مسیریابی وسایل‌نقلیه در سیستم هدایت مسیر پویا مبتنی بر یادگیری عاملهای هوشمند

عیسی نخعی کمال آبادی*، دانشیار، دانشکده فنی و مهندسی، دانشگاه تربیت مدرس، تهران، ایران
علیرضا عیدی، دانشجوی دکتری، دانشکده فنی و مهندسی، دانشگاه تربیت مدرس، تهران، ایران
E-mail:nakhai_isa@yahoo.com
تاریخ دریافت: ۸۷/۸/٦ – تاریخ پذیرش: ۸۸/۲/۱۵

چکیده

امروزه یکی از چالش‌های اصلی شبکه‌های ترافیکی، هدایت وسایل نقلیه به مقصدشان تحت وضعیت پویای ترافیک با هدف کاهش زمانهای سفر و استفاده موثرتر از ظرفیتهای موجود شبکه است. در پاسخ به مسایل بیان شده، سیستم هدایت مسیر پویا رویکردی موثر به نظر می‌رسد. این سیستم از جمله حوزه‌های مهم فعالیت سیستمهای هوشمند حمل‌ونقل است. هسته اصلی سیستم هدایت مسیر پویا، محاسبات کوتاه‌ترین مسیر بر اساس شرایط جاری (اطلاعات زمان واقعی) است. بنابراین با توجه به ضرورتهای بیان شده، هدف کلی تحقیق را می‌توان در قالب توسعه راهبرد قوی مسیریابی برای سیستم‌های هدایت مسیر تحت وضعیت پویای محیط تعریف کرد. به این منظور در این مقاله با بیان یک چارچوب مفهومی‌هدایت مسیر مبتنی بر ساختار مسیریابی غیرمتمرکز، به چگونگی کاربرد تکنیکهای عامل گرا با تاکید بر یادگیری تقویتی به عنوان یک راه حل در مواجهه با نامعینی‌های مسأله مسیریابی وسایل نقلیه در شبکه‌های ترافیکی پرداخته شده است. از نتایج مهم تحقیق ارایه شده می‌توان به توانائی مدلهای یادگیری پیشنهاد شده در ارایه سیاست یا استراتژی انتخاب مسیر در تطبیق با شرایط پویای ترافیکی و نیز ارایه آلترناتیوهای مختلف پیشنهادی طی مسیر برای رانندگان با هدف حداقل کردن معیار زمانهای سفر وسایل نقلیه اشاره کرد.

واژه‌های کلیدی:شبکه‌های حمل‌ونقل پویا، هدایت مسیر، الگوریتم‌های کوتاه‌ترین مسیر، عامل‌های هوشمند، یادگیری تقویتی، شبیه‌سازی

۱. مقدمه

امروزه یکی از مشکلات شهرهای بزرگ افزایش جمعیت و به تبع آن افزایش تعداد وسایل نقلیه به منظور جابجایی کالاها و مردم است که این امر موجب تراکم[1] و ازدحام در شبکه‌های حمل‌ونقل شهری می‌گردد. ساختن خیابانهای جدید شدیداً زمان بر و پرهزینه است. از سوی دیگر در خصوص تسهیلات موجود حمل‌ونقل شهری نیز به دلیل کمبود تجربه یا ناآشنایی برخی از رانندگان وسایل نقلیه(به ویژه وسایل نقلیه شخصی) با راهکارهای مختلف موجود برای طی مسیر، نمی‌توان سریع ترین مسیرها را انتخاب کرده و اغلب سفرها طولانی می شوند. نتیجتاً پیچیدگی سفرها در حال افزایش است. بنابراین از جمله چالشهای اصلی شبکه‌های ترافیکی، هدایت وسایل نقلیه به مقصدشان در وضعیت پویای ترافیک، با هدف کاهش زمانهای سفر و استفاده مؤثر از ظرفیتهای موجود شبکه است. با توجه به شرایط ذکر شده، رانندگان به منظور اجتناب از تراکم و انجام سفرهای راحت‌تر به خدماتی نظیر هدایت ترافیکی نیاز دارند. هدایت ترافیکی در پی توزیع مناسب جریانهای ترافیکی برروی کلیه مسیرهای شبکه حمل‌ونقل است که در نتیجه، کاهش آلودگی، کاهش مصرف سوخت و... را نیز در پی خواهد داشت.

در مجموع به‌منظور حل مسایل بیان شده، سیستم هدایت مسیر پویا[2] رویکردی مؤثر به نظر می‌رسد. سیستم هدایت مسیر پویا از جمله حوزه‌های مهم فعالیت سامانه‌های هوشمند حمل‌ونقل[3] است. این سامانه موجب ارتقای مطلوبیت زیرساختهای موجود شبکه حمل‌ونقل شده و به کنترل تراکم بر روی زمان و فضای در دسترس کمک می‌کند. سیستم هدایت مسیر همچنین قادر به اطلاع‌رسانی به

* نویسنده مسئول

رانندگان وسایل نقلیه مختلف درباره شرایط شبکه است. این اطلاع رسانی معمولاً از طریق تابلوهای پیام متغیر(VMS) یا به وسیله نمایشگرهای داخل وسیله نقلیه قابل انجام است. بازیابی اطلاعات همچنین می‌تواند توسط وب، موبایل و... برای کاربران فراهم شود.

سیستم هدایت مسیر به استفاده کنندگان از این سیستم، مسیری را از مکان فعلی به مقصد نهائی آنها پیشنهاد می‌کند، به گونه‌ای که زمان سفر برای رانندگان کمینه شود. هسته اصلی سامانه هدایت مسیر پویا، محاسبات کوتاه‌ترین مسیر در شبکه‌های پویاست. با استفاده از اطلاعات زمان واقعی، سیستم مورد اشاره قادر است کوتاه‌ترین مسیر از یک گره یا زون (منطقه) مبدأ را به گره یا زون مقصد برای رانندگان بر اساس شرایط جاری پیدا و به وسایل نقلیه توصیه کند. موثر بودن این وظیفه نیز عمدتاً به الگوریتم‌های کوتاه‌ترین مسیر وابسته است[۱،۲،۳،٤،٥]. به طور خلاصه، چارچوب کلی سامانه هدایت مسیر را می‌توان بوسیله شکل۱ بیان کرد.

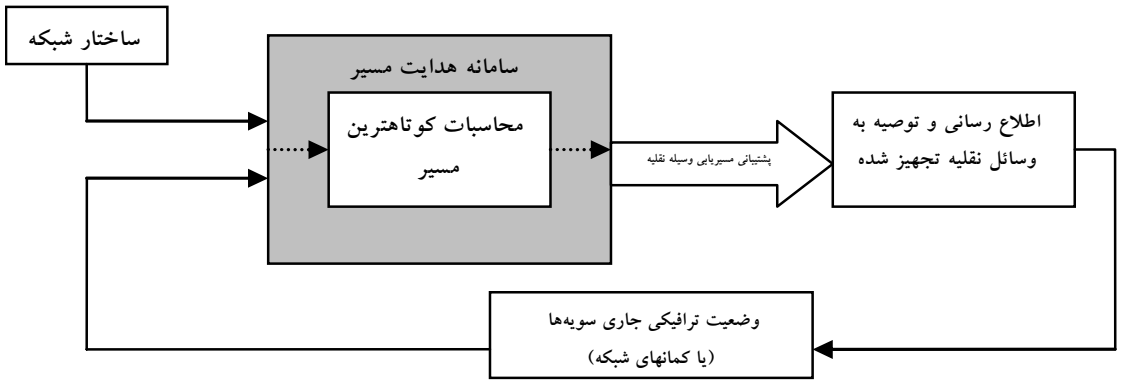
با توجه به طرح چالش‌ها و ضرورت‌های انجام تحقیق، مسأله تحقیق را می‌توان در قالب توسعه استراتژی قوی[٤] مسیریابی برای سامانه‌های هدایت مسیر تحت وضعیت پویای محیط تعریف کرد. بنابراین در این تحقیق، الگوریتم‌ها یا شیوه‌های محاسباتی مسیریابی در سیستم هدایت مسیر پویا بر مبنای تکنیکهای هوش مصنوعی مورد مطالعه دقیق تر و توسعه قرار خواهد گرفت تا در مواجهه با پویایی‌های محیط، سناریوها یا شق‌های مختلف طی مسیر برای رانندگان وسایل نقلیه فراهم شود.

در ادامه، ساختار مقاله شامل بخش‌های ذیل است: در بخش ۲ مرور ادبیات تحقیق مشتمل بر فرموله نمودن هدایت مسیر پویا، رویکردهای کوتاه‌ترین مسیر پویا در شبکه‌های حمل‌ونقل شهری مورد مطالعه قرار خواهد گرفت. متدولوژی تحقیق، فرضیات و مدل‌های ریاضی در بخش‌های ۳و٤ مقاله شامل ارایه چارچوب مفهومی‌سیستم هدایت مسیر وسایل نقلیه مبتنی بر تکنیکهای عامل‌گرا، بررسی مکانیزم یادگیری عامل‌های هوشمند با تاکید بر یادگیری تقویتی RL ارایه خواهد شد. به منظور مدلسازی کمی‌مسأله تحقیق در بخش ٥ پیاده سازی الگوریتم RL بر روی چارچوب مفهومی‌شامل اجزای RL برای مسأله تحقیق، فرآیند یادگیری عامل‌های هوشمند و ارایه الگوریتم یادگیری برای مسأله تحقیق انجام شده است. در بخش ٦ و۷ نیز با طرح یک مثال، ارزیابی مدلهای یادگیری در دسته‌ای از شبکه‌های واقعی (Grid Network) و قسمتی از شبکه ترافیکی شهر تهران همراه با محاسبات، ارایه خروجی‌ها و تحلیل‌های لازم انجام شده است. نهایتاً در بخش ۸ جمع بندی، طرح نتایج مهم، محدودیتها و برخی از افقهای تحقیقاتی آتی در ارتباط با موضوع مقاله بیان شده است.

۲. مرور ادبیات تحقیق

۲-۱ فرموله کردن هدایت مسیر پویا

پیش از ارایه تعریف ریاضی هدایت مسیر پویا، به معرفی زمانهای سفر زمان واقعی[٥] پرداخته می شود. یکی از مقولات اطلاعات زمان سفر مبتنی بر سامانه‌های هوشمند حمل‌ونقل، اطلاعات زمانهای سفر زمان واقعی است. زمانهای سفر زمان واقعی، اطلاعات پویای مرتبط با تغییرات در زمانهای سفر در زمان جاری را می‌تواند ارایه کند. استفاده از اطلاعات زمان واقعی با بکارگیری سیستمهای ویدئویی یا دوربین‌ها، حلقه‌های مغناطیسی، سیستم‌های موقعیت‌یاب جهانی(GPS) و سایر حسگرهای ترافیکی روی مسیرهای شبکه حمل‌ونقل امکان پذیر است[٦، ۷]. مسیریابی پویا که مورد نظر این تحقیق است به این دسته از اطلاعات نیاز دارد (مراجعه به ضمیمه A).



شکل۱. چارچوب کلی سامانه هدایت مسیر مبتنی بر اطلاعات زمان واقعی

در ادامه مجموعه‌های زیر تعریف می شوند: $S = \{s_{ij} | (i,j) \in A\}$ مجموعه مسافت سویه‌ها که از قبل در دسترس است و $D = \{d_{ij}(t) | (i,j) \in A\}$ مجموعه زمانهای سفر (وابسته به زمان) اکنون فرمولاسیون هدایت مسیر پویا بیان می‌شود:

فرض کنید G شبکه پویای جهت دار مبتنی بر یک نقشه الکترونیکی است به طوری که $N = \{1,...,n\}$ مجموعه گره‌ها و $A = \{1,...,m\}$ مجموعه سویه‌های جهت دار شبکه است. روی سویه‌های شبکه، تابع گسسته $d_{ij}(t)$ پس از تعداد مشخصی از بازه‌های زمانی M، عدد ثابتی را به عنوان مقدار یا ارزش می‌گیرد. از این رو $T = \{0,...,M-1\}$ مجموعه‌ای از بازه‌های زمانی حرکت یا عزیمت برای زمانهای سفر سویه است. همچنین MV حداکثر سرعت وسیله نقلیه در شبکه تعریف می‌شود.

