یازدهمین کنفرانس سیستم‌های هوشمند ایران
۹ و ۱۰ اسفند ۱۳۹۱
(ICIS2013)
11th Iranian Conference on Intelligent Systems
February 27th & 28th, 2013

استفاده از عاملهای یادگیرنده در کنترل هوشمند ترافیک شهری

منیره عبدوس[۱] و ناصر مزینی[۲]
[۱]دانشکده مهندسی کامپیوتر، دانشگاه علم و صنعت ایران، abdoos@iust.ac.ir
[۲]دانشکده مهندسی کامپیوتر، دانشگاه علم و صنعت ایران، mozayani@iust.ac.ir

چکیده – سیستمهای چندعامله ابزار مناسبی جهت مدلسازی سیستمهای توزیع شده به شمار می رود. در اغلب این سیستمها اطلاعات دقیق و از پیش مشخص از محیط در دسترس نیست و نیاز به روشی است که عاملها بدون داشتن اطلاعات قبلی از محیط بتوانند در آن عمل نمایند. یادگیری Q این امکان را برای عاملها فراهم می آورد که بتوانند با استفاده از تجربیات خود، خط مشی مناسبی را یاد بگیرند. از طرفی دیگر، 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 چند عامله مبتنی بر یادگیری Q ارائه شده است. یک شبکه شامل تعداد زیادی از تقاطعها، با یک سیستم چندعامله مدل شده است که در آن عاملها از یادگیری Q برای یافتن بهترین زمانبندی چراغهای راهنمایی بهره می برند. نتایج حاصل از پیاده سازی روش بر روی شبکه گرید شامل تقاطعهای سه فاز و چهار فاز نشان می‌دهند که روش پیشنهادی باعث کاهش میزان تاخیر در زمان سفر شده و همچنین در حجم ترافیک سنگین زمان اشباع شبکه را به تعویق می اندازد.

کلید واژه- زمانبندی چراغهای راهنمایی، عاملهای یادگیرنده، کنترل هوشمند ترافیک، یادگیریQ.

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 عاملها از یک مدل یادگیری تقویتی، یادگیریQ، جهت یافتن خط مشی بهینه استفاده می‌کنند. 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.

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 [۲] 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 [۳] 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 [۴] 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 [۵] 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 [۶] و [۷] جهت کنترل محیط مستقل از مدل ارائه شده است. کنترل تطبیقی در [۸] 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 [۹] در یک تقاطع و برخی دیگر مانند روشهای [۶] و [۸] 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 [۱۰] 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 [۹].

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 چندعامله مدل می‌شود. به این صورت که هر یک از عاملها مسئول زمانبندی یک چراغ راهنمایی می‌باشد. عاملها از یادگیری تقویتی برای زمانبندی استفاده می‌کنند که وابسته به مدل از پیش تعیین شده محیط ندارند. نحوه نمایش فضای حالت و عمل عاملها در روش ارائه شده به صورت پارامتری است که با تغییر پارامترها می توان ابعاد فضا را تغییر داد.

در بخش بعدی، مقدمه ای کوتاه در مورد یادگیری تقویتی مورد استفاده در این مقاله که یادگیری Q می باشد ذکر می گردد. 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

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 ناظر 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 یادگیریQ [۱۱]، 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:

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

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.

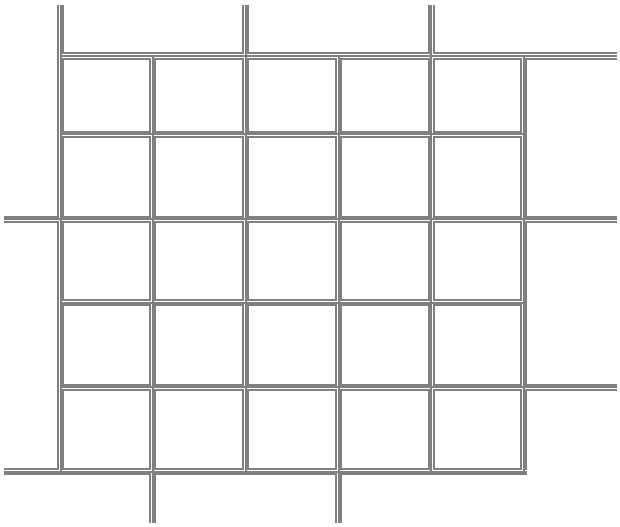
اثبات می‌شود که اگر تمامی زوج‌های حالت-عمل به صورت مکرر تجربه شوند و نرخ یادگیری در طول زمان کاهش یابد، یادگیری-Q با احتمال ۱ به مقدار بهینه‌ی $Q^*(s,a)$ همگرا می‌شود. با توجه به این امر، سیاست بهینه به صورت زیر تعریف می‌شود:

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

در این مقاله از روش یادگیریQ 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- یادگیری Q جهت کنترل هوشمند ترافیک

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. سیستم چندعامله متشکل از ۳۶ عامل است که هر کدام مسئول کنترل یک تقاطع می‌باشند. تقاطع ها به صورت چهار راه و سه راه می‌باشند که در یک شبکه گرید قرار گرفته اند.

