کنترل تطبیقی چراغ‌راهنمایی با استفاده از یادگیری تقویتی

امیرحسین خلیلی[۱]، رضا صفابخش[۲]

۱. دانشکدهٔ مهندسی کامپیوتر, دانشگاه صنعتی شریف.

۲. استاد و عضو هیئت علمی دانشکدهٔ مهندسی کامپیوتر و فنآوری اطلاعات, دانشگاه صنعتی امیرکبیر.

a_khalili@ce.sharif.edu

۱- مقدمه

در دهه‌های اخیر، با افزایش چشمگیر وسایل نقلیه و حجم نقل و انتقالات, زمانهای سفرهای درون شهری افزایش یافته است. کند شدن شریان توسعه و اقتصاد شهری، افسردگی، کاهش توان فردی، مصرف بسیار زیاد سوخت، افزایش سطح دی اکسید کربن و سایر آلاینده‌های محیطی، ترافیک را به صورت یک معضل اجتماعی-اقتصادی درآورده است، به‌گونه‌ای که مدیریت و کنترل صحیح آن اهمیت زیادی یافته و در دستور کار مسئولین قرار گرفته است.

در حال حاضر اکثر سیستم‌های کنترل ترافیک از قوانین ثابت و از پیش تعیین شده‌ایی استفاده می‌کنند که تحت عنوان آیین-نامه‌ها، دارای جامعیت عمومی بوده و کمتر باشرایط و تغییرات محلی سازگارند. درحالیکه، مطالعات جدید نشان می‌دهد که اتخاذ استراتژیها به صورت بی‌درنگ و متناسب با تغییرات محیط، زمانهای انتظار در پشت چراغها را ۵ تا ۱۵ درصد کاهش می‌دهد.

سیستم‌های کنترل ترافیک به منظور بهبود عملکرد خود می‌بایست تغییرات تدریجی در محیط ترافیکی را شناسایی کرده، نحوه تغییرات را تخمین زنند، و خود را با آن تطبیق دهند. همچنین اینگونه سیستم‌ها می‌بایست قادر به کنترل شرایط غیر قابل پیش‌بینی نظیر تصادفات نیز بوده، وآنها را به وسیله ابزار ترافیکی (نظیر چراغهای راهنمایی) و همکاری بین آنها کنترل نماید. به منظور طراحی و پیاده‌سازیِ چنین عامل‌هایی روش‌های یادگیری ماشینی هوشمند انتخاب مناسبی می‌نماید.

این مقاله با ادغام روش‌های تعمیم و تقریب خطی با داده‌های اتخاذ شده از حس‌گرهای متعارف ترافیکی و نیز استفاده از روش یادگیری تقویتی, روشی کارا با بار حافظه‌ای و محاسباتی کم به منظور کنترل جریان ترافیک ارائه می‌دهد.

در ادامه, در بخش ۲ مروری بر روش‌های استفاده شدهٔ پیشین ارائه می‌گردد. بخش‌های ۳ و ۴ به معرفی مختصر یادگیری تقویتی و روش تعمیم و تقریب می پردازند. بخش ۵, شبیه ساز ترافیکی مورد استفاده را معرفی می‌کند و بخش ۶ به معرفی سناریوی یادگیری می‌پردازد. بخش ۷ نتایج تجربی به دست آمده در محیط شبیه‌سازی شده را نشان داده و اثر پارامترهای یادگیری معرفی شده را از نظر شهودی مورد بررسی قرار می‌دهد. در آخر, در بخش ۸ مطالب جمع‌بندی و نتیجه‌گیری می‌شوند.

۲- مروری بر روش‌های پیشین

با توجه به اهمیت و گسترش کنترل مکانیزهٔ ترافیک, تحقیقات صورت گرفته در این حوزه بسیار است آنچنان که بررسی تمامی آنها از حوزهٔ این مقاله خارج است. لذا تنها به گزیدهٔ منتخب از مقالات و روش‌های معرفی شده در خصوص بکاربری تکنیک‌های هوش-مصنوعی در کنترل ترافیک اکتفا می‌شود.

مقالهٔ [۱] یک سیستم کنترل فازی به منظور کنترل جریان ترافیک معرفی می‌کند. این سیستم قادر به تشخیص تعداد ماشین‌های عبوری از تقاطع در زمان سبز بودن چراغ، بوده و داده‌ها را به ۴۰ قانون افراز می‌کرد. به وسیلهٔ این قوانین زمان چراغ‌ها به گونه‌ایی تنظیم می‌گردد که جریان روان‌تر ترافیک حاصل گردد. با استفاده از طرح پیشنهادی در [۱], متوسط زمان‌های انتظار و تعداد توقف‌ها به شدت کاسته شد. عملکرد این گونه سیستم‌ها به شدت وابسته کیفیت قوانین انتخابی است. این قوانین عمدتاً توسط کاربر بیرونی برای سیستم معرفی می‌شود و پس از تعریف ثابت بوده و قابل تطبیق بر اساس شرایط جدید محیط نمی‌باشند.

مقالهٔ [۲] از سیستم‌های خبره به منظور طراحی چراغ‌های راهنمایی تقاطع‌ها استفاده کرد. سیستم آنها بعلت حجم محاسبات بالا نیاز به ساده‌سازی‌هایی در فرضیات مسئله است. سیستم‌های خبره نیز از یک سری قوانین از پیش تعیین شده به منظور تصمیم‌گیری عمل بعدی استفاده می‌کند. قابلیت این گونه سیستم‌ها نیز به کیفیت قوانین در نظر گرفته شده بستگی دارد.

در کار [۳و۴] از پیش‌بینی روند ترافیک به منظور بهینه‌سازی تصمیم‌گیری‌ها استفاده شد. [۵] از شبکه‌های عصبی بازگشتی به منظور پیش‌بینی جریان ترافیک بهره‌بردند.

