامل زیاد باشد، همگرایی بسیار کند و در برخی موارد نیز حتی عدم همگرایی رخ

می‌دهد. در این روش، تعداد تمدیدها را برابر با ۴ در نظر گرفته ایم. به این ترتیب، برای سه فاز تعداد جوابهای معادله برابر با ۱۵ و برای چهار فاز این تعداد برابر با ۳۵ می‌شود. از محدودیت زیر برای کاهش تعداد جوابهای معادله و محدود کردن تغییرات زمان فازهای مختلف می‌توان استفاده نمود.

$$\sum_{i=1}^{n_{ph}} x_i = n_{ex}, x_i \in \mathrm{N}, \qquad x_i \le \rho, \ 1 \le \rho \le n_{ex} \tag{۷}$$

حداکثر تعداد تمدیدهایی که به هر فاز می‌توان اختصاص داد با پارامتر $\rho$ کنترل می‌شود. چنانچه $\rho$ مقدار کوچکی در نظر گرفته شود، از تعداد عمل های ممکن کاسته می‌گردد. برای انتخاب $\rho$، حالتهای مختلفی را می‌توان در نظر گرفت که در جدول ۱ آورده شده است:

*جدول ۱- تاثیر $\rho$ بر روی تعداد جوابهای معادله (۷)*

| $\rho$ | تعداد جوابهای معادله (سه فاز) | تعداد جوابهای معادله (چهار فاز) |
|---|---|---|
| ۱ | غیرقابل حل | ۱ |
| ۲ | ۶ | ۱۹ |
| ۳ | ۱۲ | ۳۱ |
| ۴ | ۱۵ | ۳۵ |

در این مقاله برای تقاطعهای سه فاز $\rho = 4$ و تقاطعهای چهارفاز $\rho = 2$ در نظر گرفته شده است. بنابراین تعداد حالتهای مختلفی که برای تخصیص زمان سبز می‌توان در نظر گرفت، برای تقاطعهای مختلف با روابط زیر به دست می‌آید:

$$x_1 + x_2 + x_3 = 4, \quad x_i \in 0,1,2,3,4 \tag{۸}$$

$$x_1 + x_2 + x_3 + x_4 = 4, \ x_i \in 0,1,2 \tag{۹}$$

زمان سبز که به هر فاز اختصاص می‌یابد برابر است با مجموع حداقل زمان سبز و میزان تمدید های زمان سبز که با رابطه زیر حاصل می‌گردد:

$$d_i = t_{\min} + x_i \times h_{ex} \tag{۱۰}$$

به طور خلاصه، مقادیر پارامترهایی که در یادگیری سطح اول مورد استفاده قرار گرفته است، در جدول ۲ نشان داده شده است.

جدول ۲- پارامترهای مورد استفاده در یادگیری عاملهای سطح اول

| پارامتر | توضیح | مقدار(سه فاز) | مقدار(چهارفاز) |
|---|---|---|---|
| $\delta$ | طول دوره زمانی موثر | ۹۴ | ۹۲ |
| $n_{ph}$ | تعداد فازها | ۳ | ۴ |
| $t_{\min}$ | حداقل زمان سبز برای هر فاز | ۱۸ | ۱۳ |
| $n_{ex}$ | تعداد بازه های زمانی تمدید فاز | ۴ | ۴ |
| $h_{ex}$ | طول زمان تمدید | ۱۰ | ۱۰ |
| $\rho$ | حداکثر تعداد تمدید برای هر فاز | ۴ | ۲ |
| $\|S\|$ | تعداد حالات | ۶ | ۲۴ |
| $\|A\|$ | تعداد اعمال | ۱۵ | ۱۹ |

با توجه به این مقادیر، مجموعه اعمال برای عاملهای مربوط به تقاطعهای سه فاز شامل ۱۵ عمل و برای عاملهای مربوط به تقاطعهای چهار فاز ۱۹ عمل می‌باشد که در جدول ۳ آورده شده است. توجه داشته باشید که ۲ ثانیه برای زمان زرد پس از هر فاز در نظر گرفته شده است. یک عمل برابر با یک دوره کامل از هر فاز خواهد بود که در یک ترتیب ثابت اجرا می گردد. به عنوان مثال، شکل ۵ یک نمونه از عمل عاملها را نشان می‌دهد. برای کلیه عاملها طول یک دوره زمانی با احتساب زمان زرد برابر با ۱۰۰ ثانیه می‌شود. در روش یادگیری تقویتی، برای محاسبه پاداش، مقدار نرمال شده میانگین طول صف خیابانها مورد استفاده قرار گرفته است.