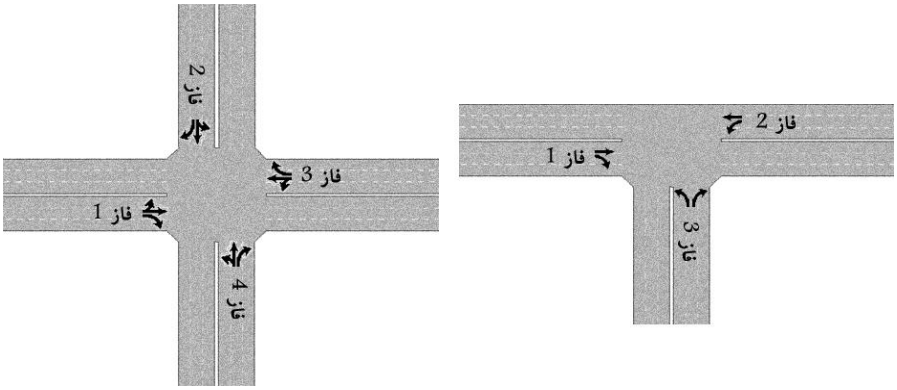


شکل ۱- شبکه ترافیک شهری مورد استفاده در شبیه سازی

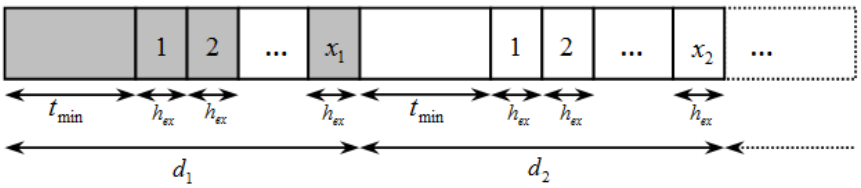
در این روش، به هر خیابان ورودی یک فاز اختصاص داده شده

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 α^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 باشند، بر اساس ترتیب فاز آن در طراحی سیگنال مرتب می‌شوند. شکل ۲ 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



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



شکل ۳- نمایش گرافیکی زمان اختصاص داده شده به فازها

تقاطعهایی که دارای سه فاز هستند، ۶ حالت و برای تقاطعهای چهار راه، ۲۴ حالت خواهیم داشت.

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 δ نشان داده می‌شود به صورت زیر قابل محاسبه است.

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

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

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

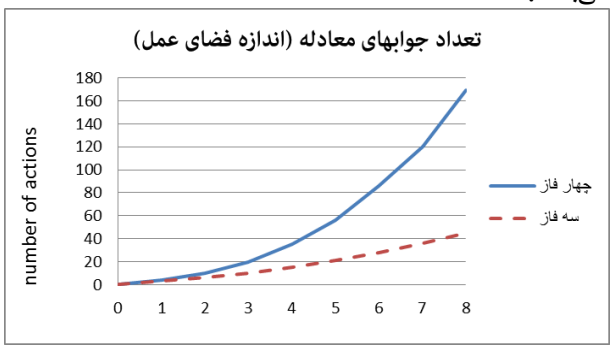
$$\sum_{i=1}^{n_{ph}} x_i = n_{ex},\ x_i \in \mathrm{N} \qquad (۴)$$

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. همانطور که در شکل ۴ دیده می‌شود، روند افزایش تعداد اعمال نسبت به n_{ex} نمایی می‌باشد.



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

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.

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

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

جدول ۱- تاثیر ρ بر روی تعداد جوابهای معادله (۷)

ρ	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:

$$x_1 + x_2 + x_3 = 4,\quad x_i \in 0,1,2,3,4 \qquad (۸)$$

$$x_1 + x_2 + x_3 + x_4 = 4,\ x_i \in 0,1,2 \qquad (۹)$$

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:

$$d_i = t_{\min} + x_i \times h_{ex} \qquad (۱۰)$$

به طور خلاصه، مقادیر پارامترهایی که در یادگیری سطح اول مورد استفاده قرار گرفته است، در جدول ۲ نشان داده شده است. 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 چند نمونه از آنها در جدول ۳ آورده شده است. توجه داشته باشید که ۲ ثانیه برای زمان زرد پس از هر فاز در نظر گرفته شده است.