در مقایسه با سایر روش‌های هوشمند استفاده شده در کنترل چراغ‌های راهنمایی, گونه‌هایی از روش تقویتی, نظیر روش تفاضل‌زمانی سارسا لاندا [۶], دارای مزایای ویژه‌ایی است. روشهای تفاضل زمانی نیاز به پیش تعریف مدل محیط برای انتخاب عمل موردنظر ندارند. در عوض ارتباط بین وضعیت، عمل و پاداش در تعامل پویایی که عامل با محیط دارد آموخته می‌شود. این در حالی است که اکثر روشهای دیگر کنترل هوشمند چراغهای راهنمایی نیاز به تعریفی از مدل محیط دارند تا برآوردی از نحوهٔ جریان عبور و مرور در آیندهٔ نزدیک به دست آورند. ایراد وارد بر این روشها آن است که تنها در محیط تعریف شده کاربری داشته و با تغییرات محیط (نظیر جریانات پیش‌بینی نشدهٔ ترافیکی، تغییر وضعیت خیابانها, وقوع تصادف و دیگر موارد اضطراری) که مغایر با مدل اولیه است سیستم توانایی ادارهٔ خود را از دست می‌دهد. دیگر مزیت روشهای تقویتی آن است نظارت و هدایتی در فرآیند یادگیری لازم ندارند. در روشهای یادگیری نظارتی (مانند شبکه‌های عصبی) نمونه‌های زیادی از وضعیتهای محیط و عملکرد مطلوب متناظر با آن نیاز است. این نمونه‌ها می‌بایست از نظر تنوع تمامی شرایط احتمالی از محیط که عامل با آنها روبرو می‌شود را در برگیرند. در مورد روشهای تقویتی کیفیت تصمیم‌گیری در برابر وضعیتهای مختلف از طریق روش آزمون و خطاو به صورت پویا فرا گرفته می‌شود. دیگر مزیت یادگیری تقویتی این است که نه تنها قابل تطبیق هستند, بلکه این تطبیق می‌تواند در تعامل پیوسته و بی‌درنگ با محیط انجام گیرد. اینگونه یادگیری, یادگیری *برخط*[1] نامیده می‌شود.

استفاده از یادگیری تقویتی در مسئلهٔ کنترل چراغ راهنمایی اولین بار توسط [۷] معرفی شد. نویسندهٔ این مقاله, چراغ راهنمایی را بصورت منفرد آموزش داده، و از آن در یک شبکهٔ ۴ در ۴ ترافیکی استفاده کرد. [۱۰,۹,۸] نیز از یادگیری تقویتی به منظور زمان-بندی چراغ‌راهنمایی استفاده کردند. مشکل تمامی روش‌های بالا حجم زیاد وضعیت‌ها است. آنها در مدل‌سازی روش خود از تعداد دقیق وسایل نقلیه عبوری در هر یک از مسیرها به عنوان وضعیت استفاده کردند. اتخاذ چنین مدلسازی علاوه بر افزایش تعداد وضعیت‌ها, براساس حس‌گرهای متعارف موجود در سطح شهرها , قابل پیاده‌سازی و استفاده عملی نمی‌باشد. بیشتر شهرها از حس-گرهای الکترومغناطیسی استفاده می‌کنند که در فاصل مختلف در کف خیابان بکار گذارده می‌شود. با توجه به تعداد کم این حس‌گرها در سطح خیابان‌ها و فاصلهٔ زیاد بین آنها, حس‌گرهای مذکور قادر به تشخیص دقیق تعداد ماشین‌های عبوری نبوده و تنها برآوردی از سرعت و حجم ترافیک را مهیا می‌سازند. در ادامه به معرفی دقیق‌تر روش یادگیری تقویتی مورد استفاده در راهکار پیشنهادی می-پردازیم.

۳- یادگیری تقویتی

در ساده‌ترین بیان، عامل یادگیری تقویتی شامل عاملی است که قصد دارد چگونگی نزدیک شدن به هدف را بوسیلهٔ تعامل پویایی که با محیط دارد فراگیرد. عامل، عمل های مختلف را در وضعیت‌های گوناگون امتحان می‌کند و پاداش (یا جریمه‌ایی) از محیط می‌گیرد که به نوعی معرف کیفیت وضعیت عامل تا این لحظه است. براساس پاداش‌های حاصل عامل در می‌یابد تا چه میزان هر عمل در نزدیک شدن به هدف نهایی اثر بخش است، بدین ترتیب عامل قادر خواهد بود بهترین عمل، یا سلسلهٔ اعمال در هر وضعیت ممکن، که سبب ماکزیمم شدن پاداش‌های دریافتی او می‌شود را کشف کند و از آنها برای رسیدن به هدف استفاده نماید.

مسئلهٔ زمان‌بندی چراغ‌راهنمایی با توجه به خصوصیت ذاتی محیطِ ترافیک، یک مسئلهٔ ممتد است. ما خواهان آنیم که عامل چراغ تقاطع به تنهایی, مدل محیط اطرافش را کشف کند، تا در صورت تغییرات در مدل نیاز به تغییر ساختار عامل نباشد. همچنین مایلیم عامل, در حین کنترل و در تعامل پیوسته و بی‌درنگِ خود با محیط, قابلیت یادگیری‌اش را حفظ کند, بنابراین از روش برخط

[1] Online

سارسا لاندا[۲] [۴], از سری روش‌های یادگیری تقویتی, بهره می‌گیریم. شکل (۱) نمایش دهندهٔ الگوریتم یادگیری سارسا لاندا جدولی, است.

- برای تمامی s,a ها, $Q(s,a)$ را به یک مقدار دلخواه و $e(s,a)$ را به مقدار صفر مقدار دهی اولیه کن.
- برای هر بازه, عملیات زیر را تکرار کن:
  - برای هر مرحله از بازه عملیات ۱ تا ۶ را تکرار کن:

۱. عمل a را از سری اعمال ممکن در وضعیت کنونی s و براساس خطِ مشی وابسته به $Q(s,a)$ برگزین, پاداش r و وضعیت آتی s' را مشاهده کن.