جدول۳- مقادیر زمان سبز اختصاص داده شده به فازهای مختلف

| عاملهای دارای سه فاز | | | | عاملهای دارای چهار فاز | | | |
|---|---|---|---|---|---|---|---|
| شماره عمل | فاز ۱ | فاز ۲ | فاز ۳ | فاز ۴ | شماره عمل | فاز ۱ | فاز ۲ | فاز ۳ |
| ۱ | ۳۳ | ۳۳ | ۱۳ | ۱۳ | ۱ | ۵۸ | ۱۸ | ۱۸ |
| ۲ | ۳۳ | ۱۳ | ۳۳ | ۱۳ | ۲ | ۱۸ | ۵۸ | ۱۸ |
| ۳ | ۳۳ | ۱۳ | ۱۳ | ۳۳ | ۳ | ۱۸ | ۱۸ | ۵۸ |
| ۴ | ۱۳ | ۳۳ | ۳۳ | ۱۳ | ۴ | ۴۸ | ۱۸ | ۲۸ |
| ۵ | ۱۳ | ۳۳ | ۱۳ | ۳۳ | ۵ | ۴۸ | ۲۸ | ۱۸ |
| ۶ | ۱۳ | ۱۳ | ۳۳ | ۳۳ | ۶ | ۱۸ | ۴۸ | ۲۸ |
| ۷ | ۳۳ | ۲۳ | ۲۳ | ۱۳ | ۷ | ۲۸ | ۴۸ | ۱۸ |
| ۸ | ۳۳ | ۲۳ | ۱۳ | ۲۳ | ۸ | ۱۸ | ۲۸ | ۴۸ |
| ۹ | ۳۳ | ۱۳ | ۲۳ | ۲۳ | ۹ | ۲۸ | ۱۸ | ۴۸ |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| ۱۰ | ۲۳ | ۳۳ | ۲۳ | ۱۳ | ۱۰ | ۳۸ | ۳۸ | ۱۸ |
| ۱۱ | ۲۳ | ۳۳ | ۱۳ | ۲۳ | ۱۱ | ۳۸ | ۱۸ | ۳۸ |
| ۱۲ | ۱۳ | ۳۳ | ۲۳ | ۲۳ | ۱۲ | ۱۸ | ۳۸ | ۳۸ |
| ۱۳ | ۲۳ | ۲۳ | ۳۳ | ۱۳ | ۱۳ | ۳۸ | ۲۸ | ۲۸ |
| ۱۴ | ۲۳ | ۱۳ | ۳۳ | ۲۳ | ۱۴ | ۲۸ | ۳۸ | ۲۸ |
| ۱۵ | ۱۳ | ۲۳ | ۳۳ | ۲۳ | ۱۵ | ۲۸ | ۲۸ | ۳۸ |
| ۱۶ | ۲۳ | ۲۳ | ۱۳ | ۳۳ | | | | |
| ۱۷ | ۲۳ | ۱۳ | ۲۳ | ۳۳ | | | | |
| ۱۸ | ۱۳ | ۲۳ | ۲۳ | ۳۳ | | | | |
| ۱۹ | ۲۳ | ۲۳ | ۲۳ | ۲۳ | | | | |