جدول ۲- پارامترهای مورد استفاده در یادگیری عاملهای سطح اول

پارامتر	توضیح	مقدار (۳ فاز)	مقدار (۴فاز)
δ	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
ρ	حداکثر تعداد تمدید برای هر فاز	۴	۲
$\|S\|$	تعداد حالات	۶	۲۴
$\|A\|$	تعداد اعمال	۱۵	۱۹

یک عمل برابر با یک دوره کامل از هر فاز خواهد بود که در یک ترتیب ثابت اجرا می گردد. 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- مثالی از مقادیر زمان سبز اختصاص داده شده به فازهای مختلف

عاملهای دارای چهار فاز					عاملهای دارای سه فاز			
شماره عمل	فاز۱	فاز۲	فاز۳	فاز۴	شماره عمل	فاز۱	فاز۲	فاز۳
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- 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 روش در این بخش آورده شده است. پارامترهایی که در شبیه سازی ترافیک شهری در نظر گرفته شده اند، در جدول ۴ آورده شده است.

نتایج حاصل از اجرای روش پیشنهادی برای حجم ترافیک ۱ در شکل ۵ با روش زمانبندی ثابت مقایسه شده است. این نتایج میانگین حاصل از ۱۰ اجرای مستقل روشها بر روی شبکه مذکور می‌باشد. نتایج نشان می‌دهد که میانگین تاخیر در کل شبکه کاهش یافته است و در نتیجه کارایی بهبود یافته است.

جدول۴- پارامترهای مورد استفاده در شبیه سازی شبکه ترافیک شهری

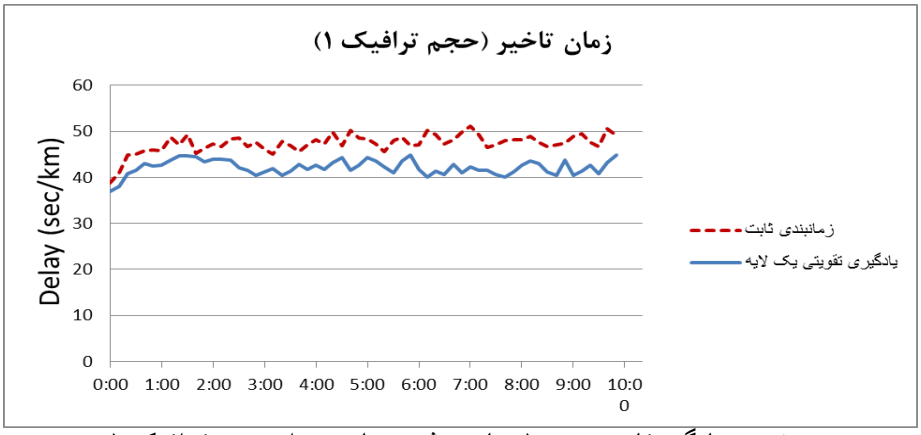
خصوصیات	مقدار
تعداد تقاطع ها	۳۶
تعداد خیابانها	۷۰
میانگین طول خیابانها (متر)	۶۱۲/۸۵
تعداد خطوط عبوری خیابانها	۳
حداکثر سرعت (کیلومتر بر ساعت)	۵۰
تعداد مراکز ورودی/خروجی	۱۰
توزیع ورودی خودروها	نرمال
مدت زمان شبیه سازی(ساعت)	۱۰
حجم ترافیک ۱ (خودرو در ساعت)	۱۰۰۰۰
حجم ترافیک ۲ (خودرو در ساعت)	۲۰۰۰۰

نتایج حاصل از اجرای روشهای مذکور برای حجم ترافیک ۲ در شکل ۶ نشان داده شده است. برای حجم ترافیک بالا اگرچه با استفاده از هر دو روش، شبکه بعد از مدتی اشباع شده، اما میانگین تاخیر قبل از اشباع و همچنین زمان رسیدن به اشباع متفاوت می‌باشد.