مسأله هدایت مسیر پویا به دنبال انتقال وسایل نقلیه از یک مبدا به مقصد تعیین شده در امتداد کوتاه‌ترین مسیر تحت وضعیت پویای شبکه با هدف کمینه کردن مجموع زمانهای سفر است، بنابر این این مسأله را (با مبدا o و مقصد g) می‌توان در قالب یک مدل تحقیق در عملیات به شرح زیر معرفی کرد:

تابع هدف: $$\min \sum_{i=o}^{g} \sum_{j=o}^{g} d_{ij}(t).U_{ij}$$

محدودیتها: $$(s_{ij}/MV) \le d_{ij}(t) \quad ;t \in [0, M-1]$$

متغیرهای تصمیم: U_{ij} متغیر باینری : ۱ اگر سویه از گره i به گره j در مسیر طی شده وجود داشته باشد و الا صفر

به علت آنکه G مبتنی بر یک شبکه حمل‌ونقل واقعی است، شرط محدودیتها به این معنی است که s_{ij}/MV حد پائین زمان سفر بر روی سویه $i-j$ است[۴،۸،۹]. از جمله فرضیات دیگر که می‌توان به فرمولاسیون فوق اضافه کرد: شرط مستقل بودن سویه‌ها از یکدیگر است، به این معنی که تأثیر وضعیت ترافیکی سایر سویه‌ها بر سویه ای که عبور وسایل نقلیه از آن در حال بررسی است در نظر گرفته نمی شود.

۲-۱-۱ الگوریتم‌های کوتاه‌ترین مسیر پویا (DSP)[۶] در شبکه‌های حمل‌ونقل

در شبکه‌های حمل‌ونقل پویا، ارزش داده‌های شبکه یعنی هزینه سویه‌ها به زمان وابسته بوده و نسبت به زمان تغییر می‌کند. در کاربردهای حمل‌ونقل، به ویژه در شبکه‌های شهری که زمان گذر از سویه‌ها بستگی به حجم ترافیک دارد، به دلیل ضعف محاسباتی الگوریتمهای ایستا در شرایط زمان واقعی؛ می‌بایست شرایط شبکه‌های پویا برای محاسبه کوتاه‌ترین مسیر مورد بررسی قرار گیرد [۱۰، ۱۱،۱۲].

الگوریتم‌های DSP به طور وسیع در ادبیات موضوع مورد بررسی قرار گرفته و اثبات می‌شود که الگوریتم‌های استاندارد کوتاه‌ترین مسیر نظیر دیکسترا نمی‌تواند در یافتن مسیر با حداقل هزینه مورد انتظار روی شبکه غیرایستا استفاده شود و در این گونه مسایل انتخاب مسیر بهینه یک مسیر ساده نیست بلکه یک سیاست است که در برخی از تحقیقات از قواعد تصمیم گیری انطباقی در تعیین سیاست یا استراتژیهای مسیریابی استفاده شده است[۱۳].

نهایتاً اینکه به دلیل غیرکافی بودن مدلهای موجود، در حال حاضر جوابی مؤثر برای مسأله کوتاه‌ترین مسیر پویا در وضعیتی که ارزش سویه‌ها تحت تأثیر برخی از نامعینی‌ها نظیر اتفاق افتادن حوادث به شکل پویا تغییر می‌کند (که در عین حال ساختار شبکه ثابت[۷] است) وجود ندارد. نکته دیگر آنکه در این وضعیت به دلیل شرایط پیچیده حاکم بر شبکه؛ اطلاعات کاملی از تغییرات ارزش سویه‌ها در دسترس نیست. بنابراین به دلیل مشکل بودن محاسبات کوتاه‌ترین مسیر پویا در وضعیت اخیر، پیدا کردن فرمولاسیون جایگزین برای الگوریتمهای یافتن مسیر که تخمینهای خوبی از جواب بهینه فراهم کرده و زمان اجرای محاسبات را کاهش دهد، یکی از انگیزه‌های قوی برای تحقیقات در این زمینه است که فرمولاسیون جایگزین پیشنهادی در این تحقیق، بهره‌گیری از دسته ای از روشهای هوش مصنوعی یعنی عاملهای هوشمند است. در این تحقیق به دنبال سطحی از هوشمندی هستیم که عملکرد سامانه‌های هدایت مسیر از طریق ترکیب خودتشخیصی و یادگیری بر پایه اعمال گذشته بهبود یابد.

۳. چارچوب مفهومی[۸] سیستم هدایت مسیر وسایل نقلیه مبتنی بر عاملهای هوشمند

پیش از ارایه چارچوب مفهومی، مروری بر مطالعات انجام شده در زمینه کاربرد سیستمهای مبتنی بر عامل در مدیریت ترافیک ارایه می‌شود. بورمیستر و همکاران [۱۴] کاربرد سامانه‌های مبتنی بر عامل را در کنترل و بهینه سازی مدیریت ترافیک مطالعه کردند. در زمینه مسیریابی وسایل نقلیه، آدلر و بلو[۱۵] ضمن پیشنهاد رویکرد

سیستم چندعاملی، مطرح کردند وسایل نقلیه را می‌توان به‌عنوان عاملهای متحرک مدل کردند. آنها همچنین ادعا نمودند بهترین مسیریابی و زمانبندی وسایل نقلیه از طریق مذاکره بین عاملهای تامین کنندگان خدمات اطلاعاتی(ISP)[۹] و عاملهایی که ترجیحات رانندگان را بازنمایی می‌کنند به‌دست می‌آید.

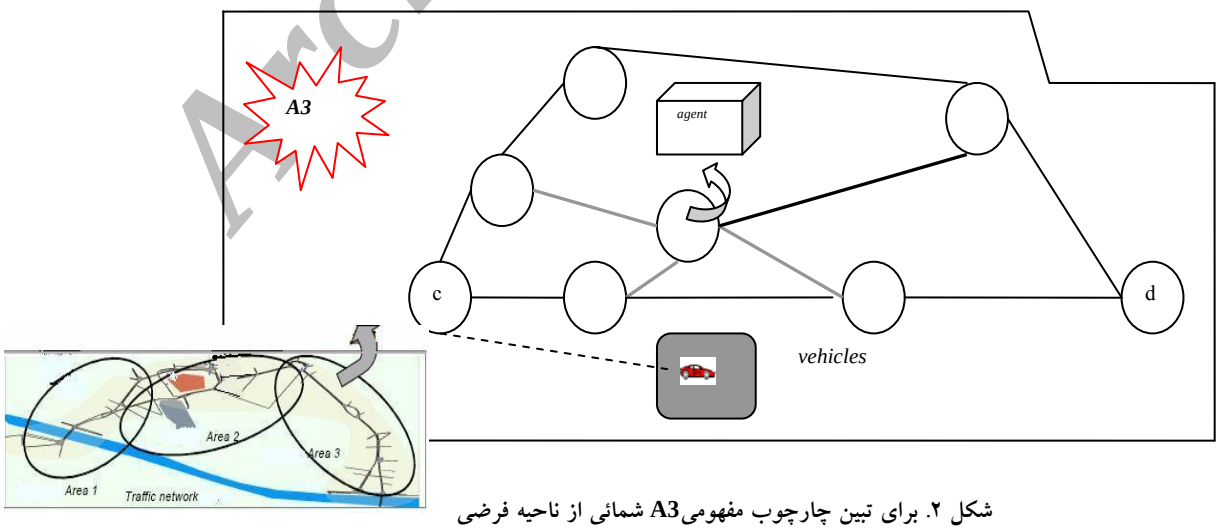
همچنین چابرول و همکاران [۱۶] در مطالعه‌ای از رویکرد چند عاملی به‌منظور مدلسازی و حل سیستمهای ترافیک شهری دفاع کرد. سی آیی و یانگ[۱۷] نیز در زمینه مدیریت هوشمند ترافیک به مطالعه مدیریت ترافیک شهری مبتنی بر سیستمهای چندعاملی پرداختند.

به هر حال مطالعات ذکر شده سعی کرده‌اند با بهره‌گیری از انعطاف‌پذیری مدلهای مبتنی بر عامل، مدیریت ترافیک را بهبود بخشند، هر چند که یک وظیفه مهم یعنی مسیریابی موثر وسایل نقلیه بر روی شبکه حمل‌ونقل در دسترس کمتر مورد تأکید بوده است. بنابراین این مسأله می‌تواند به‌طور دقیق تر مورد مطالعه و توسعه قرار گیرد.

اما در مورد چارچوب مدنظر این تحقیق، با توجه به قدمهای اولیه فرآیند برنامه ریزی حمل‌ونقل شهری، شبکه حمل‌ونقل شهری را به چند ناحیه جغرافیایی تقسیم کرده[۱۸] سپس یکی از نواحی مثلاً ناحیه فرضی و محدود(غیرگسترده) A۳ را که نقشه آن در اختیار است به عنوان یک گراف یا شبکه به شکل ۲ بازنمایی می‌کنیم. در این شبکه گره‌ها به عنوان تقاطع‌ها و سویه‌ها به عنوان آزادراهها یا خیابانهای بین تقاطع‌ها در نظر گرفته می‌شوند.

پس از ورود وسایل نقلیه به ناحیه فرضی A۳، وسایل نقلیه بر مبنای رویکرد مسیریابی غیرمتمرکز[۱۹،۲۰] از طریق تصمیم‌گیری‌های محلی توسط گره‌های شبکه، به مقصد یعنی گره d هدایت می‌شود. این شیوه مسیریابی نیازی به دسترسی به همه اطلاعات شبکه ندارد. گره شروع سفر وسایل نقلیه نیز گره c است.

حال به منظور تأثیرگذاری بر سیستم حمل‌ونقل، درک شرایط پویای محیط و توانایی تطبیق با برخی از نامعینی‌ها نظیر تغییرات در شرایط شبکه مثلاً اتفاق افتادن تصادفات- خرابی وسایل نقلیه- تعمیرات اضطراری خیابانها-... و نتیجتا تغییر زمانهای سفر روی سویه‌ها تحت تأثیر این قبیل نامعینی‌ها، بهنگام‌سازی اطلاعات ترافیکی گره‌های تصمیم‌گیر شبکه (اطلاعات محلی) بر اساس شرایط جاری به منظور انتخاب مسیرهای جایگزین یا آلترناتیو؛ می‌توان از تکنیکهای عامل‌گرا[۱۰] به عنوان یکی از مباحث متداول هوش مصنوعی(AI) استفاده کرد. به این منظور فرض می‌کنیم تعدادی عامل هوشمند(سیستم نرم افزاری یا سخت افزاری با قابلیت خودگردانی، یادگیری و...) به منظور کنترل و تصمیم گیری محلی در گره‌های مورد بررسی مستقر شده و در واقع گره‌های شبکه؛ هوشمند خواهند شد. این عاملها به منظور استفاده از تجربیات گذشته علاوه بر جمع آوری و ذخیره اطلاعات حمل‌ونقل، به یادگیری نیز می‌پردازند. از دیگر ویژگیهای مهم چارچوب ارایه شده در نظر گرفتن ارتباطات عاملها یا به عبارت دیگر بهره‌گیری از سیستمهای تعاملی یا چند عاملی به منظور مسیریابی وسایل نقلیه در شبکه است که در بخشهای آتی به آن پرداخته خواهد شد.



شکل ۲. برای تبین چارچوب مفهومی A3 شمائی از ناحیه فرضی

نکته دیگر آنکه در این چارچوب هر عامل به طور مستقل وظیفه اش را انجام می‌دهد، امّا برای رسیدن به هدف کلی با دیگر عاملها اندرکنش دارد. شکل ۳ چارچوب مفهومی پیشنهادی را نمایش می‌دهد.

بنابراین با بکارگیری عاملهای هوشمند در چارچوب ارایه شده، محاسبات لازم بر اساس اطلاعات ترافیکی زمان واقعی به منظور هدایت وسایل نقلیه در گره‌ها انجام شده و هر گره، مسیر بعدی را به وسایل نقلیه پیشنهاد می‌کند. بنابراین در بخشهای آتی به بررسی روشهای یادگیری عاملها پرداخته و سعی می‌شود این گونه روشها در بهنگام‌سازی دانش عاملهای هوشمند در مسأله تحقیق بکارگرفته شوند. ضمناً وسایل نقلیه از نظر فنآوری به گونه‌ای تجهیز می‌شود که امکان برقراری ارتباط با عاملها را داشته باشد. از دیگر مفروضات مسأله نیز این است که از ترجیحات مسیریابی رانندگان وسایل نقلیه صرفنظر می‌شود.