### ۳-۲- سیستم مبتنی بر قانون در سطح دوم

در سطح دوم، ۱۳ عامل قرار دارند که هریک مسئول کنترل ناحیه ای متشکل از چهار عامل می‌باشند. عاملهای سطح دوم نقش کنترلی خود را طوری ایفا می‌کنند که در مواقع ضروری تا حد ممکن از اشباع شدن خیابانهای ناحیه جلوگیری نمایند. برای این منظور در این عاملها از سیستم مبتنی بر قانون جهت کنترل و جلوگیری از وقوع بحران استفاده می‌نمائیم. عاملهای سطح دوم پارامترهایی جهت تعیین وضعیت ناحیه محاسبه کرده و با توجه به آن برای عاملهای سطوح پائین تر سیاستی را مشخص می‌نمایند. عاملهای سطح دوم تراکم خودروها را در ناحیه محاسبه کرده و بر اساس قوانین موجود به محدود کردن فضای عمل عاملهای عضو خود می‌پردازند. فرض کنید عامل $\beta_k$ مسئول کنترل عاملهای $\alpha_i$، $i=1,\ldots,m$ باشد. همچنین فرض کنید عاملهای $\alpha_i$ تقاطعهای حاصل از چهار خیابان ورودی هستند که با $l_j^i$، $i=1,\ldots,4$ نشان داده می‌شود. عامل $\beta_k$ تراکم خودروها را در ناحیه اندازه گیری می‌کند و بر اساس قانون زیر به محدود کردن فضای عمل عامل می‌پردازد:

$$if \;\; d \geq \lambda_1 \;\; \& \;\; s_i(t) \in \{s(l_1,\ldots,l_k) \mid l_j = \max_{i=1,\ldots,k}(l_i)\} \Rightarrow a_i(t) \in \{a(x_1,\ldots,x_k) \mid x_j = \max_{i=1,\ldots,k}(x_i)\} \qquad (۱۱)$$

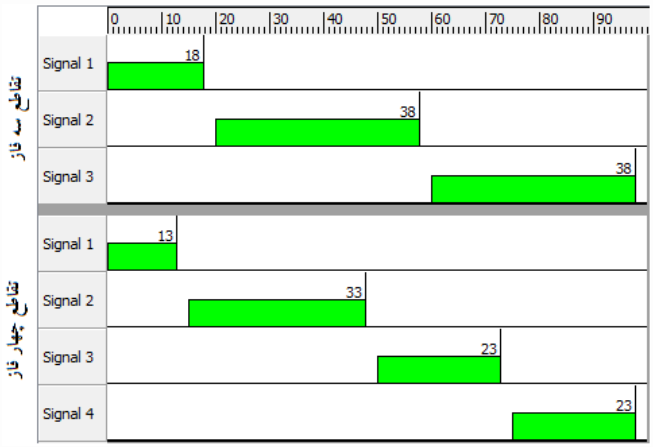
$$if \;\; d \geq \lambda_2 \;\; \& \;\; s_i(t) \in \{s(l_1,\ldots,l_k) \mid l_j = \max_{i=1,\ldots,k}(l_i)\} \Rightarrow a_i(t) \in \{a(x_1,\ldots,x_k) \mid x_j = n_{ex}\} \qquad (۱۲)$$

$\lambda_1$ و $\lambda_2$ حدود آستانه ای هستند که با استفاده از وضعیت شبکه تعیین می‌شوند. در شبیه سازی این مقادیر برابر با ۳۰٪ و ۶۰٪ در نظر گفته شده است. $d$ تراکم خودروها در ناحیه مربوطه می‌باشد که در هر دوره زمانی با در نظر داشتن ظرفیت خیابانها محاسبه و به روز می‌شود. قوانین فوق محدودیت های رابطه (۴) افزایش می‌دهند و لذا عامل به جای اینکه از میان مجموعه عمل ممکن خود عملی را انتخاب نماید، از مجموعه محدودتری انتخاب می‌نماید که مغایرتی با قانون عامل سطح بالاتر نداشته باشد. قانون اول به این صورت است که در صورتی که تراکم خودروها در ناحیه بیشتر از $\lambda_1$ گردد، محدودیتی به عاملها ابلاغ می گردد که باعث می‌شود زمان بیشتری به

فازهایی اختصاص یابد که مربوط به خیابانی با بیشترین طول صف خودروها می‌باشد. از میان عملهای موجود در مجموعه عمل عاملها تنها ۶ عمل مطابق با این قانون می باشد. بدین ترتیب عاملهای مربوط به تقاطعهای چهار فاز و سه فاز به ترتیب به جای اینکه از میان ۱۹ و ۱۵ عمل، عملی را انتخاب نمایند، از میان ۶ عمل به انتخاب عمل می‌پردازند.