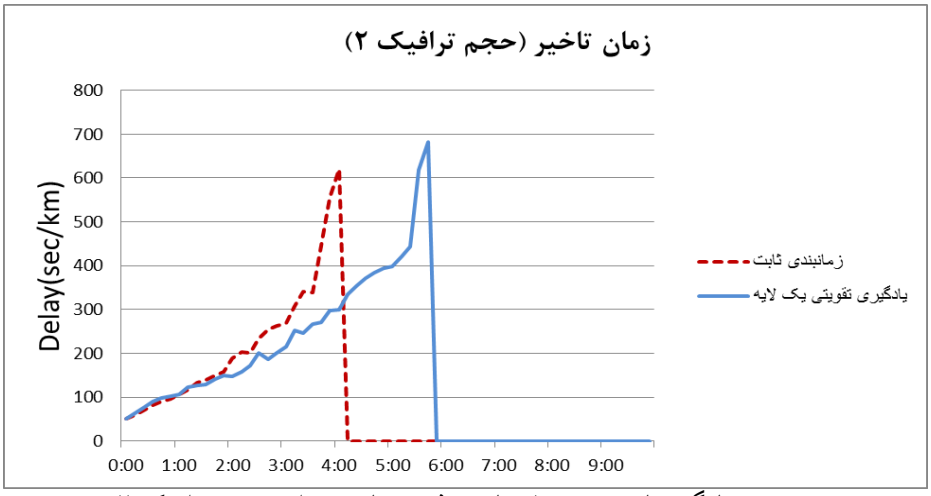
۵- نتیجه‌گیری

سیستمهای چندعامله اغلب برای مدل کردن سیستمهای توزیع شده در مقیاس بالا مورد استفاده قرار می‌گیرد. در اغلب این سیستمها اطلاعات دقیق و از پیش مشخص از محیط در دسترس نیست و نیاز به روشی است که عاملها بدون داشتن اطلاعات قبلی از محیط بتوانند در آن عمل نمایند که یادگیری Q این امر را ممکن می سازد.

در این مقاله یک شبکه ترافیکی شامل ۳۶ تقاطع با استفاده از سیستمهای چندعامله مدل گردید که در آن عاملها با به کارگیری یادگیری Q به کنترل و زمانبندی چراغهای راهنمایی در هر تقاطع می‌پردازند.



شکل ۵- میانگین تاخیر در ۱۰ ساعت شبیه سازی برای حجم ترافیک ۱



شکل۶- میانگین تاخیر در ۱۰ ساعت شبیه سازی برای حجم ترافیک ۲

مدل یادگیری 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. 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 روش پیشنهادی بهتر از زمانبندی ثابت بوده است. در حجم ترافیک بالا، اگرچه هر سه روش در نهایت منجر به اشباع شبکه می‌شوند اما زمان رسیدن به اشباع فرق می‌کند و روش پیشنهادی باعث تعویق زمان اشباع شبکه شده است.

کارهای آینده شامل به کارگیری روش پیشنهادی در شبکه های واقعی و تعمیم روش برای تقاطع هایی با تعداد فازهای مختلف می‌باشد.

سپاسگزاری

این پژوهش با حمایت مالی مرکز تحقیقات مخابرات ایران صورت گرفته است.

مراجع

[1] B. Chen, H. H. Cheng and J. Palen, “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, 2009.

[2] P. Grégoire, C. Desjardins, J. Laumônier and B. Chaib-draa, “Urban traffic control based on learning agents”, Intelligent Transportation Systems Conference, Laval University, Quebec, pp. 916–921, 2007.

[3] M. Weiring, “Multi-agent reinforcement learning for traffic light control, 7th International Conference on Machine Learning, Stanford University, US, pp. 1151–1158, 2000.

[4] M. Steingröver, R. Schouten, S. Peelen, E. Nijhuis, B. Bakker, “Reinforcement learning of traffic light controllers adapting to

[8] L.A. Prashanth and S. Bhatnagar, "Reinforcement learning with function approximation for traffic signal control", IEEE Transactions on Intelligent Transportation Systems, No. 99, pp. 1–10, 2011.

[9] P.G. Balaji, X. German and D. Srinivasan, "Urban traffic signal control using reinforcement learning agents", IEEE Transactions on Intelligent Transportation Systems, No. 4, Issue 3, pp. 177–188, 2010.

[10] A. Salkham, R. Cunningham, A. Garg and V. Cahill, "A collaborative reinforcement learning approach to urban traffic control optimization", International Conference on Web Intelligence and Intelligent Agent Technology, Sydney, Australia, pp. 560–566, 2008.

[11] C. Watkins and P. Dayan, "Q-learning", Machine learning, Vol. 8, pp. 279-292, 1992.

traffic congestion", 17th Belgium-Netherlands Conference on Artificial Intelligence (BNAIC 2005), Brussels, Belgium , pp. 216–223, 2005.

[5] B.C.d. Silva, E.W. Basso, A.L.C. Bazzan and P.M. Engel, "Improving reinforcement learning with context detection", 5th International Joint Conference on Autonomous Agents and Multi-agent Systems (AAMAS), Hakodate, Japan, pp. 811–812, 2006.

[6] Y. Dai, D. Zhao and J.a Yi, "A comparative study of urban traffic signal control with reinforcement learning and adaptive dynamic programming", International Joint Conference on Neural Networks. Barcelona, pp. 1–7, 2010.

[7] K. Wen, S. Qu and Y. Zhang, "A stochastic adaptive control model for isolated intersections", IEEE International Conference on Robotics and Biomimetics, Sanya, China, pp. 2256–2260, 2008.