به طور خلاصه عاملهای بکارگرفته شده در سیستم هدایت مسیر وظایف زیر را برعهده دارند: انتقال اطلاعات به دست آمده از حسگرها، دریافت درخواست مسیر از وسایل نقلیه، محاسبه بهترین مسیر، فرستادن اطلاعات مسیر برای رانندگان وسایل نقلیه و دریافت / ارسال اطلاعات به دیگر عاملها (مشارکت با دیگر عاملها).همچنین چارچوب پیشنهادی دارای مزایای عمده زیر است:

الف. سیستم پیشنهادی دارای ساختار توزیع شده‌ای است که این امر سرعت محاسبات را در شبکه‌های بزرگ مقیاس افزایش می‌دهد.

ب. سیستم پیشنهادی از توانایی یادگیری برخوردار است لذا می‌تواند تحت تأثیر نامعینی‌ها و تغییرات محیط تطبیق یابد.

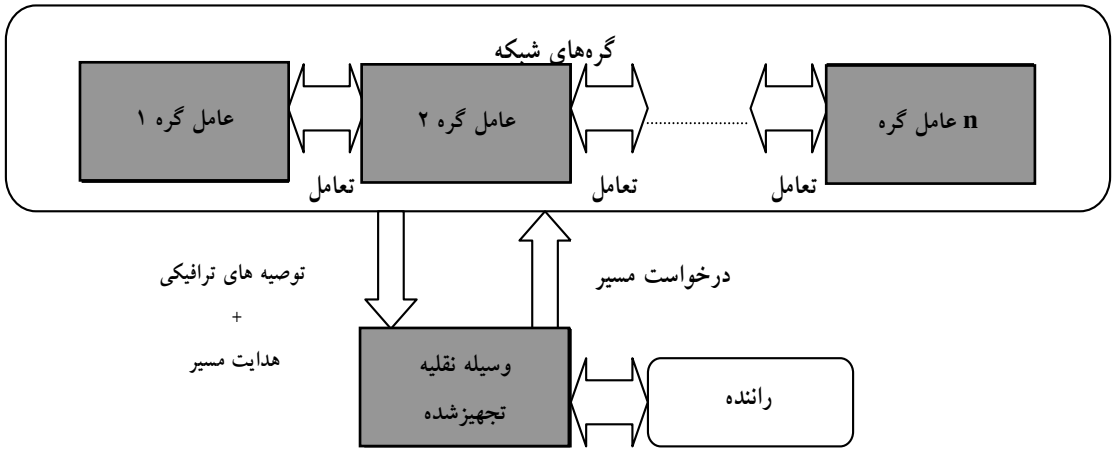
ج. برتری دیگر سیستم پیشنهادی اینکه می‌تواند با بهره‌گیری از همکاری و هماهنگی بین عاملها از ثبات لازم برخوردار باشد.

در مجموع از اهداف مهم تحقیق، کنترل جریان ترافیکی از طریق پیاده سازی یک مدل یادگیری هوشمند روی گره‌های شبکه حمل‌ونقل پویاست که وظیفه عبور وسایل نقلیه را برای طی مسیر تا رسیدن به مقصد برعهده دارند.

٤. بررسی مکانیزم یادگیری عاملها و مدلسازی کمی مسأله

در اغلب سیستمها و حتی سیستمهای نسبتاً ساده، تعیین دقیق رفتار و فعالیت یک مجموعه چند عاملی غیر ممکن است. تعیین دقیق رفتار مستلزم آگاهی از شرایط (چه بیرونی و چه درونی) است که در محیط‌های پویا تقریباً ناممکن است. این قبیل مشکلات سیستمهای چند عاملی را می‌توان به وسیله توانایی یادگیری و سازگاری عاملها برطرف ساخت.

یکی ازمکانیزم‌های یادگیری سیستمهای چند عاملی روش یادگیری تقویتی یا تشدیدی[11] است که در آن بازخورد سیستم یادگیری، سودمندی فعالیت انجام شده را نشان داده و هدف از سیستم یادگیری، بیشینه کردن سودمندی فعالیت در طولانی مدت است[۲۱، ۲۲، ۲۳].



شکل۳. نمایشی از چارچوب مفهومی پیشنهادی مبتنی بر سیستمهای چند عاملی

٤-١ تعامل عامل- محیط و اجزای RL

در مسأله یادگیری تشدیدی موجودیت‌هایی که تعامل عامل با آنها صورت می‌گیرد مشتمل بر هر چیزی بیرون از عامل است که محیط نامیده می‌شود. در این تعامل، عامل دائما اعمالی را انتخاب کرده و محیط به این اعمال پاسخ داده؛ موقعیت یا حالت جدید را به عامل عرضه می‌کند. شکل ٤ نشان دهنده محیط تعاملی عامل است.

چهار جزء اصلی سیستم‌های یادگیری تشدیدی، ماورای عامل و محیط، عبارتند از: سیاست[١٢] عامل، تابع پاداش[١٣] ، تابع ارزش[١٤] و مدلی از محیط. سیاست عامل با $\pi_t(s,a)$ نشان داده می‌شود که عبارتست از احتمال اینکه در دوره t که حالت سیستم s است عمل a انجام شود. روشهای یادگیری تشدیدی مشخص می‌کنند که عامل چگونه سیاست‌های خود را بر طبق تجربیاتش تغییر دهد.

تابع پاداش تعریف کننده اهداف در مسأله یادگیری تشدیدی است. در حقیقت یک تابع پاداش نگاشتی است که حالت مشاهده شده (یا زوج حالت–عمل) را به یک عدد واحد که پاداش نامیده می‌شود می‌نگارد که این عدد نشان دهنده میزان مطلوبیت حالت است.

در حالی که تابع پاداش، نشان دهنده انتخاب‌های خوب به صورت بلافاصله فوری است، تابع ارزش مشخص کننده انتخاب‌های خوب در طولانی مدت است. ارزش یک حالت عبارتست از حجم کلی پاداشی که یک عامل می‌تواند توقع داشته باشد تا در آینده با شروع از حالت مذبور اندوخته شود. انتخاب اعمال بر اساس قضاوت در مورد ارزش آنها است[٢٤، ٢٥].

٤-٢ فرآیند تصمیم‌گیری مارکوفی[١٥]

در مسأله یادگیری تشدیدی در صورتی که تعداد محدودی حالت و ارزش وجود داشته باشند و فرض شود که خاصیت مارکوفی برای حالات وجود دارد، می‌توان گفت احتمال اینکه در حالت بعدی یک عمل خاص انجام شود تنها وابسته به حالت فعلی بوده و مستقل از مسیری است که به حالت فعلی منجر شده است، به عبارت ریاضی:

$$\Pr\{s_{t+1}=s', r_{t+1}=r \mid s_t, a_t\} \qquad (١)$$

حتی وقتی سیگنال حالت غیرمارکوفی باشد مناسب است در یادگیری تشدیدی فرض شود که این سیگنال شباهت زیادی به حالت مارکوفی دارد[٢٥].

٤-٣ تخمین تابع‌های ارزش

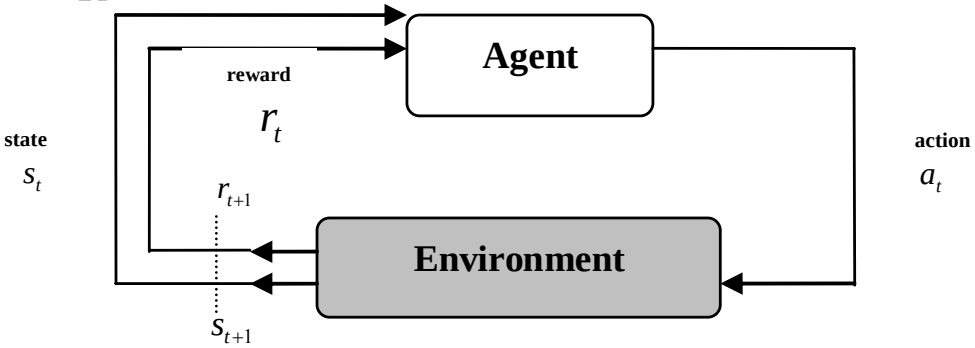
حل اغلب الگوریتمهای یادگیری تشدیدی بر پایه تخمین تابع‌های ارزش است. برای فرآیندهای تصمیم گیری مارکوفی، ارزش عمل انجام شده a تحت سیاست π با $Q^{\pi}(s,a)$ معرفی شده و به عنوان بازگشت مورد توقع در زمانی که حالت اولیه s بوده و عمل a انجام شده و سیاست π دنبال شود به صورت زیر تعریف می‌شود:

(٢)

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

در رابطه ٢، R_t اصطلاحاً بازگشت مورد انتظار و γ ضریب تنزیل پاداش در طول زمان (که ارزش حال حاضر پاداش‌های آینده را تعیین می‌کند) و دارای مقداری بین صفر و یک است. k نیز تعداد گامهای زمانی در آینده است.

همان طور که مشخص است به منظور تعیین ارزش یک عمل در یک حالت خاص می‌بایست یک دنباله از پاداش‌ها در یک سیاست محاسبه شوند و از آنجاکه دنباله پاداش‌ها بستگی به حالت سیستم دارد و حالت سیستم نیز خود تابعی از شرایط محیط است که در اغلب مسایل مورد بررسی نامعین است، در نتیجه نمی‌توان به صورت قطعی مقدار رابطه ارایه شده برای تابع ارزش عمل را محاسبه کرد. به همین دلیل به منظور تعیین ارزش عمل a در حالت s از روابط تخمینی استفاده می‌شود[٢٥].



شکل ٤. محیط تعاملی عامل در یادگیری تشدیدی

٤-٤ روش‌های یادگیری اختلاف موقتی(TD)[16]

روش‌های TD به عنوان یکی از روشهای پایه‌ای حل مسأله یادگیری تشدیدی ایده‌ای است که به عنوان هسته مرکزی در یادگیری تشدیدی شناخته شده است. روش‌های مذکور می‌توانند مستقیماً از تجربیات و بدون نیاز به مدلی از پویایی محیط بیاموزند و تخمین‌ها را بر پایه سایر تخمین‌های یاد گرفته شده بهنگام‌سازی کنند[۲۵].

یکی از مهم‌ترین روشهای اختلاف موقتی، الگوریتم *SARSA* است که شیوه یا سیاست on-policy (تخمین جاری از یک سیاست غیر بهینه برای حرکت به سمت سیاستهای موجود بهینه) را در استفاده از تجربیات بکار می‌گیرد. در روش *SARSA*، گذار از یک جفت عمل- حالت به جفت عمل- حالت جدید بررسی شده، ارزش جفت عمل- حالت آموخته می‌شود. ایده روش مذکور، بهنگام‌سازی تخمینهای $Q(s,a)$ برمبنای تخمینهای قبلی $Q(s,a)$ است. رابطه کلی بهنگام‌سازی مورد استفاده این روش به صورت زیر(رابطه۳) قابل بیان است:

(۳)

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t)+\alpha[r_{t+1}+\gamma Q(s_{t+1},a_{t+1})-Q(s_t,a_t)]$$

که در آن α نرخ یادگیری بوده و چگونگی بهنگام‌سازی تابع ارزش را کنترل می‌کند. نرخ یادگیری بزرگ‌تر به عامل اجازه می‌دهد که عمل بهینه را در زمان آموزشی کوتاه‌تر تشخیص دهد. همچنین مقادیر بزرگ γ به این معنی است که عامل یاد بگیرد که برای پاداشهای طولانی مدت عمل کند[۲٦، ۲۷، ۲۸]. در حقیقت رابطه ۳ بعد از هر گذار از یک حالت غیر پایایی بکار می‌رود. در صورتی که s_{t+1} حالت پایانی باشد آنگاه $Q(s_{t+1},a_{t+1})$ معادل با صفر تعریف می‌شود. از لحاظ ریاضی ثابت شده است که این روش با احتمال ۱ به سیاست بهینه و به توابع ارزش- عمل بهینه همگرا می‌شود[۲۵، ۲۹].