۲. عمل a' از وضعیت s' با استفاده از خطِ‌مشی وابسته به $Q(s,a)$ (نظیر خطِ مشی ε- حریصانه) برگزین.

۳. $\delta \leftarrow r + \gamma Q(s',a') - Q(s,a)$

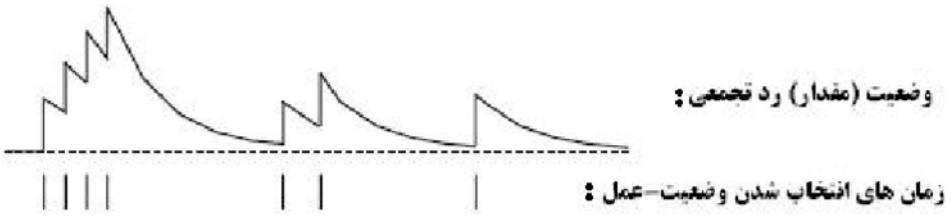
۴. $e(s,a) \leftarrow e(s,a) + 1$

۵. برای تمامی s,a ها موارد زیر را انجام بده:

$$Q(s,a) \leftarrow Q(s,a) + \alpha \delta e(s,a)$$
$$e(s,a) \leftarrow \gamma \lambda e(s,a)$$

شکل (۱): شبه‌کد سارسا لاندا جدولی[۴]

در این روش $Q(s,a)$ ارزش انتخاب عمل a در وضعیت s است. و α را اندازهٔ گام[۳] می نامند . معمولاً این پارامتر به صورت ثابت و در محدودهٔ $0 < \alpha \leq 1$ اختیار می‌شود و نرخ یادگیری را مشخص می‌کند. پارامتر γ، $(0 \leq \gamma \leq 1)$ نرخ کاهندگی نامیده می‌شود. نرخ کاهندگی[۴]، ارزش فعلی پاداش‌های آینده را مشخص می‌کند: ارزش پاداشی که در k بازه زمانی بعد دریافت می‌شود γ^{k-1} برابر ارزش پاداشی است که مستقیماً دریافت شده است. براین اساس اگر $\gamma < 1$ انتخاب گردد، مجموع نامتناهی پاداشها چنانچه دنباله پاداشهای $\{r_k\}$ کراندار باشد- مقداری متناهی خواهد بود. اگر $\gamma = 0$ اختیار گردد عامل نزدیک‌بین است, چرا که تنها سعی در ماکزیمم نمودن پاداش بلافاصل خود دارد. با نزدیک شدن مقدار γ به ۱ پاداش‌های آینده با قدرت بیشتری در تصمیم‌گیری‌ها شرکت کرده و عامل دورنگر خواهد بود. $Q(s,a)$ همچنین متأثر از $e_t(s,a)$ است که آنرا ردّ شایستگی[۵]می‌نامیم. ردّ شایستگی تاریخچه‌ای از وضعیت-عمل‌های انتخاب شده است و میزان تغییرات ارزش آنها را نسبت به هر ارزشی که عامل در آینده بدست می‌آورد, مشخص می‌کند. بروزرسانی ردّ شایستگی نشان داده شده در شکل (۱) را ردّ تجمعی[۶] گویند, زیرا هر بار که یک وضعیت و عمل رویت شد مقدار آن یک واحد افزایش می‌یابد و هنگامی که دیده نمی‌شود مقدار آن که با ضرب $\gamma\lambda$ کاهش می‌یابد و کم اثر می‌گردد. پارامتر λ را پارامتر ردّ کاهش[۷] می‌نامیم. شکل (۲) فرآیند مذکور را نمایش می‌دهد. عبارت جدولی در نام الگوریتم فوق به این معنا است که به منظور ذخیرهٔ ارزش و ردّ شایستگی هر جفت وضعیت-عمل یک خانهٔ حافظه در نظر گرفته می‌شود.



[۲] State Action Reward State Action (SARSA) (λ)

[۳] Step size

[۴] Discount rate

[۵] Eligibility trace

[۶] Accumulating trace

[۷] Trace-decay parameter

شکل (۲): ردّ تجمعی. با رویت هر وضعیت ردّ مربوطه یک واحد افزایش می‌یابد، اگر وضعیت رویت نشود با ضریب $\gamma\lambda$ کم اثر می‌گردد[۴].

۴- تعمیم[۸] و تقریب

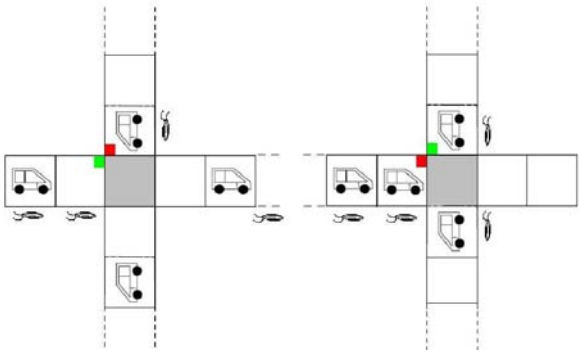
الگوریتم جدولی ارائه شده در شکل (۱) به علت حافظه, زمان و حجم دادهٔ مورد پردازش برای هر وضعیت-عمل (s,a), تنها در مسائل با دامنهٔ محدود وضعیت‌ها و عملگرها کاربرد دارد. در زمان‌بندی چراغ‌راهنمایی به علت بزرگی دامنه, با وضعیت‌های بسیاری مواجه می‌شویم که قبلاً به طور دقیق تجربه نشده‌اند. مفهوم کلیدی تعمیم چگونگی تقریب مناسبِ مجموعهٔ بزرگی از وضعیت‌ها بوسیلهٔ زیرمجموعهٔ کوچکتری از وضعیت‌های تجربه شده را بیان می‌کند.