قانون دوم محدودیت بیشتری روی فضای عمل اعمال می‌نماید. بدین ترتیب که همه تمدیدها را به فازی اختصاص می‌دهد که بیشترین طول صف را دارد. بدین ترتیب هر عامل تنها یک عمل می‌تواند انجام دهد و آن هم عملی است که ۴ تمدید برای فاز مربوطه در نظر می‌گیرد و سایر فازها حداقل زمان سبز را که برابر با $t_{\min}$ می‌باشد اجرا می‌نمایند. چنانچه این عمل در فضای عامل وجود داشته باشد (عاملهای سه فاز) تاثیری در روند یادگیری تقویتی ایجاد نمی‌شود و فرآیند به روز رسانی خط مشی می‌تواند انجام گیرد، اما چنانچه در فضای عمل وجود نداشته باشد (عاملهای چهارفاز) خط مشی ثابت مانده و به روز نمی‌شود.

بدین ترتیب یک سیستم سلسله مراتبی در دو سطح خواهیم داشت که عاملهای سطح اول از یادگیری تقویتی برای یافتن کنترل بهینه استفاده می‌نمایند در حالی که عاملهای سطح دوم، وضعیت ناحیه را بررسی کرده و در مواقع لازم با محدود کردن فضای انتخاب عاملها، نقش کنترلی خود را ایفا می‌نمایند.




*شکل۵- نمایش گرافیکی عمل شماره ۱۲ برای عاملهای مربوط به تقاطعهای سه فاز و چهار فاز*

## ۴- نتایج تجربی

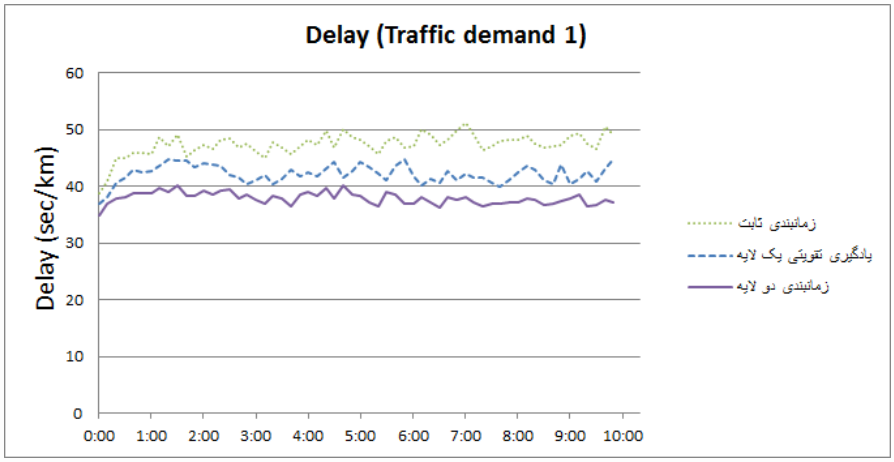
روش ارائه شده در این مقاله با استفاده از نرم افزار Aimsun بر روی شبکه شکل ۱ پیاده سازی شده است. نتایج تجربی حاصل از به کارگیری این در این بخش آورده شده است. پارامترهایی که در شبیه سازی ترافیک شهری در نظر گرفته شده اند، در جدول ۴ آورده شده است.