٤-٥ استراتژیهای انتخاب عمل

یکی از مهم‌ترین تنظیمات یک سیستم یادگیری تشدیدی ایجاد تعادل بین جستجو[17] و بهره‌برداری[18] است. یک عامل علاوه بر اینکه از دانش جاری درباره محیط و از نتایج یادگیریهای خود بهره برداری می‌کند می‌بایست همواره در پی جوابهای بهتر از طریق جستجو به منظور بهبود در کیفیت تصمیماتش باشد[۲۵، ۳۰].

انتخاب عمل از سوی عامل به چندین روش مختلف صورت می‌پذیرد. ساده‌ترین آنها انتخاب عمل به روش حریصانه[19] است. در این روش معمولاً بهترین عمل یعنی عملی با بالاترین ارزش ممکن برای حالت – عمل انتخاب می‌شود (روش بهره برداری خالص). دو روش پیچیده دیگرعبارتست از: در روش $\varepsilon-greedy$ عامل با احتمال کوچک اپسیلون، یک عمل را به طور تصادفی و مستقل از تخمین ارزش عمل انتخاب می‌کند که با افزایش تعداد تجربیات یا آزمایشهای در طول زمان، احتمال انتخاب بهترین عمل به عددی بزرگتر از $\varepsilon-1$ که $\varepsilon \in [0,1]$ همگرا می‌شود. اما در روش سوم یعنی *Softmax* توزیع احتمال انتخاب عمل به اسکالری تحت عنوان درجه حرارت T بستگی دارد. دما پارامتری مثبت است و عامل بر اساس توزیع احتمال زیرکه به وسیله فرمول بالتزمن[20] تعریف می‌شود عملی را با بیشترین احتمال انتخاب می‌کند که منتهی به بیشترین ارزش مورد انتظار خواهد شد:

$$p(a_i)=(e^{Q(s,a_t)/T})/(\sum_{j=1}^{S_k} e^{Q(s,a_j)/T}) \qquad (٤)$$

$p(a_i)$ احتمال انتخاب عمل a_i و s_k تعداد اعمال انتخاب شده در حالت s است. وقتی که دما بالاست قابلیت یادگیری نیز بالاست و در ادامه که دما کاهش می‌یابد سیستم بیشتر به سمت استفاده از دانش خود یعنی بهره برداری می‌رود. از مزیتهای روش فوق اینکه تنها با استفاده از یک پارامتر T می‌توان بین جستجو و بهره برداری تعادل برقرار کرد[٤، ۳۱، ۳۲].

۵. پیاده‌سازی الگوریتم RL بر روی چارچوب مفهومی

پس از مروری نسبتاً جامع بر RL و تکنیک‌های مرتبط با آن، در این قسمت از تحقیق، به مدلسازی کمی مسأله تحقیق پرداخته خواهد شد.

۵-۱ اجزا یادگیری تشدیدی برای مسأله تحقیق

با توجه به تعاریف ارایه شده در مورد اجزای یادگیری تشدیدی در بخشهای قبلی، در جدول۱ به طور خلاصه اجزا RL برای مسأله مورد مطالعه ارایه شده است[۲۱، ۳۳، ۳٤].

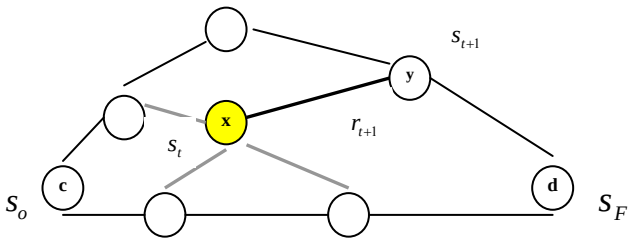
جدول۱ . اجزا RL برای مسأله مورد مطالعه

اجزاRL	اجزا مسأله
محیط	شبکه حمل ونقل(که بشکل پویا تغییر می‌کند)
عامل Agent	مستقر در گره های شبکه
حالت در زمان S_t :t	گره ای که از طریق آن وسایل نقلیه عبور می‌کند(گره جاری)
سیاست / عمل a_t	استراتژیهای مسیریابی / هدایت وسایل نقلیه از گره جاری به سوی مسیرها برای رسیدن به گره بعدی (یا آتی)
تابع پاداش r_t	زمان واقعی سفر روی سویه ها یا مسیرهای شبکه
تابع ارزش $Q_d(s_t,a_t)$	حجم کلی پاداش مورد توقع(قابل تخمین با روشهای TD) : تخمین مجموع زمان واقعی سفر از گره جاری تا گره مقصد یا هدف[۲۱]

در تعریف تابع پاداش علاوه بر عامل زمان سفر می‌توان عواملی دیگر نظیر هزینه سفر، آسایش و راحتی و سطح سرویس را نیز اضافه کرد. با وجود اینکه عوامل مذکور دارای اهمیت هستند اما فاکتور زمان سفر تقریباً در تمامی‌مدلهای مربوط به انتخاب مسیر به عنوان نماینده ای از همه فاکتورها مورد استفاده قرار می‌گیرد که دلیل اصلی آن نیز سهولت اندازه گیری زمان سفر در مقابل متغیرهای دیگر است[۱۸].

۵-۲ فرآیند یادگیری عاملهای هوشمند

با توجه به مطالب بخش‌های قبلی، اکنون فرآیند یادگیری و بهنگام‌سازی دانش عاملها (یا گره‌ها)، مبتنی بر اطلاعات و شرایط جاری بررسی می‌شوند. به‌این منظور شکل ۵ در نظر گرفته می‌شود:



شکل۵ . توپولوژی یک شبکه فرضی به منظور تشریح فرآیند یادگیری

با توجه به توضیحات بیان شده در مورد تابع ارزش، در شکل ۵ هر گره نظیر x (با حالت s_t) تخمینی از تابع ارزش در زمان t یعنی $Q_d(s_t,a_t)$را ذخیره می‌کند. بنابراین تابع ارزش برای گره y با حالت s_{t+1} معادل $Q_d(s_{t+1},a_{t+1})$خواهد بود. تابع ارزش برای گره مقصد d نیز معادل صفر تعریف می‌شود.

با توجه به هدف مسیریابی که رسیدن وسایل نقلیه به مقصد نهایی در زودترین زمان ممکن است بنابراین راهبرد یا سیاستی اتخاذ می‌شود که وسایل نقلیه از طریق گره همسایه‌ای عبور داده شوند که کمترین تخمین تابع ارزش را داشته باشد.

حال فرض کنید با توجه به شکل ۵ در نظر است هدایت وسایل نقلیه را از گره مبدا c به گره هدف یا مقصد d بررسی شود. حالت گره مبدا s_o و حالت گره نهائی یا مقصد را s_F در نظرمی‌گیریم. بر اساس سیاست تعریف شده فرض کنید در بین گره‌های همسایه x،گره y دارای کمترین تخمین زمانی برای رسیدن به مقصد باشد، بنابراین گره y برای ادامه حرکت وسایل نقلیه انتخاب شده و حالت بعدی، گره y خواهد بود. تابع پاداشی که فورا توسط گره y مشاهده می‌شود به شرح رابطه ۵ است:

$$r(s_t,a_t)=r_{t+1}=time_{link}^{(x,y)} \qquad (۵)$$

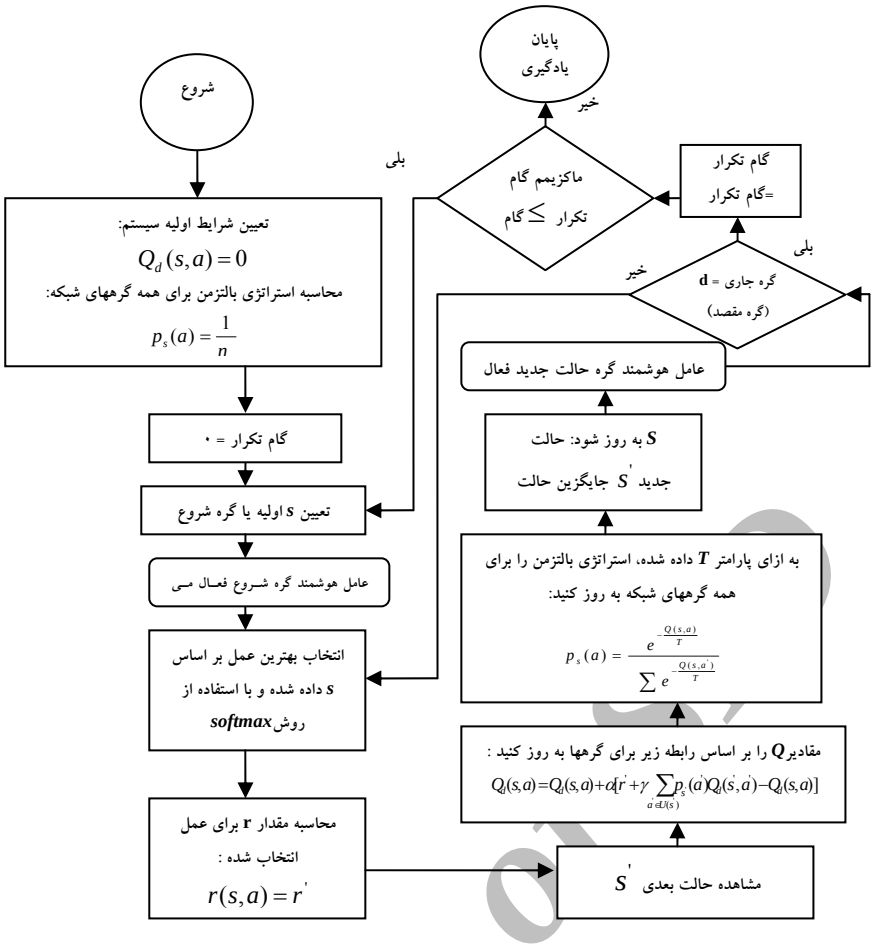
در رابطه فوق مقدار تابع پاداش معادل زمان واقعی سفر وسایل نقلیه برای عبور از سویه $x-y$است.

راهبرد انتخاب عمل:

در روش یادگیری تشدیدی، از جمله تعاریف جستجو از لحاظ تکنیکی و محاسباتی، ارایه مجموعه‌ای از اعمال در دسترس برای هر حالت است. بدون جستجو یا اکتشاف، سیستم هدایت مسیر؛ وسایل نقلیه را منحصراً در امتداد بهترین مسیرهای جاری به سوی مقصد می‌فرستد بدون آنکه مسیرهای دیگری را به عنوان آلترناتیو یا کاندیدا به منظور تطبیق با تغییرات ممکن در محیط کشف کند. با توجه به مطالب ارایه شده در مورد استراتژیهای انتخاب عمل و تحقیق انجام شده توسط آچیانی و همکاران[۳۵] در مورد بهترین استراتژی انتخاب عمل در مسایل کوتاه‌ترین مسیر در یک شبکه، در ادامه استراتژی *softmax* با توزیع احتمال بالتزمن به منظور انتخاب عمل ترجیح داده شده و بر اساس فرمول زیر مورد استفاده قرار می‌گیرد:

$$p_{s_t}(a_t)=\frac{e^{-Q(s_t,a_t)/T}}{\sum_{a_{t+1}\in U(s_t)} e^{-Q(s_t,a_{t+1})/T}} \qquad (۶)$$

در این فرمول که برای موارد هزینه‌ای ارایه شده است $U(s_t)$ مجموعه همه اعمال در دسترس در حالت s_t است (مراجعه به ضمیمهB).



شکل٦. فلوچارت الگوریتم یادگیری برای مساله تحقیق

نهایتاً وقتی در مورد استراتژی ادامه حرکت وسایل نقلیه تصمیم گیری شد؛ لازم است مقادیر تابع ارزش روی گره x، گره y و سایر گره‌های همسایه بهنگام‌سازی شود. با بهره‌گیری از روش *SARSA* به عنوان یکی از روشهای اختلاف موقتی، بهنگام‌سازی اطلاعات مسیریابی طبق رابطه ۳ انجام می‌شود.