توسعهٔ یادگیری تقویتی به استفاده از روش تعمیم و تقریب کاهش شیب [۴] امری آسان و کم هزینه است. در بین روش‌های کاهش شیب, نوع خطی از نظر تئوری ارضاء کنندهٔ نیاز است و در عمل چنانچه ویژگی‌های مناسبی انتخاب شوند, بخوبی کار خواهد کرد. از روش‌های خطی، کدگذاری کاشی‌کاری‌های مشبک همپوشان چندگانه[۹] [۴] که کاربرهای فراوانی دارد، را می‌توان نام برد. این روش دارای هزینهٔ پردازش بسیار کم و ثابت نسبت به بسیاری از روش‌های تعمیم نظیر کدگذاری درشت است و انتخاب ما در این مقاله قرار گرفت.

۵- شبیه‌سازی محیط ترافیک

یک عامل می‌بایست از جزئیات غیرضروری در حل مسئله پرهیز نماید و تنها قوهٔ درکی مناسب با کاربرد داشته باشد. بر این اساس ما از محیطی انتزاعی و متناسب با درک و عمل عامل و قابلیت‌های واقعی حس‌گرهای بکاررفته در شهرهایی نظیر شهر تهران و مشهد استفاده کردیم.

در محیط شبیه‌ساز مورد استفاده، وسایل نقلیه تنها براساس موجودیتشان به صورت میکروسکوپی و یا اتوماتیک نشان داده شده‌اند و از جزئیاتی نظیر نوع وسیلهٔ نقلیه، رنگ آن، طول آن و ... پرهیز شده است, شکل(۳)، چراکه فرض بر آن است که عامل محیط پیرامون خود را توسط حس‌گرهای ترافیکی متعارف که حساس به فلز بدنهٔ ماشین هستند، درک می‌کند. این حس‌گرها سازوکاری مانند دستگاه فلزیاب دارند و بعلت سادگی، کمی هزینه و قابلیت تغییری که دارند، در بسیاری از شهرها از جمله تهران از آنها استفاده می‌شود. جریان ترافیک شباهت بسیاری با جریان سیّال یا شن در لوله‌ها دارد [۱۱].



شکل (۳): دو تقاطع در محیط شبیه‌سازی.

[۸] Generalization

[۹] Multiple, overlapping grid tiling

مسیرها دارای اندازهٔ یکسان هستند و هیچ اولویتی بین آنها در نظر گرفته نمی‌شود. زیرا آنچه کـه بـرای عامـل اهمیـت دارد, نسـبت جریـان ترافیک مسیرهای منتهی به تقاطع به جریان اشباع آنها است و نه عرض و شکل مسیر.

در کنترل یک تقاطع منفرد می‌توان جریان‌ها و صف‌های مسیرهای شمالی- جنوبی و جنوبی- شمالی را با هم ادغام کرد و یا اثـر بخـش‌تـرین آنها را در تصمیم‌گیری منظور کرد. این موضوع در مورد مسیرهای شرقی- غربی و برعکس نیز صادق است. بنابراین در شبیه ساز مـورد اسـتفاده از مسیرهای یکطرفه استفاده کردیم.

در رنگ بندی چراغ‌های شبیه‌ساز چراغ زرد، از آنجاکه دارای زمان ثابت (برابر زمان تخلیهٔ تقاطع) اسـت و عملیـات یـادگیری بـر آن تـأثیر ندارد، نادیده گرفته شده و در شبیه‌سازی منظور نمی‌گردد.

۶- سناریوی یادگیری عامل چراغ راهنمایی

برای آموزش عامل‌ها، هر یک از آنها را به یک تقاطع شبکهٔ ترافیکی منسوب کرده و نرخ عبور را بوسیلهٔ مشخص کردن میانگین توزیع پواسون در ابتدای شریانهای شبکه مشخص می‌کنیم. در این سناریو لازم است عامل در هر قدم فراخوانی شود، تا بر اساس آخرین تغییرات در چگونگی ورود وسایل نقلیه تصمیمات خود را اتخاذ کند. وضعیتهای این مسئله از چگونگی ورود ماشین‌ها در هر مسیر از تقاطع و وضعیت چراغ تشکیل شده است. بنابراین وضعیت برداری سه مؤلفه‌ای شامل تخمینی از جریان شمالی–جنوبی, تخمین جریان شرقی–غربی و وضعیت چراغ خواهد بود. وضعیت چراغ دارای دو حالت است: ۱- سبز بودن در جهت شریان شمالی- جنوبی و ۲- قرمز بودن در جهت شریان شمالی–جنوبی. چگونگی ورود ماشین‌ها عمدتاً توسط کاراندازی حسگرهایی که به بدنهٔ اتومبیل حساسند, قابل حصول است.

از آنجا که بردار وضعیت ما در این مسئله برداری با طول ۳ است لازم است که فضای کاشی‌کاری ما, به منظور عملیات تعمیم و تقریب, نیز ۳ بعدی باشد. بعد سوم تنها شامل دو لبه است که با دو وضعیت چراغ متناظر است. از دو بعد دیگر یکی به روند عبور اتومبیل‌ها در مسیر شمالی–جنوبی و دیگری به روند عبور اتومبیل‌ها در مسیر شرقی–غربی اختصاص می‌یابد, شکل (۴).

شکل (۴): تناظر وضعیت به کاشی‌ها که حالتی نمایی دارند.

