ارائه یک روش کنترل بهینه برای بهینه‌سازی مصرف انرژی در سیستم‌های سیگنالینگ راه‌آهن با استفاده از یادگیری تقویتی

مقاله علمی – پژوهشی

محمدعلی صندیدزاده*، دانشیار، دانشکده راه‌آهن، دانشگاه علم و صنعت ایران، تهران، ایران
مجید آذین‌فر، دانش‌آموخته کارشناسی ارشد، دانشکده راه‌آهن، دانشگاه علم و صنعت ایران، تهران، ایران
ناصر مزینی، دانشیار، دانشکده مهندسی کامپیوتر، دانشگاه علم و صنعت ایران، تهران، ایران
فرزاد سلیمانی، پژوهشگر پسا دکتری، دانشکده راه‌آهن، دانشگاه علم و صنعت ایران، تهران، ایران
*پست الکترونیکی نویسنده مسئول: sandidzadeh@iust.ac.ir
دریافت: ۱٤۰۳/۰۳/۰۹– پذیرش: ۱٤۰۳/۱۱/۰۱
صفحه ٤۲۲-٤۰۹

چکیده

امروزه بهینه‌سازی مصرف انرژی در سیستم‌های حمل و نقل عمومی یک موضوع بسیار مهم تلقی می‌شود. از آنجایی که بخش بزرگی از انرژی در سیستم‌های حمل و نقل توسط مترو مصرف می‌شود، رویکرد جدیدی برای کنترل بهینه قطار برای کاهش مصرف انرژی پیشنهاد شده است. مدل پیشنهادی مبتنی بر الگوریتم یادگیری تقویتی است. فرض بر این است که یک قطار بین دو ایستگاه در امتداد خطی با شیب، منحنی و محدودیت سرعت غیر ثابت حرکت می‌کند. علاوه بر این، قطار باید سفر خود را در یک بازه زمانی معین کامل کند. یادگیری تقویتی متغیرهای حالت و پاداش‌ها بر اساس اقدامات انتخاب شده است. در روش پیشنهادی، متغیرهای حالت قطار، سرعت و موقعیت قطار هستند و عمل، شتاب یا حرکت خلاص است. برخلاف تکنیک‌های قبلی، اکثر مراحل بهینه‌سازی در این روش به صورت آفلاین بوده و برای هر مسیری فقط یک بار اجرا می‌شود. پس از تشکیل ماتریس پاداش، می‌توانیم از این روش به صورت آنلاین استفاده کنیم و سپس مشخصات سرعت را در حداقل زمان تولید کنیم. شبیه‌سازی‌های روش پیشنهادی در متلب پیاده‌سازی شده و در نهایت با الگوریتم ژنتیک مقایسه شده است.

واژه‌های کلیدی: پروفیل سرعت قطار، کنترل بهینه، بهینه‌سازی مصرف انرژی، روش یادگیری تقویتی، سیستم حمل و نقل ریلی

۱- مقدمه

امروزه انرژی یکی از حیاتی‌ترین عوامل شکل‌گیری و توسعه جوامع صنعتی است. اهمیت آن به گونه‌ای است که دسترسی به انواع انرژی عامل رشد، قدرت سیاسی و صنعتی ملت‌هاست. هزینه‌های بالای تولید انرژی از منابع مختلف، صنعتی شدن ملت‌ها و گرسنگی روزافزون آنها برای دریافت انرژی بیشتر، بسیاری از کشورها را مجبور به اجرای استراتژی‌های بهینه‌سازی مصرف انرژی به منظور جلوگیری از تخلیه اضافی انرژی، کاهش هزینه‌های تولید انرژی و افزایش رفاه عمومی کرده است. از آنجایی که صنعت حمل و نقل ریلی دارای تقاضای انرژی بیش از حد است که بخش اصلی هزینه‌های آن را تشکیل می‌دهد، بهینه‌سازی مصرف انرژی جزء ضروری هر سیستم حمل و نقل ریلی است. طراحی و نصب یک سیستم سیگنالینگ جدید برای ارائه خدمات با کیفیت بالا نه تنها گران است بلکه باعث افزایش هزینه‌های نگهداری و سایر هزینه‌های