در رابطه مذکور ضریب تنزیل γ معین می‌کند که پاداش‌های آتی را تا چه فاصله ای (از لحاظ زمانی) در محاسبات وارد کنیم. همان طور که قبلا گفته شد مقادیر γ بین صفر و یک است. در وضعیت $\gamma=0$ عامل به دنبال به حداکثر رساندن پاداشهای فوری است. همچنین در وضعیت $\gamma\to1$ هدف این است که پاداشهای آتی قویا یا بدون تنزیل در محاسبات تخمین تابع ارزش وارد شوند. بنابراین به منظور بررسی مسأله در وضعیت بدون تنزیل رابطه ۳ به صورت زیر اصلاح می‌شود:

$$Q_d(s_t,a_t)=Q_d(s_t,a_t)+\alpha[r_{t+1}+Q_d(s_{t+1},a_{t+1})-Q_d(s_t,a_t)] \quad (۷)$$

از جمله ویژگیهای مهم چارچوب *SARSA* جستجو یا اکتشاف online در مورد انتخاب عمل است، به این معنی که به طور پیوسته یا دوره‌ای می‌تواند اکتشاف مجدد را بر روی محیط ناشناخته انجام دهد. بنابراین به منظور تخمین online حالت آتی و با بهره‌گیری از استراتژی بالتزمن، قانون بهنگام‌سازی *SARSA* را به شکل زیر تعدیل کرده و در ادامه برای انجام محاسبات از آن استفاده می‌شود:

$$Q_d(s_t,a_t)=Q_d(s_t,a_t)+\alpha[r_{t+1}+\gamma\sum_{a_{t+1}\in U(s_{t+1})}p_{s_{t+1}}(a_{t+1})Q_d(s_{t+1},a_{t+1})-Q_d(s_t,a_t)] \quad (۸)$$

مفهوم فرمول فوق این است که می‌توان عدم قطعیت در حالت آتی یعنی در گره S_{t+1} را از طریق بکارگیری توزیع احتمالات عمل‌های مختلف و امکان پذیر در آن حالت با استفاده از تعریف استراتژی بالتزمن تخمین زد.

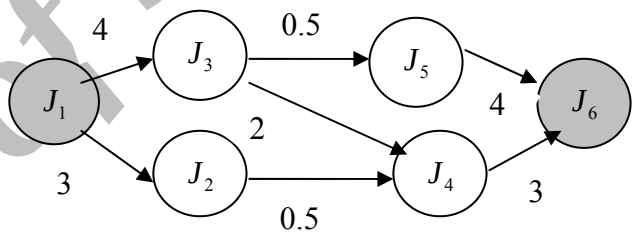
۵-۳ الگوریتم یادگیری برای مسأله مورد بررسی

در این قسمت با توجه به شناسایی اجزای الگوریتم RL در ارتباط با مسأله تحقیق و نیز فرآیند یادگیری ارایه شده در بخش قبل، فلوچارت الگوریتم یادگیری عاملهای شبکه حمل‌ونقل را به منظور هدایت وسایل نقلیه برای رسیدن به گره مقصد ارایه می‌شود.

بر طبق فلوچارت الگوریتم شکل٦، قدمهای آن به تعداد دفعات مشخصی تکرار می‌شود و هدف از تکرارها تعیین مقادیر نهایی $Q_d(s,a)$است. در واقع این مقادیر سیاست یا استراتژی بهینه عامل را در قبال حالتهای مختلف محیط توصیه می کنند، به قسمی‌که بیشترین ارزش را در پی داشته باشد.

٦. انجام کدگذاریهای مسأله با ارایه یک مثال

به منظور آشنایی با کدگذاریهایی که طی پیاده سازی فلوچارت الگوریتم یادگیری انجام می‌شود شبکه مثال زیر (شکل۷) در نظر گرفته می‌شود:



شکل۷. توپولوژی شبکه مورد استفاده در مثال بخش ٦

مقادیر اولیه Q را برای هریک از جفت‌های در دسترس حالت-عمل در مورد عامل‌های هوشمند مستقر در گره‌های شبکه به وسیله ماتریس دو بعدی $Q_d(s,a)=0$ در نظر می‌گیریم. در این ماتریس با ۵ سطر و ۷ ستون، هر سطر یک حالت مجزا و هر ستون نیز یک عمل مجزا را بازنمایی می‌کند. کدبندی انجام شده در مورد حالت‌ها و عمل‌های مختلف نیز به شرح جداول زیر است:

کدبندی حالت‌ها برای مثال مورد بررسی

کد حالت	شماره گره در شبکه
۱	۱
۲	۲
۳	۳
٤	٤
۵	۵
٦	٦(گره مقصد یا نهایی)

کدبندی عمل‌ها برای مثال مورد بررسی

کد عمل	نام سویه یا مسیر در شبکه
۱	۱-۲
۲	۱-۳
۳	۲-٤
٤	۳-٤
۵	۳-۵
٦	٤-٦
۷	۵-٦

همچنین با توجه به بخشهای قبلی و تعریف اجزای RL برای مسأله مورد بررسی، ماتریس $p_s(a)$ را برای توزیع احتمال بالتزمن جفت‌های در دسترس حالت-عمل و ماتریس تابع پاداش r را برای هر جفت حالت-عمل در دسترس می‌توان به شکل زیر ارایه نمود:

$$Q_d(s,a)=0=\begin{pmatrix}0&0&-&-&-&-&-\\-&-&0&-&-&-&-\\-&-&-&0&0&-&-\\-&-&-&-&-&0&-\\-&-&-&-&-&-&0\end{pmatrix}_{5*7}$$

$$r(s,a)=\begin{pmatrix}3&4&-&-&-&-&-\\-&-&0.5&-&-&-&-\\-&-&-&2&0.5&-&-\\-&-&-&-&-&3&-\\-&-&-&-&-&-&4\end{pmatrix}_{5*7}$$

$$p_s(a)=\begin{pmatrix}0.5&0.5&-&-&-&-&-\\-&-&1&-&-&-&-\\-&-&-&0.5&0.5&-&-\\-&-&-&-&-&1&-\\-&-&-&-&-&-&1\end{pmatrix}_{5*7}$$

۷. ارزیابی مدلهای یادگیری در دسته‌ای از شبکه‌های واقعی(Grid Network)

۷-۱ توصیف شبکه

یکی از کلاسهای شبکه‌ها، شبکه grid است. در این نوع از شبکه‌ها، گره‌ها در یک مستطیل مسطح توری شکل آرایش می‌یابند. بنابراین روش پیشنهادی را می‌توان برای یافتن مسیر بهینه به منظور هدایت وسایل نقلیه در شبکه با ساختار grid (مشابه شبکه‌های حمل‌ونقل واقعی) بکار گرفت. در این ساختار هر گره از طریق سویه‌های ورودی و خروجی در جهت‌های شمال، جنوب، شرق و غرب با سایر گره‌های همسایه خود در ارتباط است. شبکه ارایه شده در شکل ۸ را مشاهده کنید.

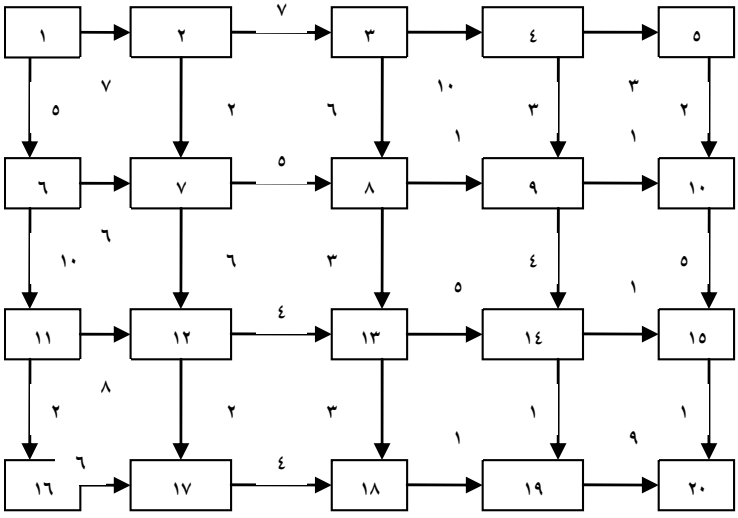
شبکه ارایه شده؛ یک شبکه حمل‌ونقل شهری است که دارای ١٦ مسیر افقی و ۱۵ مسیر عمودی جهت دار است که در مجموع ۲۰ تقاطع را به وجود می‌آورند. تقاطع‌ها از ۱ تا ۲۰ شماره گذاری شده و گره ۱ مبدا (O) و گره ۲۰ مقصد (D) شبکه مذکور در نظر گرفته شده است. زمانهای سفر روی سویه‌های شبکه شکل ۸ ارایه شده است. نکته قابل ذکر اینکه؛ مقادیر زمانهای سفر روی سویه‌ها

اعدادی بین ۱ و ۱۰ به طور تصادفی تولید شده است. بنابراین ارزش یا وزن سویه‌ها در زمانهای مختلف t به طور تصادفی تغییر می‌کند.

نهایتاً آنکه دیگر شبکه‌های واقعی را نیز می‌توان به صورت شبکه‌ای با ساختار grid تبدیل کرد. از جمله روشهای چگونگی استخراج grid از شبکه‌های شهری بکارگیری رویکردی جدید است که از همپوشانی تصاویر سه بعدی نواحی شهری استفاده می‌کند[۳۶].

۷-۲ شبیه‌سازی الگوریتم یادگیری

در بخشهای قبلی مقاله، قدمهای الگوریتم یادگیری ارایه شد. در این قسمت به منظور بررسی رفتار الگوریتم پیشنهادی، نتایج حاصل از اجرای مدل و انجام شبیه‌سازیها یعنی مقادیر نهایی در دسترس از ماتریس $Q_d(s,a)$ بر مبنای پارامترهای تنظیم شده α نرخ یادگیری، T درجه حرارت یا دما، γ ضریب تنزیل پاداش بستگی و حداکثر تعداد تکرار شبیه‌سازی(count) با اجتناب از ذکر جزئیات برنامه شبیه‌سازی[۲۲] برای شبکه grid مورد بررسی ارایه می‌شود. پارامترها یا مشخصه‌های شبیه‌سازی از طریق تحلیل حساسیت و به شکل دستی تنظیم می‌شوند، به این طریق که مقادیر مختلفی برای مشخصه‌ها تا رسیدن به مقدار بهینه مورد آزمون قرار می‌گیرد.



شکل ۸. شبکه حمل‌ونقل واقعی تبدیل شده به شبکه grid

$\gamma = 1, \alpha = 0.5, T = 1, count = 300$

زمان مورد نیاز برای انجام محاسبات ۰/٦۷۰۹۹۱ ثانیه

$Q_d(s,a)$	مقدار	$Q_d(s,a)$	مقدار
$Q(1,1)$	۲۲/۰۹٥۰	$Q(9,16)$	٦/٤۷٤۰
$Q(1,2)$	۲۲/۳۱۲۲	$Q(9,17)$	٦/۲٤۹۳
$Q(2,3)$	۱٥/٥٥۷۹	$Q(10,18)$	٥/۹٦۹۳
$Q(2,4)$	۱٤/۸٦۳۸	$Q(11,19)$	۱۳/٤٦۱۲
$Q(3,5)$	۱۰/۷۰۲٤	$Q(11,20)$	۱٤/۱۳۱۸
$Q(3,6)$	۱۱/۹۸٦۸	$Q(12,21)$	۱۰/۲٥۱۰
$Q(4,7)$	۳/۷۸۱۳	$Q(12,22)$	۱۲/۲۱۱۲
$Q(4,8)$	٥/٤٦۹۷	$Q(13,23)$	۷/۱٤۰۸
$Q(5,9)$	٥/٤٦۹۹	$Q(13,24)$	۹/۰۰۲۰
$Q(6,10)$	۱۸/۲۲۹٦	$Q(14,25)$	۲/۰۰۰۰
$Q(6,11)$	۱۹/۱۸٤۳	$Q(14,26)$	۳/۸۷٥۰
$Q(7,12)$	۱۲/٥۹٥۸	$Q(15,27)$	۱
$Q(7,13)$	۱۳/٥۸۳۲	$Q(16,28)$	۱۷/٦۱۹۸
$Q(8,14)$	۷/۳٤۹۱	$Q(17,29)$	۱۳/۷۹٤٤
$Q(8,15)$	۸/۱۳۹۷	$Q(18,30)$	۹/۹۸۲۲
		$Q(19,31)$	۸/۹۹۸۹