به کاشی‌های متناظر با بعد روند ترافیک, فاصلهٔ نزدیک‌ترین وسایل نقلیه به صورت نمایی(توانی از ۲) نگاشته می‌شود. به این معنی که اگر نزدیک‌ترین ماشین به تقاطع در فاصلهٔ ۱ اتومبیل تا تقاطع باشد، وضعیت به کاشی شمارهٔ یک نظیر می‌شود، و اگر نزدیک‌ترین ماشین به تقاطع در یکی از فاصله‌های ۲ ویا ۳ اتومبیل تا تقاطع باشد کاشی شمارهٔ ۲ فعال می‌شود. به همین ترتیب بسته به تعداد حس‌گرهای بکاربرده شده در مسیر و محل آنها تا تقاطع, کاشی‌ها تناظر می‌یابند. سایر وسایل نقلیه بنا بر فاصلهٔ حدودی که تا

تقاطع دارند وضعیت را به سمت مبدأ کاش‌کاری, که در منتهی‌الیه پایین و سمت چپ شکل (۴) است, می‌کشانند. به عنوان مثال, اگر ماشینی که نزدیک‌ترین ماشین به تقاطع نباشد در بین حس‌گر t و $t+1$ قرار گیرد، وضعیت را به اندازهٔ $\lfloor 10/2^{t+1} \rfloor$ به سمت مبدأ منتقل می‌کند. ۶ کاشی‌کاری از این نوع در روش تقریب کاشی‌کاری‌های مشبک همپوشان چندگانه در نظر گرفته شد. هریک از کاشی‌کاری‌ها در دو بعدِ مشخص کنندهٔ روند ترافیک, به اندازهٔ یک واحد کاشی با کاشی‌کاری قبلی خود فاصله داشته و در بعد سوم یعنی وضعیت چراغ با یک دیگر تفاوتی نمی‌کردند. اندازهٔ کاشی‌ها ۱۰ در نظر گرفته شد.

ویژگی‌های انتخابی تقریباً به چگونگی ورود ماشین‌ها و وضعیت چراغ تقاطع وابسته است و مستقل از وضعیت‌ها و حالات گذشته می‌باشد. علت بکار بردن کلمهٔ تقریباً و عدم قطعیت موضوع به دلیل آن است که ورود وسایل نقلیه در محدودهٔ نظارت عامل از دست عامل خارج است، ولی از آنجا که ماشین‌های تازه وارد به محدودهٔ سیستم معمولاً بعلت فاصلهٔ زیاد تا تقاطع اثر کمتری بر تصمیم‌گیری دارند، از این رو محیط را تقریباً ارضاء کنندهٔ خاصیت مارکوف می‌نامیم. بنابراین قادر خواهیم بود از الگوریتم تقویتی سارسا لاندا به منظور کنترل چراغ‌راهنمایی بر اساس تخمینی از شدت جریان عبوری ماشین‌ها استفاده کنیم.

برای هر وضعیت تنها دو عمل در نظر می‌گیریم:

سبز شدن چراغ برای مسیرهای شمالی- جنوبی و قرمز شدن چراغ برای مسیرهای شرقی- غربی.

سبز شدن چراغ برای مسیرهای شرقی – غربی و قرمز شدن چراغ برای مسیرهای شمالی-جنوبی.

پاداش‌ها به صورت جریمه معرفی می‌شوند. در این مسئله تعداد کل اتومبیل‌های صف انتظار عنصری از جریمه را تشکیل می‌دهد. تعداد کل اتومبیل‌های صف انتظار را بتوان ۳ می‌رسانیم تا بین وضعیت یک صف طولانی در یک شریان و یک صف کوچک در شریان دیگر تقاطع با و ضعیت صف‌های مساوی، ولی با همان مجموع طول صف قبلی، تمایز قائل شویم. عنصر دیگری که در جریمه‌ها مؤثر است تغییر رنگ چراغ است، زیرا هر تغییر رنگ چراغ زمان تلف شده‌ای را در بر دارد که هزینه بر است و مطلوب نمی‌باشد. برای هر تغییر رنگ در چراغ جریمه ۱ در نظر گرفته شد.

۷- نتایج تجربی

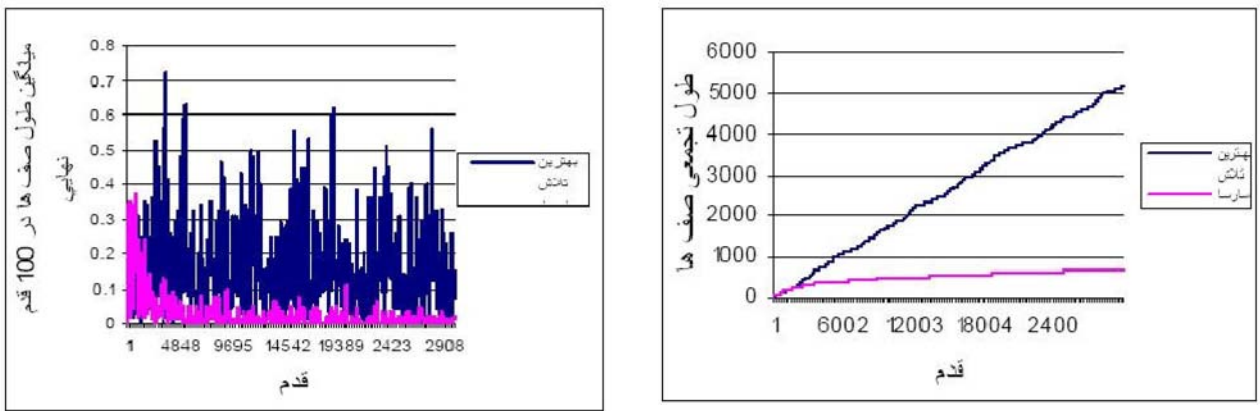
یکی از معیارهایی که می تواند ملاک ارزیابی ما از چگونگی عملکرد عامل باشد, طول صف‌های انتظار در پشت چراغ‌های قرمز است که خود نمادی از زمانهای تلف شده، مصرف بی‌جهت سوخت و تنشهای عصبی در جامعه است. شبیه‌ساز ما برای نمایش روند یادگیری از این پارامتر به دو صورت تجمعی و میانگین گیری در ۱۰۰ سیکل آخر استفاده می‌کند. انتظار می رود با همگرا شدن سیستم و نزدیک شدن به پایان فاز یادگیری، طول صف‌های انتظار کاهش یابد و شیب نمودار به سمت صفر مایل شود. واضح است که در حالت عادی این شیب هیچگاه صفر نخواهد شد, زیرا هنگامیکه دو ماشین همزمان در دو مسیر به یک چهار راه می رسند، چراغ راهنمایی تنها به یکی از آنها اجازهٔ عبور از تقاطع را داده و دیگری مجبور به ایستادن در پشت خط تقاطع می‌شود.