*جدول ۴- پارامترهای مورد استفاده در شبیه سازی شبکه ترافیک شهری*

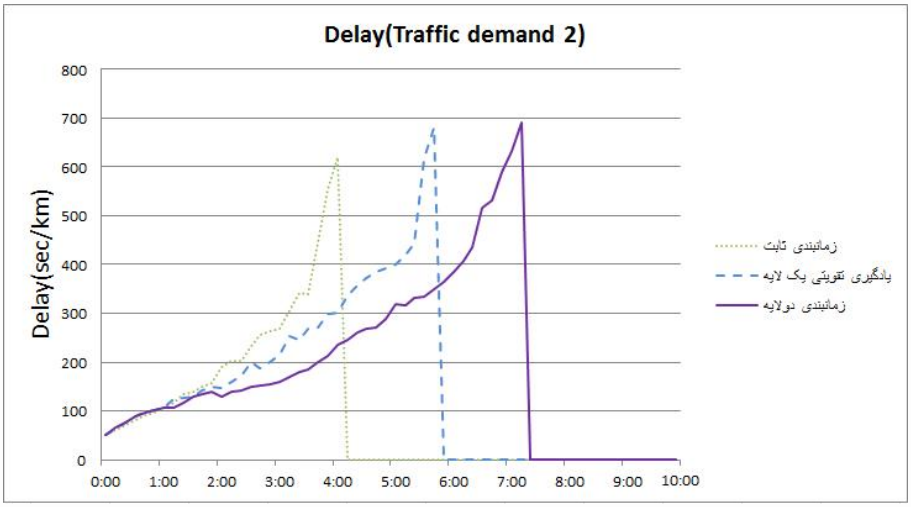
| خصوصیات | مقدار |
|---|---|
| تعداد تقاطع ها | ۳۶ |
| تعداد خیابانها | ۷۰ |
| میانگین طول خیابانها (متر) | ۶۱۲/۸۵ |
| تعداد خطوط عبوری خیابانها | ۳ |
| حداکثر سرعت (کیلومتر بر ساعت) | ۵۰ |
| تعداد مراکز ورودی/خروجی | ۱۰ |
| توزیع ورودی خودروها | نرمال |
| مدت زمان شبیه سازی(ساعت) | ۱۰ |
| حجم ترافیک ۱ (خودرو در ساعت) | ۱۰۰۰۰ |
| حجم ترافیک ۲ (خودرو در ساعت) | ۲۰۰۰۰ |

نتایج حاصل از اجرای روش پیشنهادی در دو حالت یک سطحی و دو سطحی برای حجم ترافیک ۱ در شکل ۶ با روش زمانبندی ثابت مقایسه شده است. این نتایج میانگین حاصل از ۱۰ اجرای مستقل روشها بر روی شبکه مذکور می‌باشد. در شبیه سازی، ابتدا روش یادگیری تقویتی در سطح اول برای زمانبندی تقاطع ها مورد استفاده قرار گرفته است و نتایج نشان می‌دهد که میانگین تاخیر در کل شبکه کاهش یافته است. سپس سیستم مبتنی بر قانون سطح دوم به یادگیری تقویتی عاملهای سطح اول اضافه و مورد آزمایش قرار گرفته است. همانطور که در شکل ۶ مشاهده می‌شود، روش زمانبندی دو سطحی باعث کاهش بیشتر میانگین تاخیر در شبکه و در نتیجه افزایش کارایی شده است.

نتایج حاصل از اجرای روشهای مذکور برای حجم ترافیک ۲ در شکل ۷ نشان داده شده است. برای حجم ترافیک بالا اگرچه با استفاده از هر سه روش، شبکه بعد از مدتی اشباع شده اما میانگین تاخیر قبل از اشباع و همچنین زمان رسیدن به اشباع متفاوت می‌باشد. نمودار شکل ۷ نشان می‌دهد که کارایی زمانبندی دو سطحی بهتر از یادگیری تقویتی یک سطحی است و این روش نیز بهتر از زمانبندی ثابت می‌باشد.