اضافی نیز می‌شود. برای غلبه بر این امر، به نظر می‌رسد مدیریت ترافیک پیشرفته بهترین انتخاب برای افزایش بهینه‌سازی و ظرفیت خطوط موجود با توجه به منابع محدود موجود باشد. به منظور بهینه‌سازی عملکرد بالاتر و تقاضای انرژی کمتر در سیستم‌های حمل‌ونقل ریلی، روش‌های بسیاری از جمله کنترل پیک زمان منبع تغذیه، کاهش مقاومت چرخ به ریل (غلتان)، استفاده از ترمزهای احیاکننده و لکوموتیوهایی با نیاز انرژی بهینه در گذشته ارائه شده است. با این حال، پیاده‌سازی تکنیک‌های ذکر شده بسیار گران بوده است. در سال‌های اخیر، سرمایه‌گذاری در بهینه‌سازی الگوریتم‌های پروفیل سرعت قطار به منظور کاهش مصرف انرژی سیستم‌های حمل‌ونقل عمومی افزایش چشمگیری داشته است. نتیجه کاهش قابل توجهی در هزینه‌ها با پیاده‌سازی الگوریتم‌های ذکر شده در سیستم‌های حمل و نقل عمومی عظیم است. همچنین جدول زمانی قطارهای مدیریت شده با برنامه‌ریزی سفرهای قطار هوشمند از محصولات پیاده‌سازی‌های مذکور است.

حرکت قطار تحت تأثیر عوامل متعددی از جمله هندسه مسیر، سیستم سیگنالینگ، ویژگی‌های موتور کششی، شبکه‌های قدرت و محدودیت‌های سرعت است. در نتیجه، استفاده از روش‌های بهینه‌سازی سنتی با در نظر گرفتن پروفیل‌های مختلف سرعت موجود و محاسبه زمان و انرژی آن‌ها، اگر غیرممکن نباشد، کاری دلهره‌آور است.

از آنجایی که ایجاد پروفایل سرعت قطار به معنای طراحی استراتژی مناسب برای حرکت قطار است، می‌توان از الگوریتم یادگیری تقویتی (*RL*) که یکی از قوی‌ترین روش‌ها برای یافتن استراتژی‌های حرکتی بهینه است استفاده کرد. در این مقاله، با استفاده از الگوریتم *RL*، روشی جدید برای ایجاد یک استراتژی مناسب برای کنترل بهینه قطار با در نظر گرفتن تمامی مشخصات خط از جمله شیب، منحنی و محدودیت سرعت، معرفی شده است.

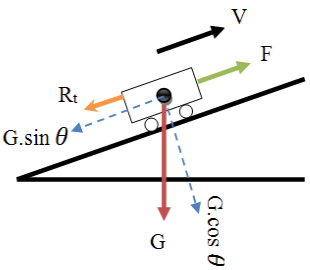
۲- پیشینه تحقیق

تعداد زیادی از تحقیقات برای کنترل بهینه قطار و صرفه‌جویی در انرژی انجام شده است. حرکت قطار بین دو ایستگاه در یک بازه زمانی معین به منظور صرفه‌جویی در انرژی مورد مطالعه قرار گرفته است (Milroy, 1980). در سال ۱۹۸۲، گروه تحقیقاتی برنامه ریزی و کنترل (SCG) یک مدل مصرف سوخت برای کنترل انرژی قطار ارائه کرد و پس از آن راه‌حل مناسبی ارائه شد (Howlett, 1996). در این تحقیق از اصل پونتریاگین و ضریب لاگرانژ برای بهینه‌سازی مصرف انرژی قطار در حالی که قطار بین دو ایستگاه در حال حرکت بود استفاده شد. نتایج تحقیقات وی برای خط بدون شیب و محدودیت سرعت عبارتند از: شتاب با حداکثر توان، کوستینگ و کاهش سرعت با حداکثر توان. لکوموتیوهای دیزلی-الکتریکی با این فرض که تنظیمات کنترل گسسته محدود بوده و نسبت مستقیمی با نرخ تحویل انرژی دارد، مدلسازی شدند. بعدها تحقیقات متعددی به نام «مدل‌های مصرف سوخت» بر اساس این مدل انجام شد. مسئله کنترل بهینه در یک خط با گرادیان ۰ مورد مطالعه قرار گرفت که طی آن این نتیجه حاصل شد که کنترل قطار بهینه برابر است با یافتن سری زمانی تغییرات حالت حرکت (Zhu et al., 2022)، (Howlett et al., 1997). در تمامی تحقیقات ذکر شده، شتاب یک متغیر کنترلی و دائماً محدود در نظر گرفته شد. این فرض در تضاد با عملکرد واقعی قطار بود. در نتیجه، نتایج نظری این تحقیقات چندان مورد توجه جامعه صنعتی قرار نگرفت. روش جدیدی برای شناسایی یک مدل کنترل فازی به منظور ایجاد یک چارچوب اقتصادی برای راه‌آهن‌های پرسرعت با استفاده از یک تعامل منطقی بین زمان سفر و مصرف انرژی ارائه شد (Hwang, 1998)، و بعدها، نظریه کنترل مستمر معرفی شد (Howlett, 2000)، (Howlett et al., 2001)، (Khmelnitsky, 2000). از روش‌های تحلیل عددی شامل برنامه‌ریزی پویا (*DP*) و برنامه‌ریزی غیرخطی (NLP) نیز برای حل مسئله استفاده شد (Yeo et al., 2002). الگوریتم ژنتیک (*GA*) برای جستجوی نقاط خلاصی مناسب آزمایش شده بود. با اختصاص ژن به هر نقطه شروع خلاصی، امکان وجود یک یا چند نقطه تغییر حالت (از شتاب به خلاصی و برعکس) در یک مسیر کوتاه بین دو ایستگاه بررسی شد. همچنین الگوریتم بهینه‌سازی ازدحام ذرات برای حل مسئله معرفی شد (Hu et al., 2010) و نشان داد که این الگوریتم راه‌حل بهتری نسبت به *GA* ارائه می‌دهد. سپس، از روش بهینه‌سازی سطح پاسخ دوگانه مبتنی بر *GA* برای کنترل نقطه خلاصی (Blanco et al., 2022)، (Liao et al., 2021)، (Xu et al., 2023)، (Ding et al., 2011) استفاده شد. مطالعات موردی نشان داد که این روش