عامل را به صورت منفرد در یک محیط که از یک شریان آن ۶۰٪ جریان اشباع و از شریان دیگر آن ۴۰٪ جریان اشباع می‌گذشت قرار دادیم و محیط را به اندازهٔ ۳۰۰۰۰ واحد شبیه ساز اجرا کردیم. برای پارامترهای روش تقویتی سارسا لاندای مورد استفاده، مقادیر $\alpha = 0.4, \lambda = 0.4, \gamma = 0.3$ نتایج بهتری را به ما دادند. در خط‌مشی رفتاری نیز از شیوهٔ حریصانهٔ ε استفاده شد. در ابتدا اندازهٔ $\varepsilon = 1.0$ قرار داده شد که در هر بار اجرای عامل با ضریب ۰/۹۹۹۶ کاهش می‌یافت. باید توجه داشت که لازم است ε صفر نشود و دست‌کم دارای مقداری اندک باشد تا ضمن انتخاب بهترین عملکرد، جستجوگری خود را نیز در محیط پویایی مانند محیط ترافیک حفظ کند. شکل (۵) روند یادگیری را نشان می‌دهد.

(ب) (الف)

شکل (۵): نحوهٔ یادگیری عامل آزمایش بوسیلهٔ الف. نمایش چگونگی تغییر شیب نمودار طول تجمعی صف‌ها. ب. نمایش میانگین طول صف‌ها در ۱۰۰ قدم نهایی.

نتایج را با نتایج الگوریتم *بهترین تلاش*[۱۰] که در آن چراغ برای مسیر با بیشترین تعداد ماشین‌ها سبز می‌شود (زیرا امکان بوجود آمدن صف‌های طولانی در آن مسیر بیشتر است) مقایسه می‌کنیم , شکل (۶). نمودارها بهبودی در حدود۱۹.۶٪ توسط روش پیشنهادی را نشان می‌دهند.



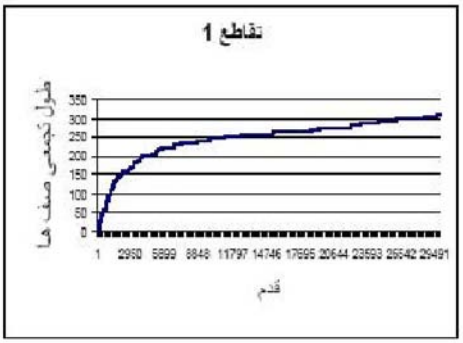
(ب) (الف)

شکل (۶): مقایسهٔ نمودار های روش سارسا با بهترین تلاش. رنگ آبی نتایج حاصل از اعمال الگوریتم بهترین تلاش و رنگ صورتی روش پیشنهادی ما رانشان می دهد. الف. طول تجمعی صف‌ها را به مرور زمان نشان می‌دهد. ب. میانگین طول صف‌ها در ۱۰۰ قدم آخر در هر قدم را نشان می‌دهد.

در آزمایش بعدی عامل فوق را در شبکه‌ایی مانند شبکهٔ واقعی شهری اجرای کرده و نتایج آنرا مشاهده می‌کنیم. شبکهٔ در نظر گرفته شده شامل ۶ تقاطع است که در شکل (۷) نشان داده شده است. در این شکل جریان ترافیک هر مسیر بر اساس درصدی از جریان اشباع در دایره‌های ابتدای مسیرها درج شده است. نتایج حاصل از یادگیری در شکل (۸) و شکل (۹) نشان داده شده است. همانطور که مشاهده می‌شود نمودارهای تمامی تقاطع‌ها همگرا شده و یادگیری عامل‌ها را نشان می‌دهند و در نتیجه آن در کل جریان ترافیک بهبود حاصل شده است.

[۱۰] Best efforts

شکل (۷): شبکهٔ تقاطع‌های مورد آزمایش.



شکل (۸): نمودارهای نمایش دهندهٔ طول تجمعی صف‌های تقاطع‌های شبکهٔ مورد آزمایش

شکل (۹): نمودارهای نمایش دهندهٔ میانگین طول صف‌های تقاطع‌های شبکهٔ مورد آزمایش در ۱۰۰ قدم نهایی

۷-۱- تأثیر پارامترهای مختلف

در این بخش به بررسی نحوهٔ تأثیر گذاری پارامترهای مختلف روش تقویتی در چگونگی یادگیری می‌پردازیم. آثار این پارامترها در زمان معرفی آنها در بخش‌های قبل از نظر تئوری بررسی شد، لازم است بعد از پیاده‌سازی عامل اثر هر یک عملاً دیده و بررسی شود. برای هر چه واضح‌تر شدن تغییرات میزان تعمیم کاسته شد تا اثر تغییرات در همپوشانی‌های تعمیم، پنهان نگردد. بنابراین به تعمیم اولیهٔ خود مانند آنچه در فصل قبل انجام دادیم روی آوردیم. عاملی که رفتار آنرا مورد مطالعه قرار می‌دهیم، عامل شمارهٔ ۳ از شبکهٔ شکل (۷) است. این انتخاب بدون هیچ دلیل خاصی انجام گرفت. نمودارهای چگونگی تغییرات طول صف‌های انتظار آن برای پارامترهای $\varepsilon_{1\to 0}, \alpha = 0.5, \lambda = 0.1, \gamma = 0.3$ در شکل (۱۰) نشان داده شده است. این نمودارها مبنای مقایسه با سایر شرایط که در ادامه خواهد آمد، قرار گرفته اند.