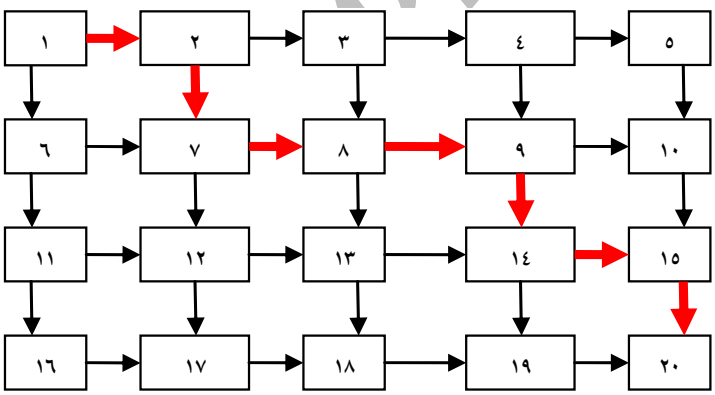
با بهره‌گیری از تفسیر خروجی و اینکه خروجی فوق، سیاست یا استراتژی مناسب را می‌تواند به ازای حالت‌های مختلف محیط بر مبنای بیشترین ارزش $Q_d(s,a)$ (در اینجا کمترین تخمین هزینه زمان) توصیه کند. مراحل طی کردن مسیر توسط یک وسیله نقلیه از گره شروع تا رسیدن به گره مقصد شبکه؛ یعنی مسیر ۲۰-۱۵-۱۴-۹-۸-۷-۲-۱ در شکل۹ با سویه‌های پررنگ متمایز شده است:

جدول۲. سیاست عامل ها در مدل یادگیری اجرا شده برروی شبکه grid

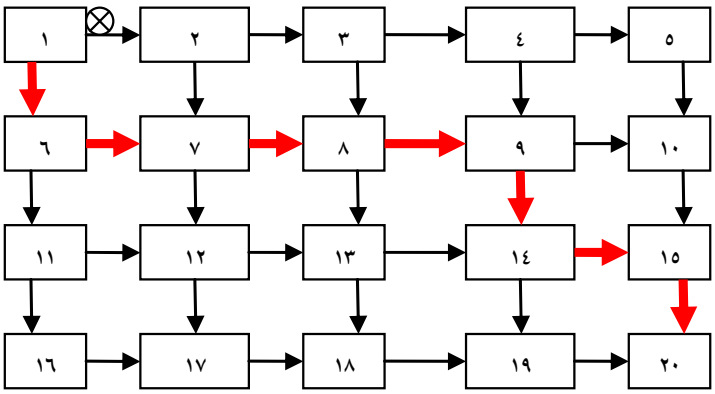
کد حالت	عمل توصیه شده	مقدار $Q_d(s,a)$
۱	حرکت در امتداد سویه ۲-۱	۲۲/۰۹۵
۲	حرکت در امتداد سویه ۷-۲	۱۴/۸۶۳۸
۷	حرکت در امتداد سویه ۸-۷	۱۲/۵۹۵۸
۸	حرکت در امتداد سویه ۹-۸	۷/۳۴۹۱
۹	حرکت در امتداد سویه ۱۴-۹	۶/۲۴۹۳
۱۴	حرکت در امتداد سویه ۱۵-۱۴	۲
۱۵	حرکت در امتداد سویه ۲۰-۱۵	۱

۷-۳اجرای مدل یادگیری در وضعیت پویا:تغییر زمانهای سفر

به منظور در نظر گرفتن وضعیت پویایی‌های محیط(یا پویایی شبکه مورد بررسی) در طی فرآیند یادگیری، تغییرات در برخی از ویژگیهای شبکه در گامهای زمانی معین طی شبیه‌سازی ناشی از تغییر (به طور تصادفی) در مقادیر زمانهای سفر را در الگوریتم یادگیری وارد می کنیم. توپولوژی شبکه نیز طی این تغییرات، ثابت است(یعنی همان شکل ۸). در ادامه به عنوان مثال فرض کنید در زمان عزیمت t_1 بر اثر حوادثی نظیر تصادف ترافیکی، زمان سفر برروی سویه زمان- وابسته۱-۲ به بینهایت میل کند. یعنی سناریوی تراکم ترافیک اتفاق بیافتد، یا به عبارتی دیگر حجم ترافیکی سویه مذکور بیش از ظرفیت ترافیکی آن شود. اکنون برای به دست آوردن جواب یا مسیر بهینه جدید از آخرین مقادیر به دست آمده برای $Q_d(s,a)$ از ماتریس مرحله قبل؛ برای شروع الگوریتم یادگیری با درنظر گرفتن تغییر بیان شده استفاده می‌کنیم. پس از اجرای شبیه‌سازی الگوریتم یادگیری، جواب یا مسیر بهینه جدید (با مسدود نمودن سویه ۱-۲) برروی شکل۱۰ قابل مشاهده است (یعنی مسیر۲۰-۱۵-۱۴-۹-۸-۷-۶-۱).



شکل۹ - عبور جریان ترافیکی از سویه های متمایز(پر رنگ) شبکه grid بر مبنای خروجی



شکل ۱۰. مسیر بهینه جدید برروی شبکه grid با لحاظ نمودن تراکم ترافیک

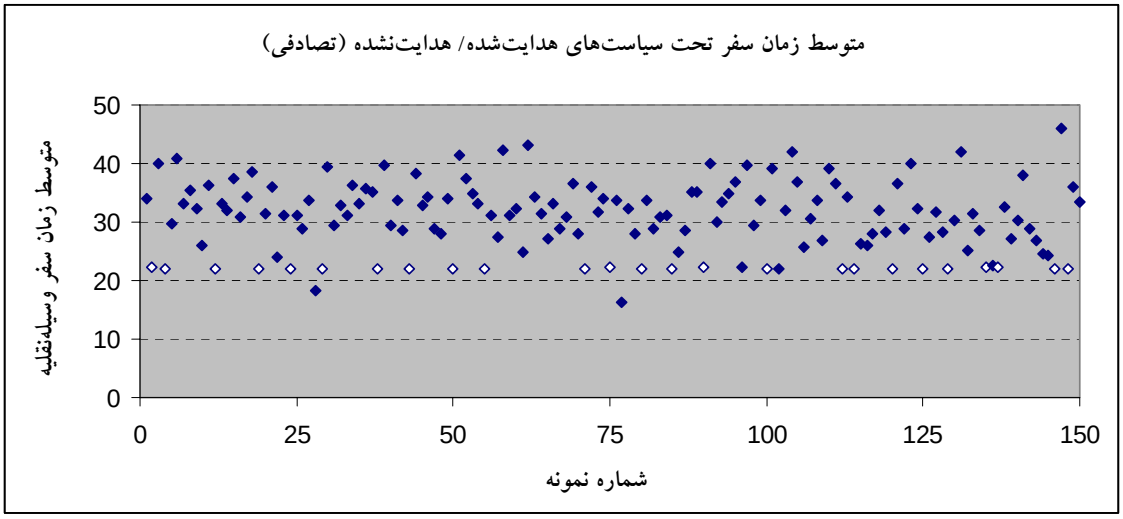
با استفاده از شیوه فوق می‌توان سایر تغییرات آتی برروی زمان سفر سویه‌های شبکه را تحلیل کرده و پاسخ‌های لازم را به دست آورد.

۷-٤ اعتبار سنجی یا ارزیابی مدلها

در این قسمت به منظور اعتبارسنجی مدلهای یادگیری ارایه شده در هدایت مسیر وسایل نقلیه؛ از مقایسه هزینه‌ای(در واقع هزینه زمان) بین سیاست پیشنهاد شده یعنی انتخاب عمل بر اساس استراتژی بالتزمن با دیگر سیاست‌ها (سیاست تصادفی) یعنی انتخاب عمل به شکل تصادفی استفاده می‌شود. به عبارت دیگر ارزیابی عملکرد[۲۳] وسایل نقلیه هدایت شده که از تکنیک هدایت شده جستجوی مسیر پیروی می‌کنند انجام می‌شود تا صرفه‌جویی‌های هزینه زمان سفر اثبات شود. برای انجام ارزیابی‌ها، وسایل نقلیه را به ۲ گروه هدایت شده و هدایت نشده با تعاریف ذیل؛ تقسیم می‌شوند. وسیله نقلیه هدایت شده، سفر خود را از تقاطع مبدا شبکه شروع کرده و از تقاطع مقصد از شبکه حمل‌ونقل خارج می‌شود. این دسته از وسایل نقلیه می‌توانند از طریق تجهیزاتی،اطلاعات مسیر را برای مبدا-مقصد مشخص در هر زمان دریافت کنند و وقتی که به تقاطعی می‌رسند می‌توانند در تطبیق با اطلاعات دریافتی در حین سفر، مسیر خود را بر اساس استراتژی بهینه تغییر دهند. در مقابل، وسیله نقلیه هدایت نشده تا قبل از اینکه وارد شبکه حمل‌ونقل شوند(سفر خود را آغاز کنند) اطلاعات مسیر را برای مبدا - مقصد مشخص دریافت کرده و در کل سفر نیز از آن پیروی می‌کنند و چنانکه بخواهند در تقاطعی تغییر مسیر دهند این کار بر اساس استراتژی تصادفی انجام می‌شود.

اکنون به منظور بررسی نتایج شبیه‌سازی، در حدود ۱۵۰ نمونه یا وسیله نقلیه برای شبکه grid تولید می‌کنیم و آزمایشی را به این صورت طراحی می‌کنیم که ۲۵ وسیله نقلیه (در حدود۱٦,۷ درصد از نمونه‌ها) از استراتژی بالتزمن استفاده کنند، یعنی هدایت شده باشند و مابقی(۱۲۵وسیله نقلیه دیگر) از سیاست تصادفی پیروی می‌کنند. مسیریابی هر وسیله نقلیه را بر حسب نوع آن برای شبکه شکل۸ به طور کامل انجام داده و ارزیابی عملکرد هر وسیله نقلیه را با معیار متوسط زمان (هزینه زمان) مسیریابی در شکل ۱۱ نمایش می‌دهیم.

نقاط مربعی توخالی، وسایل نقلیه ای است که از سیاست هدایت شده پیروی می‌کنند. سایر نقاط نیز نماینده وسایل نقلیه ای است که از سیاست تصادفی پیروی می‌کنند. همان طور که از روی شکل ۱۱مشخص است متوسط زمان سفر یا هزینه زمان سفر وسایل نقلیه هدایت شده در اطراف حدودا ۲۲ واحد زمان تغییر می‌کند که از زمان سفر اکثر سایر نقاط یعنی وسایل نقلیه‌ای که از سیاست تصادفی پیروی می‌نماید بسیار کمتر است. همچنین میانگین زمان سفر برای وسایل نقلیه هدایت نشده حدوداً ۳۲,۲۵ واحد زمان است[۲٤] نتیجتاً صرفه جویی زمان سفر(هزینه زمان سفر) وسایل نقلیه هدایت شده نسبت به وسایل نقلیه هدایت نشده تقریباً معادل درصد $\frac{32.25-22}{32.25}*100=31.8$ تخمین زده می‌شود. بنابراین سیاست پیشنهادی قابل قبول است. در ادامه صرفه جویی زمان سفر برای سایر درصدهای وسایل نقلیه هدایت شده از کل وسایل نقلیه به همین شیوه قابل تخمین زدن است.