برای قطارهای شهری مسافت طولانی در خطوط با ایستگاه‌های متعدد بسیار موثر بود. همچنین الگوریتم‌های بهینه‌سازی برای تولید پروفیل سرعت قطار با هدف به حداقل رساندن مصرف انرژی ارائه شد که این الگوریتم‌ها از GA استفاده می‌کردند و بر اساس روشی بودند که حرکت قطار را به ۳ فاز تقسیم می‌کرد: حرکت با حداکثر توان شتاب، حرکت در حین کوستینگ و حرکت با حداقل قدرت کاهش سرعت با استفاده از کنترل‌های چندگانه نقطه خلاصی با حداقل کردن مصرف انرژی (Kang, 2011). تابع هدف به گونه‌ای مشخص شده است که قطار فاصله بین دو ایستگاه را در زمان تعیین شده با حداقل انرژی طی کند (Yang et al., 2011)، (Lu et al., 2011).

۳- مدل‌سازی حرکت قطار

در این قسمت به رفتار قطار در حین حرکت بر روی شیب و محاسبه نیروهای شتاب و زمان سفر پرداخته شده است. شکل (۱) نیروهای وارد بر قطار را حین حرکت بر روی مسیری با گرادیان θ نشان می‌دهد (Sandidzadeh et al., 2020). که در آن F نیروی کشش، R_t مقاومت قطار در سرعت V و G وزن قطار است (Esveld, 2001).



شکل ۱. نیروهای وارد بر قطار در حال حرکت

۳-۱- نیروی کشنده

نیروی کشنده به موارد مختلفی از جمله وزن قطار بستگی دارد که تعیین کننده نیروی چسبندگی چرخ به ریل برای حرکت با سرعت مطمئن قبل از وقوع لغزش چرخ است. یکی دیگر از عوامل مهم نیروی کشنده سرعت حرکت است. هنگامی که سرعت افزایش می‌یابد، جریان موتور کششی کاهش می‌یابد. در نتیجه، نیروی کشنده نیز کاهش می‌یابد. کارخانه‌های سازنده لوکوموتیو نیروی کشنده را به عنوان تابع سرعت نشان می‌دهند که قدرت لکوموتیوها را نشان می‌دهد. معادله (۱) یک معادله معروف است که نیروی کشنده قطار را به عنوان تابعی از سرعت آن محاسبه می‌کند (Manajem, 2007).

$$F_t = \frac{270N}{V} n_1 n \qquad (۱)$$

که در آن F_t *(kg.f)* نیروی کشنده، N (اسب بخار) قدرت لوکوموتیو، V *(km/h)* سرعت قطار، η_1 تعداد لکوموتیوها و η ضریب کارایی لکوموتیو (ساییدگی لکوموتیو) است. از رابطه (۱) مشخص است که افزایش سرعت باعث کاهش نیروی کشنده می‌شود.