(الف) (ب)

شکل (۱۰): نمودارهای نمایش دهندهٔ روند یادگیری برای عامل تقاطع ۳ شکل (۷). الف. طول تجمعی صف‌ها. ب. میانگین طول صف‌ها در ۱۰۰ گام آخر در هر گام است

۷-۱-۱ تأثیر پارامتر ε

برای بررسی عملکرد پارامتر ε مقایسه‌ایی بین حالت $\varepsilon = 0$ و حالت شروع از $\varepsilon = 1$ و همگرا شدن به $\varepsilon = 0$ (که آنرا با $\varepsilon_{1\to 0}$ نشان می‌دهیم) ، انجام می‌دهیم. نتایج حاصل از $\varepsilon = 0$ در شکل (۱۱) قابل مشاهده است. در حالت $\varepsilon = 0$ زمان شروع همگرایی سریعتر از زمان شروع همگرایی در $\varepsilon_{1\to 0}$ است, زیرا در حالت $\varepsilon_{1\to 0}$ در زمانهای اولیه بیشتر انتخابها به صورت تصادفی است. هر چند در ابتدای فرآیند، عملگرهای انتخابی بیشتر اتفاقی و از روی تصادف هستند, ولی در بطن کار سیستم در حال یادگیری است.

(الف) (ب)

شکل (۱۱): نمودارهای نمایش دهندهٔ روند یادگیری برای عامل تقاطع ۳ شکل (۷) با $\varepsilon = 0$. الف. طول تجمعی صف‌ها. ب. میانگین طول صف‌ها در ۱۰۰ گام آخر در هر گام است

به عبارت دیگر یادگیری در مراحل اولیه تأثیر کمی در انتخابها دارد، بنابراین در مراحل اولیه حالت $\varepsilon_{1\to 0}$ دارای انباشتگی بیشتر در طول صفها خواهد بود که همگرایی کندتر را باعث می‌گردد.

اما همگرایی سریعتر $\varepsilon = 0$ نمایانگر کیفیت بالای همگرایی نیست، زیرا از بین عمل خوب و عمل بهتر ممکن است به تصادف عمل خوب ابتدا انتخاب شود و به علت عدم بکارگیری عمل بهتر و امتیاز گرفتن مضاعف عمل خوب از عمل بهتر این حالت تثبیت گردد. در سوی دیگر از آنجاکه در $\varepsilon_{1\to 0}$ اکثر قریب بالاتفاق حالات امتحان می‌گردند, این عامل تعدیل شده و نمودار از کیفیت همگرایی بهتری برخوردار می‌گردد به گونه‌ای که انباشتگی کمتری در طول صفها در ادامهٔ روند دیده می‌شود و به عبارتی $\varepsilon_{1\to 0}$ از $\varepsilon = 0$ از نظر کیفیت پیشی می‌گیرد. همچنین درحالت $\varepsilon = 0$ نمودار دارای شکستگی‌هایی بوده که نشان از تصمیمات نامناسب و عدم کیفیت مطلوب در تصمیمات است. این حالت در $\varepsilon_{1\to 0}$ از بین می‌رود و نمودار کمتر و در حد کوچک‌تر شکستگی می‌یابد، که حاکی از کیفیت بهتر در یادگیری است.

۷-۱-۲ تأثیر پارامتر α

برای هر محیط غیرپویا و نسبتاً یکنواخت اثبات می‌شود که اگر در روش تفاضل زمانی پارامتر α، اندازه قدم، به اندازهٔ کافی کوچک باشد به V^{π} همگرا خواهیم شد[۴]. نمودارهای بدست آمده از $\alpha = 0.1$ در شکل (۱۲), نتایج بهتری را نسبت به $\alpha = 0.5$ بدست می‌دهد، که گفتهٔ فوق را تأیید می‌نماید.

اما α کم دارای مقاومت بیشتر نسبت به اتخاذ تصمیمات جدید در برابر وضعیت‌های امتحان شده در قبل است. این استقامت در برابر یادگیری‌های جدید، عملکرد سیستم را در مقابل محیط‌های پر تغییر کاهش می‌دهد، آنچنان که در نمودارهای یادگیری به ازای

$\alpha = 0.1$ و $\alpha = 0.5$ این تفاوت دیده می‌شود, نتایج حاصل از $\alpha = 0.1$ در شکل (۱۳ الف) و $\alpha = 0.5$ در شکل (۱۳ب) نشان داده شده‌اند. این نمودار ها نمایش دهندهٔ عکس‌العمل عامل شمارهٔ ۳ شکل (۷)، به ازای تغیر جریان دو مسیر منتهی به آن، از ۲۰ به ۷۵ در مسیر شمالی-جنوبی و از ۳۰ به ۵۰ در مسیر شرقی-غربی در گام ۱۵۰۰۰ هستند.

در شکل (۱۳ الف)، $\alpha = 0.1$، پس از اعمال تغییرات در محیط، نمودار انحنای کم داشته که نشان از یادگیری کند آن نسبت به وضعیت جدید است. در حالیکه با $\alpha = 0.5$, شکل (۱۳ ب) این انحنا بیشتر بوده و نشان می‌دهد که عامل با سرعت بیشتر خود را با وضعیت جاری وفق می‌دهد.

(الف) (ب)

شکل (۱۲): نمودارهای نمایش دهندهٔ روند یادگیری برای عامل تقاطع ۳ شکل (۷) با $\alpha = 0.1$.الف. طول تجمعی صف‌ها. ب. چپ میانگین طول صف‌ها در ۱۰۰ گام آخر در هر گام است

(الف)

(ب)

شکل (۱۳): نمودارهای نمایش دهندهٔ روند یادگیری برای عامل تقاطع ۳ شکل (۷) با در تغییر محیط در گام ۱۵۰۰۰. الف. $\alpha = 0.1$. ب. $\alpha = 0.5$. ستون سمت راست طول تجمعی صف‌ها و ستون سمت چپ میانگین طول صف‌ها در ۱۰۰ گام آخر در هر گام را نشان می‌دهد.