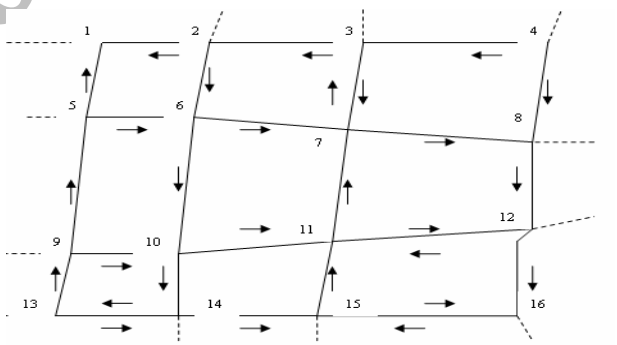
شکل ۱۱. اعتبارسنجی سیاست پیشنهادی در انتخاب مسیر وسایل نقلیه

۷-۵ کاربرد مدل‌های یادگیری در یک نمونه واقعی: شبکه خیابانهای تهران

در این قسمت به منظور کاربرد الگوریتم‌های یادگیری عاملها؛ از یک شبکه انتخاب شده از شهر تهران استفاده می‌شود و از طریق شبیه‌سازی، الگوریتم پیشنهادی برروی شبکه انتخاب شده اجرا خواهد شد.

۷-۵-۱ شبکه خیابانهای انتخاب شده از شهر تهران

به منظور شبیه‌سازی الگوریتم‌های یادگیری برروی یک شبکه خیابانهای شهری، بخشی از شبکه خیابانهای شهر تهران که در مرکز شهر واقع شده و بین ٤ خیابان اصلی محصور شده به عنوان یک بستر آزمایش[۲۵] انتخاب شده است. ناحیه مذکور از شمال به خیابان جمهوری، از غرب به خیابان ولیعصر، از جنوب به خیابان شوش و از شرق به خیابان ۱۷ شهریور محدود شده است. سپس با استفاده از نقشه ترافیکی زیر، فراهم شده توسط شرکت کنترل ترافیک شبکه ارایه شده دارای ٢٤ سویه و ١٦ تقاطع است که به منظور تهران، شبکه جهت دار شکل۱۲ توسط محققین استخراج شد.

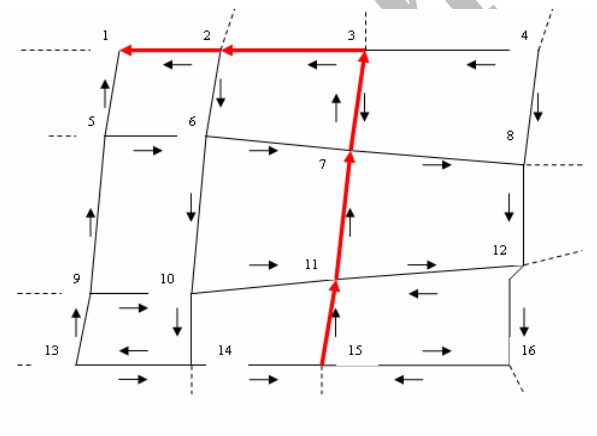


شکل ۱۲. یک شبکه خیابانی در شهر تهران

تشخیص تقاطع‌ها از یکدیگر، تقاطع‌ها از ۱ تا ١٦ شماره گذاری شده است. از دیگر ویژگیهای شبکه ارایه شده اینکه برخی از سویه‌ها دوطرفه هستند که از دو سویه در خلاف جهت هم به‌عنوان جایگزین آنها در محاسبات استفاده شده و الزاماً زمانهای سفر رفت یا برگشت برروی این قبیل سویه‌ها معادل نیست. به عبارت دیگر حرکت در امتداد هر کدام از سویه‌های در خلاف جهت هم را به عنوان یک عمل مجزا کدگذاری می‌کنیم. همچنین فرض می‌شود که وسایل نقلیه از طریق تقاطع‌های معرفی شده به شبکه وارد یا از آن خارج می‌شوند. زمانهای سفر روی سویه‌های شبکه شکل ۱۲ نیز در جدول۳ ارایه شده است. نکته قابل ذکر اینکه، مقادیر زمانهای سفر(در زمان عزیمت t) روی سویه‌ها با استفاده از داده‌های فراهم شده از سیستم هوشمند کنترل مرکزی شرکت کنترل ترافیک و روش بیان شده در ضمیمه A تولید می‌شود.

۷-۵-۲ نتایج شبیه‌سازی اجرای الگوریتم یادگیری عاملها

به منظور نشان دادن نتایج شبیه‌سازی، می‌توان سناریوهای مختلف ترافیکی را در نظرگرفت از جمله جفت مبدا- مقصد ۱۵ و ۱ که گره ۱۵ را بعنوان تقاطع شروع حرکت یک وسیله نقلیه هدایت شده و گره ۱ را به عنوان تقاطع خاتمه فرض می‌کنیم. در ادامه پس از تنظیم پارامترهای الگوریتم یادگیری ($\gamma = 1, \alpha = 0.5, T = 1, count = 100$) نتایج حاصل از اجرای مدل و انجام شبیه‌سازیها یعنی مسیر توصیه شده برای حرکت وسیله نقلیه (در زمان عزیمت t) در شکل ۱۳ ارایه شده است.



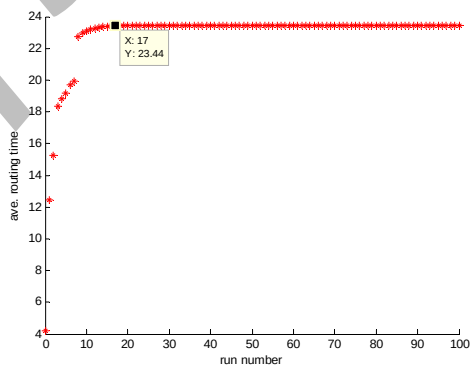
شکل ۱۳. مسیرتوصیه شده برای حرکت وسایل نقلیه برروی شبکه بخش ۷-۵ (در زمان عزیمت t برای جفت مبدا- مقصد۱۵و۱)

همچنین متوسط زمان مسیریابی در مقابل هر تکرار شبیه‌سازی با استفاده از نمودار ۱ ارایه شده است. همانطور که در این نمودار

جدول ۳. زمانهای سفر(در زمان t) روی سویه‌های جهت‌دار شبکه شکل۱۲

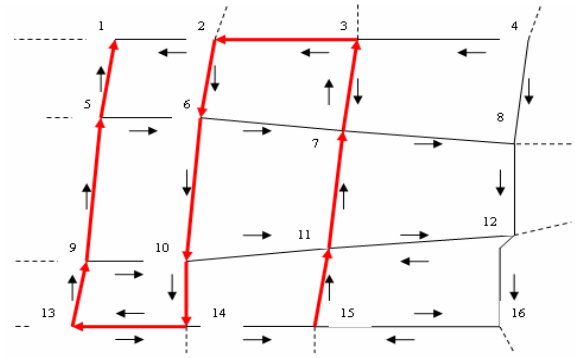
زمان سفربر روی سویه جهت دارp-q	گره خاتمه q	گره شروع p	زمان سفربر روی سویه جهت دارp-q	گره خاتمه q	گره شروع p	زمان سفربر روی سویه جهت دارp-q	گره خاتمه q	گره شروع p
٥	١١	١٢	٤/٦	٣	٧	٣/٥	١	٢
٤	١٦	١٢	٥	٨	٧	٣	٦	٢
٢	٩	١٣	٣	١٢	٨	٤	٢	٣
٣	١٤	١٣	٦/٤	٥	٩	٤/٦	٧	٣
٣	١٣	١٤	٣/٩	١٠	٩	٥/٢	٣	٤
١	١٥	١٤	٥	١١	١٠	٥	٨	٤
٥	١١	١٥	١	١٤	١٠	٢/٤	١	٥
٧	١٦	١٥	٣/٩	٧	١١	٢	٦	٥
٧	١٥	١٦	٦/٢	١٢	١١	٤	٧	٦
						٦/٢	١٠	٦

مشاهده می‌شود معیار عملکرد تعریف شده پس از تعداد مشخصی تکرار به مقدار ثابت حدودا ٢٣,٤٤ (واحد هزینه زمان) همگرا می‌گردد که نشان دهنده کامل شدن فرآیند یادگیری است.



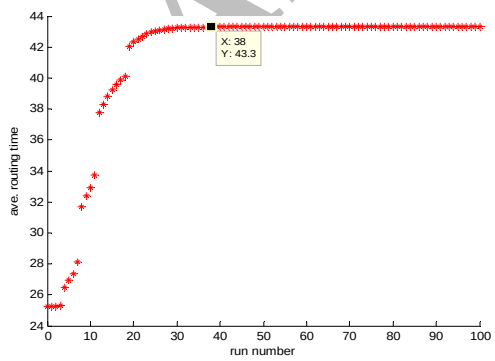
نمودار۱. روند تغییرات متوسط زمان مسیریابی در تکرارهای مختلف شبیه‌سازی(برای شبکه بخش ۷-۵)

پس از به دست آوردن مسیر بهینه برای هریک از جفت‌های مبدا-مقصد مورد نظر، می‌توان با کسب اطلاعات ترافیکی زمان واقعی در زمانهای عزیمت آتی، مسیرهای بهینه جدید را با اعمال تغییرات زمانهای سفر مشابه با توضیحات بیان شده در بخش۷-۳ به دست آورد. ضمناً مسیرهای توصیه شده به وسایل نقلیه(هدایت شده) به دلیل پیروی از استراتژی بالتزمن همواره دارای کمترین زمان سفر برای رسیدن به مقصد هستند. در ادامه فرض کنید زمان سفر سویه ۱-۲ به مقدار مثلاً ۱۰۰ واحد زمان تغییر یابد، نتایج حاصل از اجرای مدل و انجام شبیه‌سازیها یعنی مسیر توصیه شده برای حرکت وسیله نقلیه پس از این تغییر برای رسیدن به گره مقصد ۱ در شکل ١٤ ارایه شده است.



شکل ١٤. مسیرتوصیه شده برای حرکت وسایل‌نقلیه برروی شبکه بخش ۷-۵ (پس از تغییر زمان سفر سویه ۱-۲)

همچنین روند تغییرات متوسط زمان مسیریابی در مقابل هر تکرار شبیه‌سازی پس از اعمال تغییر یاد شده با استفاده از نمودار ۲ ارایه شده است. معیار عملکرد تعریف شده پس از تعداد مشخصی تکرار به مقدار ثابت حدودا ٤٣,٣ (واحد هزینه زمان) همگرا می‌شود که نشان دهنده کامل شدن مجدد فرآیند یادگیری است.



نمودار۲. روند تغییرات متوسط زمان مسیریابی در تکرارهای مختلف شبیه‌سازی(برای شبکه بخش ۷-۵ پس از تغییر زمان سفر سویه ۱-۲)

۸. جمع‌بندی و نتیجه‌گیری

در این مقاله چگونگی بکارگیری تکنیکهای عامل گرا با تأکید بر یادگیری تقویتیRL به عنوان یک راه حل در مواجهه با نامعینی‌های مسأله مسیریابی وسایل نقلیه در شبکه‌های ترافیکی ارایه شد. به این منظور با بیان یک چارچوب مفهومی‌هدایت مسیر مبتنی بر ساختار مسیریابی غیرمتمرکز و شناسایی اجزای مختلف RL در ارتباط با مسأله مورد مطالعه، محاسبات هدایت مسیر ترافیکی مبتنی بر یادگیری تعدادی از عاملهای هوشمند در وضعیت ایستای شبکه و تعمیم آن به وضعیت پویای محیط و همچنین در دسته‌ای از شبکه‌های واقعی (Grid Network) و یک نمونه واقعی (شهر تهران) انجام شده است.

از نتایج مهم تحقیق ارایه شده می‌توان به حل مسأله هدایت مسیر پویا با استفاده از روشهای شبیه‌سازی و توانایی مدلهای یادگیری ارایه شده به منظور تصمیم گیری مسیریابی در شبکه حمل‌ونقل شهری شامل به‌دست آوردن سیاست یا استراتژی انتخاب مسیر در گره‌های شبکه، در تطبیق با شرایط پویای ترافیکی و ارایه آلترناتیوهای مختلف پیشهادی طی مسیر برای رانندگان با هدف حداقل کردن معیار زمانهای سفر وسایل نقلیه اشاره کرد. همچنین در این مقاله با توجه به امکان مدل کردن زمانهای سفر روی سویه‌ها به عنوان یک فرآیند مارکوفی، به طور ضمنی از روشهای یادگیری به منظور حل مسایل کوتاه‌ترین مسیر پویا بدون در اختیار داشتن توابع انتقال حالت، استفاده شده است.