*شکل ۶- میانگین تاخیر سه روش در ۱۰ ساعت شبیه سازی برای حجم ترافیک ۱*




*شکل ۷- میانگین تاخیر سه روش در ۱۰ ساعت شبیه سازی برای حجم ترافیک ۲*

## ۵- نتیجه گیری و کارهای آینده

سیستمهای چندعامله اغلب برای مدل کردن سیستمهای توزیع شده در مقیاس بالا مورد استفاده قرار می‌گیرد. یکی از روشهایی که برای کاهش پیچیدگی در این سیستمها به کار می رود، استفاده از ساختارهای سلسله مراتبی می‌باشد. ساختار سلسله مراتبی باعث تجزیه مساله به زیر مسائل ساده تر شده و از این رو منجر به کاهش پیچیدگی می‌گردد. در این مقاله یک شبکه ترافیکی شامل ۳۶ تقاطع با استفاده از سیستمهای چندعامله سلسله مراتبی در دو سطح مدل گردید. در سطح اول عاملها با به کارگیری یادگیری تقویتی به کنترل و زمانبندی چراغهای راهنمایی در هر تقاطع می‌پردازند. مدل یادگیری تقویتی ارائه شده به صورت پارامتری بوده و لذا امکان افزایش و یا کاهش اعمال ممکن در مجموعه عمل عامل وجود دارد. از طرفی دیگر مجموعه حالت و عمل طوری

تعریف شده است که به راحتی می‌تواند برای تقاطعهایی با تعداد فازهای مختلف تعمیم یابد. شبکه ترافیک شهری به نواحی منظم تقسیم بندی شده و به هر ناحیه عاملی در سطح دوم نسبت داده شده است. عاملهای سطح دوم از سیستم مبتنی بر قانون برای کاهش فضای انتخاب عاملها و در راستای کاهش تراکم خودروها استفاده می‌کنند. این عاملها تراکم خودروها را در ناحیه کنترل می‌کنند و زمانی که از حد مشخصی بیشتر شده از قوانینی جهت کاهش آن و در نتیجه کاهش میانگین تاخیر و جلوگیری از اشباع شدن شبکه استفاده می‌کنند. نتایج تجربی حاصل از اجرای روش بر روی شبکه گرید پیاده سازی و با روش زمانبندی ثابت مقایسه گردیده است. میانگین تاخیر با استفاده از ساختار پیشنهادی دو سطحی کمتر از ساختار یک سطحی در شبکه بوده و هر دو روش بهتر از زمانبندی ثابت در محیط عمل کرده اند. در حجم ترافیک بالا، اگرچه هر سه روش در نهایت منجر به اشباع شبکه می‌شوند اما زمان رسیدن به اشباع در روشهای مختلف فرق می‌کند و روش سلسله مراتبی دو سطحی در شرایط فوق اشباع توانسته شبکه را طوری کنترل نماید که تا حد ممکن زمان اشباع را به تعویق اندازد.

کارهای آینده شامل به کارگیری روش پیشنهادی در شبکه های واقعی و تعمیم روش برای تقاطع هایی با تعداد فازهای مختلف می‌باشد. همچنین می‌توان تعداد سطوح سلسله مراتبی را افزایش داد و کارایی روش را با افزایش تعداد سطوح اندازه گیری نمود. در این مقاله، ناحیه ها به طور ثابت و مشخص تعیین شده اند، که تعیین ناحیه ها نیز خود بحث جداگانه و مستقلی است که می‌توان به آن پرداخت.