۷-۱-۳ تأثیر پارامتر γ

γ میزان اثربخشی ارزش پاداش‌های آینده در روند تصمیم‌گیری فعلی است. در آزمایشها نشان داده شد مقدار دادن آن در حدود ۰/۵ کیفیت نسبی نمودار را منجر می‌شود و از طول صفهای ناگهانی به مرور می‌کاهد.تصدیق این موضوع در نمودار شکل (۱۴) نشان داده شده است.

زیاد نمودن این مقدار باعث کوچک شدن پاداش‌های بلافاصل در مقایسه با پاداش‌های آینده می‌شود که نتیجهٔ نامطلوبی به همراه دارد و کوچک‌تر کردن آن باعث می‌گردد که سیستم بیشتر بر اساس پاداش بلافاصل و بدون در نظر گرفتن وضعیتی که بعداً با آن دچار می‌شود، تصمیم گیرد. این وضعیت اثر خود را بیشتر در زمانی آشکار می‌کند که ماشینی در پشت چراغ وجود دارد و ماشینهای دیگر در حال اضافه شدن به صف هستند. تصمیم‌گیری بلافاصل ایجاب می‌کند تک‌ماشین در پشت چراغ قرمز بماند, زیرا جریمهٔ کمی دارد و از سوی دیگر تعویض چراغ هزینه‌بر است. ولی این حالت با رسیدن ماشینهای در راه جریمهٔ زیادی را به عامل تحمیل می‌کندکه با اثر دادن این وضعیت‌ها بوسیلهٔ γ از جریمه‌های زیاد که در اثر طول صف ایجاد می‌شود می‌توان کاست.

(الف) (ب)

شکل (۱۴): نمودارهای نمایش دهندهٔ روند یادگیری برای عامل تقاطع ۳ شکل (۷) با $\gamma = 0.5$. الف. طول تجمعی صف‌ها. ب. میانگین طول صف‌ها در ۱۰۰ گام آخر در هر گام است. که نمودار های صاف تر و بدون پرش‌های ناگهانی را نشان می‌دهد

۷-۱-۴ تأثیر پارامتر λ

پاداش حاصل بوسیلهٔ توانی از λ به وضعیت‌های گذشته انتقال می‌یابد، تا به یادگیری سرعت بخشد. مقدار زیاد آن به خصوص در مراحل اولیه که سیستم بیشتر از حالت تصادفی پیروی می‌کند باعث قوت بخشیدن به وضعیتهای نابهنگام چراغ می‌شود که هر چند تأثیری روی صفها ندارد, ولی از دید ناظر انسانی چندان خوشایند به نظر نمی‌رسد.

۸- نتیجه گیری

زمانبندی مناسب چراغهای راهنمایی یکی از مهمترین عوامل کنترل ترافیک محسوب می‌شود. عدم انعطاف الگوریتم‌های محاسباتی فعلی در رویارویی با شرایط متغییر ترافیکی, متخصصین امر را برآن داشت که به استفاده از سیستم‌های هوشمند و یادگیر روآورند. توانایی برخی از تکنیک‌های روش تقویتی در حل مسائل به صورت برخط و در تعامل بی درنگ با محیط, حتی اگر مدلی از محیط در اختیار نباشد, استفاده از این روش را یکی از گزینه های مطالعه قرار داده است. نشان دادیم چگونه مسئله کنترل ترافیک, به عنوان یک مسئله با پیچیدگی زیاد با بیانی ساده به فرم مسئلهٔ تقویتی تبدیل شد. نتایج حاصل از بهبود روند کنترل زمانبندی چراغ تقاطع حکایت می‌کردند. همچنین نشان داده شد چگونه چراغ‌راهنمایی با استفاده از این روش در برابر تغییرات نرخ ورود وسایل نقلیه از خود واکنش نشان داده و در صدد مدیریت هرچه بیشتر این رویداد گام بر می‌دارد. تأثیر پارامترهای مختلف در روند و کیفیت یادگیری ارزیابی شد و مشاهده گردید نتایج حاصل همانگونه است که از روی تئوری پیشبینی می‌شد.

مراجع

[1] Tan, K. K., Khalid, M., Yusof, R., "*Intelligent traffic lights control by fuzzy logic.*" Malaysian Journal of Computer Science, 9-2, 1995.

[2] Findler, N. and Stapp, J., "*A distributed approach to optimized control of street traffic signals.*" Journal of Transportation Engineering, 118-1:99–110, 1992.

[3] Tavladakis, K. and Voulgaris, N. C., "*Development of an autonomous adaptive traffic control system.*" In ESIT '99 - The European Symposium on Intelligent Techniques, 1999.

[4] Liu, H. L., Oh, J.-S., Recker, W., "*Adaptive signal control system with on-line performance measure.*" In 81st Annual Meeting of the Transportation Research Board,2002.

[5] Zhou, C., Nelson, P.C., "*Predicting Traffic Congestion Using Recurrent Neural Networks,*" 9th World Congress on Intelligent Transport Systems, Chicago, October 14-18, 2002.

[6] Sutton, R. S., Barto, A.G., "*Reinforcement Learning: an Introduction*", Cambridge: MIT Press, 1998.

[7] Thorpe, T. L., Andersson, C., "*Traffic light control using sarsa with three state representations.*" Technical report, IBM Corporation, 1996.

[8] Abdulhai, B., Pringle, R., and Karakoulas, G. J., "*Reinforcement Learning for True Adaptive Traffic Signal Control*" Journal of transportation engineering, ASCE, May /Jun 2003.

[9] Eagan, J., Lamstein, A., Mappus, Ch., Patrick, A., "*Applying Reinforcement Learning to Traffic Signal Timing Optimization,*" Jan 2004.

[10] Camponogara, E., Kraus Jr., W., "*Distributed learning agents in urban traffic control,*" Proc. 11th Portuguese Conference on Artificial Intelligence, Lecture Notes in Artificial Intelligence, pp. 324-335, Nov 2003.

[11] Wiering, M., Veenen, J. V., Vreeken, J., Koopman, A., "*Intelligent Traffic Light Control,*" Intelligent Systems Group, Institute of Information and Computing Sciences, Utrecht University, 2004.