از جمله محدودیتهای تحقیق حاضر اینکه یادگیری عاملها با استفاده از مدلهای شبیه‌سازی می‌تواند تا حدودی شرایط ترافیکی واقعی را فراهم کند، اما کاربردی کردن مدلهای مذکور در یک سیستم واقعی نیازمند تجهیز شبکه ترافیکی به سخت افزارها و نرم افزارهای لازم و ارتباط میان مدل شبیه‌سازی و شبکه ترافیکی واقعی از طریق طراحی و تعریف پروتکل‌ها است.

سایر تحقیقات آتی نیز می‌توانند بر روی موضوعاتی مانند اجرای مدل در یک محیط حقیقی و بررسی تغییرات توپولوژی شبکه حمل‌ونقل مثلاً حذف یا مسدود شدن برخی از سویه‌ها یا مسیرها و تحلیل پاسخ‌های به‌دست آمده، در نظر گرفتن انتظار در گره‌ها یا تقاطع‌های شبکه در مدلهای یادگیری ارایه شده، تعیین تعداد بهینه عاملهای هوشمند مورد استفاده در چارچوب هدایت مسیر، متفاوت در نظر گرفتن پارامتر دما T برای گره‌های شبکه، بررسی مکانیزم اشتراک گذاری اطلاعات مسیریابی وسایل نقلیه به عنوان عاملهای هوشمند، مکانیزم پرداخت هزینه خدمات هدایت مسیر ارایه شده انجام شوند.

ضمیمه A :

محاسبه زمانهای سفر زمان واقعی یا جاری برای یک مسیر ممتد که به چندین آشکارگر یا حسگر حلقه مغناطیسی تجهیز شده باشد کاملا ساده است. مسیر مورد نظر را به N قسمت با طولهای Δ_i تقسیم می‌کنیم به طوری که هر قسمت شامل یک ایستگاه آشکارگر باشد. ایستگاه آشکارگر می‌تواند مستقیماً، سرعت متوسط وسیله نقلیه v_iرا با دقت کافی اندازه‌گیری کند. بنابراین زمان سفر جاری در قسمت i ام به وسیله رابطه زیر قابل محاسبه است:

$$\tau_i = \Delta_i / v_i$$

زمان سفر جاری روی مسیر ممتد به عنوان زمان سفر یک وسیله نقلیه که بر روی مسیر ممتد تحت این فرضیه که شرایط ترافیکی جاری در طی سفر تغییر نخواهد کرد تعریف می‌شود. بر پایه این تعریف و فرمول ارایه شده، زمان سفر جاری τ برای کل مسیر(شامل N قسمت) از رابطه زیر به دست می‌آید:

$$\tau = \sum_{i=1}^{N} \tau_i = \sum_{i=1}^{N} \Delta_i / v_i$$

ضمیمه B :

اثبات اینکه در مسأله مورد بررسی آیا استراتژی بالتزمن در انتخاب عمل، این هدف را که وسایل نقلیه در زودترین زمان ممکن به مقصد برسد تامین می‌نماید یا خیر، از طریق تحلیلی بصورت زیر امکان پذیر است:

با توجه به هدف مسأله که حداقل کردن تابع ارزش (تخمین مجموع زمان سفر از گره حالت s تا مقصدd) است تابع هدف را می‌توان به شکل $\min Q(s,a)$ ارایه کرد. در ادامه محدودیتهای زیر در نظر گرفته می‌شوند:

الف- محدودیت مربوط به گره مقصد به شکل $Q(d,a)=0$

ب- محدودیت مربوط به مجموع احتمالات انتخاب اعمال در هر گره حالت s به‌شکل $\sum_{a\in U(s)} p_s(a) = 1$

ج- محدودیت مربوط به پارامتر درجه حرارت T با توجه به تعریف ارایه شده از آنتروپی در هر گره حالت s به‌شکل $-\sum_{a\in U(s)} p_s(a)\log\ p(a_s) = E_s$ اکنون به منظور حل مدل بهینه‌سازی ارایه شده می‌توان با تشکیل تابع لاگرانژ و مشتق‌گیری از آن نسبت به $p_s(a)$ به جواب بهینه رسید که جواب همان رابطه ارایه شده (٦) خواهد بود.

۹. پانویس‌ها

1. Congestion
2. Dynamic Route Guidance System(DRGS)
3. Intelligent Transportation System(ITS)
4. Robust
5. Real-time
6. Dynamic Shortest Path
7. Fix
8. Conceptual framework
9. Information Service Provider
10. Agent-oriented
11. Reinforcement Learning (RL)
12. Policy
13. Reward function
14. Value function
15. Markov decision process(MDP)
16. Temporal Difference
17. Exploration
18. Exploitation
19. Greedy
20. Boltzmann

۲۱. با توجه به اینکه در محاسبات Q از ضریب تنزیل γ استفاده می شود بنابراین می‌توان Q را هزینه زمان*(time cost)* محسوب کرد.

۲۲. از آنجائی که مقادیر $r,Q, p_s(a)$ به شکل ماتریسی ارائه گردیده است بنابراین شبیه‌سازی مدل پیشنهادی در قالب ماتریسی امکان پذیر بوده و به دلیل توانایی های نرم افزار *MATLAB* در انجام عملیات ماتریسی،از آن برای برنامه نویسی الگوریتم یادگیری و اجرای مدل شبیه سازی استفاده شده است.

۲۳. به منظور ارزیابی عملکرد مدل می‌توان از معیار میانگین یا متوسط زمان مسیریابی وسایل‌نقلیه استفاده نمود. زمان مسیریابی به‌عنوان زمان لازم برای رسیدن به گره مقصد یا حالت نهایی با شروع حرکت از گره مبدا (حالت اولیه) تعریف می‌شود. با توجه به انکه در هر تکرار شبیه‌سازی برای ادامه حرکت از گره مبدا، ۲ مسیرپیش روست بنابراین متوسط زمان مسیر یابی را می‌توان توسط رابطه زیر تعریف کرد:

$$Average\ routing\ time = \frac{Q(1,1)+Q(1,2)}{2}$$

۲٤. این میانگین از حاصل تقسیم مجموع زمانهای سفر وسایل‌نقلیه هدایت نشده بر تعداد وسایل‌نقلیه هدایت نشده به‌دست می‌آید.

25. Test bed

۱۰. مراجع

1. Sadek, A. and Chowdhury, M.A. (2003) "Fundamentals of intelligent transportation systems planning", Boston, Artech House.

2. Adler, J.L. and Blue, V.J. (1998) "Toward the design of intelligent traveler information systems", Transportation Research Part C, vol. 6, no.3, pp.157-172.

3. Deflorio, F.P. (2003) "Evaluation of a reactive dynamic route guidance strategy", Transportation Research, Part C, vol.11, no.5, pp.375-388.

4. Liang, Z. [et al] (2007) "Application of genetic algorithm in dynamic route guidance system", Journal of Transportation Systems Engineering and Information Technology, vol.7, no.3, pp.45-48.

5.Vandebon, U. and Upadhyay, P.K. (1997) 'Simulation modeling of route guidance concept', Transportation Research Record 1573, pp.44-51.

6. Taniguchi, E. and Shimamoto, H. (2004) 'Intelligent transportation system based dynamic vehicle routing and scheduling with variable travel times", Transportation Research part C, vol.12, no.3-4, pp.235-250.

7. Zhang, Z. and Xu, J. (2005) "A dynamic route guidance arithmetic based on reinforcement learning", Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou, pp.3607-3611.

8. Chabini, I. (1998) "Discrete dynamic shortest path problems in transportation applications: Complexity and algorithm with optimal run time", Transportation Research Record, 1645, no.1150, pp.170-175.

9. Liang, Z. [et al] (2005) "Designing dynamic path guidance system based on electronic maps by using Q-learning", Proc. International Conference on Space Information Technology, vol. 5985.

10. Ahuja, R.K [et al] (2003) "Dynamic shortest paths minimizing travel times and costs", Networks, vol.41, no.4, pp.197-205.

11. Dean, B.C. (2004) "Shortest paths in FIFO time-dependent networks: theory and algorithms", Technical Report, MIT, Cambridge.

12. Kotnyek, B. (2003) "An annotated overview of dynamic network flows", Technical Report, <URL: http://www.inria.fr/rrt>.

13. Fu, L. (2001) "An adaptive routing algorithm for in-vehicle route guidance systems with real-time information ", Transportation Research, Part B, vol.35, no.8, pp.749-765.

14. Burmeister, B. [et al] (1997) "Application of multi-agent systems in traffic and transportation", In IEE Proceedings of Software Engineering, 97, pp.51-60.

15. Adler, J.L. and Blue, V.J. (2002) "A cooperative multi-agent transportation management and route guidance system", Transportation Research, Part C, vol.10, No.5-6, pp.433-454.

16. Chabrol, M. [et al] (2006) "Urban traffic systems modeling methodology", Int. J. Production Economics, vol.99, no.1-2, pp.156-176.

17. Cai, C. Q. and Yang, Z. S. (2007) "Study on urban traffic management based on multi-agent system", Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, Hong Kong, pp.25-29.

18. Sussman, J.M. (2000) "Introduction to transportation systems", Boston, Artech House.

19. Baskar, L.D. [et al](2006) "Decentralized traffic control and management with intelligent vehicles", Proceedings of the 9th TRAIL Congress, Netherlands.

20. Schmitt, E. J. and Jula, H.(2006) "Vehicle route guidance: classification and comparison', Proceedings of the IEEE Intelligent transportation systems conference, Toronto, Canada, pp.242-247.

21. Shoham, Y. [et al] (2007) "If multi-agent learning is the answer, what is the question?" Artificial Intelligence, vol. 171, no.7, pp.365-377.

22. Busoniu, L. [et al] (2005) "Learning and coordination in dynamic multiagent systems", Technical Report 05-019 of Delft University of Technology, <URL: http://www.dcsc.tudelft.nl >.

23. Lhotska [et al] (1998) "Problems of learning in multi-agent systems", Proceedings of the 3rd IEEE/IFIP International Conference on Intelligent Systems for Manufacturing".

24. Kaelbing, L. and Moore, A. (1996) "Reinforcement learning: A Survey", Journal of Artificial Intelligence Research, No.4, pp.237-285.

25. Sutton, R.S. and Barto, A.G. (1998) "Reinforcement learning-An Introduction", Cambridge, MIT Press.

26. Ling, K. and Shalaby, A.S. (2005) "A reinforcement learning approach to streetcar bunching control", Journal of Intelligent Transportation Systems, vol.9, No.2, pp.59-68.

27. Schweighofer, N. and Doya, K. (2003) "Meta-learning in reinforcement learning", Neural Networks, vol.16, No.1, pp.5-9.

28. Yen, G.G. and Hickey, T. W. (2004) "Reinforcement learning algorithms for robotic navigation in dynamic environments", ISA Transactions, vol.43, No. 2, pp.217-230.

29. Rummery, G.A. and Niranjan, M. (1994) "Online Q-learning using connectionist systems", Technical Report 166, Cambridge University.

30. Kolouriotis, D.E. and Xanthopoulos, A. (2008) "Reinforcement learning and evolutionary algorithms for non-stationary multi-armed bandit problems" "Applied Mathematics and Computation", vol.196, No.2, pp.913-922.

31. Abdulhai, B. and Kattan, L. (2003) "Reinforcement learning: Introduction to theory and potential for transport application", Canadian Journal of Civil Eng., Vol. 30, pp.981-991.

32. Stefan, P. [et al] (2001) "Reinforcement learning for solving shortest-path and dynamic scheduling problems", Proceedings of the 3rd International Workshop on Emergent Synthesis, IWES'01 Bled, Ljubljana, pp.83-88.

33. Mellouk, A. [et al] (2007) "Adaptive quality of service-based routing approaches: Development of neuro-dynamic state-dependent reinforcement learning algorithms", International Journal of Communication Systems, no.20, pp.1113-1130.

34. Valdivia Y.T. [et al] (2001) "*An adaptive network routing strategy with temporal differences*', Artificial Intelligence, no.12, pp.85-91.

35. Achbany, Y. [et al] (2008) "Tuning continual exploration in reinforcement learning: An optimality property of the Boltzmann strategy", Neurocomputing, Article In Press.

36. Price, K. (2000) "Urban street Grid description and verification", 5th IEEE workshop on applications of computer vision, pp. 148-154.