## ۶- مراجع

1- Chen, B., Cheng, H. H., Palen, J. (2009) .Integrating mobile agent technology with multi-agent systems for distributed traffic detection and management systems, Transportation Research Part C: Emerging Technologies, No. 17, Issue 1, pp. 1–10.
2- Roozemond, Danko A. (2001). Using intelligent agents for pro-active, real-time urban intersection control, European Journal of Operational Research, No.131, Issue 2, pp. 293–301.
3- Cai, C. and Yang, Z.(2007). Study on urban traffic management based on multi-agent system", 6th International Conference on Machine Learning and Cybernetics, August 19-22, 2007, Hong Kong, pp. 25–29.
4- Choy, Min Chaee, Srinivasan, Dipti and Cheu, Ruey Long (2003). Cooperative, hybrid agent architecture for real-time traffic signal control, IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans, No. 33, Issue 5, pp. 597–607.
5- Srinivasan, Dipti, Choy, Min Chaee and Cheu, Ruey Long (2006). Neural networks for real-time traffic signal control, IEEE Transactions on Intelligent Transportation Systems, No. 7, Issue 3, pp. 261–272.
6- Grégoire, P., Desjardins, C., Laumônier, J. and Chaib-draa, B. (2007), "Urban traffic control based on learning agents", Intelligent Transportation Systems Conference, September 30- October 3, 2007, Laval University, Quebec, pp. 916–921.
7- Weiring, M. (2000) "Multi-agent reinforcement learning for traffic light control, 7th International Conference on Machine Learning, Stanford University, US, June 29-July 2, 2000, pp. 1151–1158.
8- Steingröver, M., Schouten, R., Peelen, S., Nijhuis, E., Bakker, B. (2005) "Reinforcement learning of traffic light controllers adapting to traffic congestion", 17th Belgium-Netherlands Conference on Artificial Intelligence (BNAIC 2005), October 17-18, 2005, Brussels, Belgium: Koninklijke Vlaamse Academie van Belie voor Wetenschappen en Kunsten, pp. 216–223.
9- Silva, B.C.d., Basso, E.W., Bazzan, A.L.C. and Engel, P.M. (2006), "Improving reinforcement learning with context detection", 5th International Joint Conference on Autonomous Agents and Multi-agent Systems (AAMAS), May 8-12, 2006, Hakodate, Japan, pp. 811–812.
10- Wen, K., Qu, S. and Zhang, Y. (2008) "A stochastic adaptive control model for isolated intersections", 2007 IEEE International Conference on Robotics and Biomimetics, December 15-19, 2007, Sanya, China, pp. 2256–2260.
11- Dai, Y., Zhao, D. and Yi, J.a. (2010) "A comparative study of urban traffic signal control with reinforcement learning and adaptive dynamic programming", International Joint Conference on Neural Networks. Barcelona, July 18-23, 2010, pp. 1–7.
12- Mannion, P., Duggan, J. and Howley, E. (2015), "Parallel Reinforcement Leaning for Traffic Signal Control", Procedia Computer Science 52, pp. 956-961.

13- Prashanth, L.A. and Bhatnagar, Shalabh (2011). Reinforcement learning with function approximation for traffic signal control, IEEE Transactions on Intelligent Transportation Systems, No. 99, pp. 1–10.
14- Balaji, P. G., German, X. J. A., Srinivasan, D. (2010). Urban traffic signal control using reinforcement learning agents", IEEE Transactions on Intelligent Transportation Systems, No. 4, Issue 3, pp. 177–188.
15- Salkham, A., Cunningham, R., Garg, A. and Cahill, V. (2008), "A collaborative reinforcement learning approach to urban traffic control optimization", International Conference on Web Intelligence and Intelligent Agent Technology, December 9-12, 2008, Sydney, Australia, pp. 560–566.
16- Gabric, T., Howden, N., Norling, E., Tidhar, G., L. Sonensberg (1994), "Multi-agent design of a traffic flow control system", Technical Report, Department of Computer Science, The University of Melbourne.
17- Watkins, Chris and Dayan, Peter (1992). Q-learning, Machine learning, Vol. 8, pp. 279-292.

# Urban Traffic Control Using Hierarchical Multi-agent System and Reinforcement Learning

**Monireh Abdoos**
Assistant Professor, Faculty of Computer Science and Engineering, Shahid Beheshti University
m_abdoos@sbu.ac.ir

## Abstract

By increasing the number of vehicles in urban network, classic methods cannot be used for traffic control. Hence, intelligent methods have been massively used today. In this paper, a new method is proposed for urban traffic control using a two-level hierarchical multi-agent system. At the first level, a network including a large number of intersections has been modeled using multi-agent system. At the second level, each agent controls a region of network which contains a number of intersections. Reinforcement learning has been used for traffic signal timing at the first level. At the second a rule based system has been used to control the vehicle congestion control in a region. The experimental results on a grid network show that the proposed method decreases the average delay time and prevents the network to be saturated.

***Keywords:*** ***Intelligent Traffic Control, Multi-agent System, Hierarchical structures, Reinforcement Learning.***