۳-۲- نیروهای مقاوم

هر جسمی که در بالای جسم دیگر حرکت می‌کند دارای یک سطح تماس است که در آن حرکت باعث سایش و اصطکاک می‌شود. حرکت قطار در یک مسیر همان نیروهای مقاوم را ایجاد می‌کند. این نیروها حرکت قطار را کند می‌کنند. همه نیروهای مقاوم به طور همزمان در قطار در حال حرکت رخ نمی‌دهند. با این وجود، موتور کششی قطار باید آنقدر قدرتمند باشد که بتواند بر تمام نیروهای مقاوم در برابر حادثه غلبه کند.

۳-۳- نیروهای مقاوم قطار

نیروهای مقاوم قطار، نیروهایی هستند که همیشه در حین حرکت قطار به وجود می‌آیند که شامل مقاومت در برابر اصطکاک در محور چرخ (مقاومت ژورنال)، مقاومت هوا و فلنج چرخ هستند.

مقاومت قطار به ساختار بوژی، واگن‌ها و هندسه لوکوموتیو، سایش سر و فلنج ریل، تعداد محورها، سرعت قطار و ویژگی‌های مسیر بستگی دارد. نیروهای مقاوم قطار توسط تجربه به عنوان نیروهای مورد نیاز برای حرکت قطار با سرعت ثابت محاسبه می‌شود. بیشتر معادلات ارائه شده در این زمینه چندجمله‌ای درجه دو هستند که به صورت تابع سرعت بیان می‌شوند و معادلات معروف آن عبارتند از معادلات دیویس، اشمیت و توتیل.

نیروی مقاوم دیگری که روی سرعت قطار تاثیر می‌گذارد، به وزن قطار بستگی دارد و از طریق (۲) محاسبه می‌شود.

$$R_g = G\sin\theta \cong G\theta \qquad (۲)$$

که در آن R_g نیروی مقاومت درجه کل است، G تناژ ناخالص قطار و θ (رادیان) زاویه درجه است. در مقاطع سربالایی مقاومت درجه به مقاومت قطار اضافه می‌شود و در مقاطع سرازیری از مقاومت قطار کم می‌شود (Manajem et al., 2009).

مقاومت در کاهش سرعت به دو عامل چسبندگی بین ریل و چرخ‌های کاهش سرعت تا توقف و جزء نرمال (عمود) نیروی واکنش ناشی از وزن چرخ روی ریل بستگی دارد. نیروی کاهش سرعت ثابت و مستقل از سرعت قطار است. قدرت کاهش سرعت قطار بر اساس تناژ قطار و اقدامات ایمنی به صورت (۳) انتخاب می‌شود.

$$R_b = \beta \times G \qquad (۳)$$

که در آن R_b *(kg.f)* نیروی کاهش سرعت کل است، β *(kg.f/ton)* قدرت کاهش سرعت قطار در واحد وزن قطار و G تناژ ناخالص قطار است.

۳-۳-۱- آزمایشات اشمیت و توتیل

این روش بر اساس آزمایش‌هایی است که در سال ۱۹۴۰ در دانشگاه ایلینویز انجام شد. در این آزمایش‌ها، حرکت قطار برای هر قطار تحت شرایط مختلف مسیر، نیروی کششی و بوژی اندازه‌گیری شد. وزن و سرعت قطار در هر مرحله به ترتیب ۵ تن و ۱۶ کیلومتر در ساعت افزایش یافت (Rochard et al., 2000).

۳-۳-۲- مدل دیویس

شکل اصلاح شده معادله دیویس به صورت (٤) است.

$$R_t(v) = q_0 + q_1 v + q_2 v^2 \qquad (٤)$$

که در آن R_t *(kg.f/ton)* نیروی دیویس است، v *(km/h)* سرعت و q_0، q_1 و q_2 ضرایب آیرودینامیکی هستند. از (۲) مشخص است که با افزایش سرعت، مقاومت قطار افزایش می‌یابد. با این وجود، معادله مناسب باید بر اساس وضعیت مسیر و میزان سایش بوژی‌ها تعیین شود. در شبیه‌سازی‌های این مقاله از روش دیویس استفاده شده است.

۳-٤- دینامیک حرکت قطار

نیروهای وارد بر قطار در حال حرکت در شکل (۱) نشان داده شده است. معادله حرکت را با استفاده از قانون دوم حرکت نیوتن می‌توان به صورت (۵) نوشت.

$$F_a = F_t - (R_t + R_g + R_b) = m \times a \qquad (۵)$$

که در آن F_a نیروی شتاب است، R_t، R_g و R_b به ترتیب نیروهای مقاوم قطار، درجه و کاهش سرعت هستند، F_t نیروی کششی، m جرم قطار و a شتاب قطار است.

یک فاکتور مهم برای قطار، رسیدن به سرعت مطلوب است. وقتی قطار در شیب شروع به حرکت می‌کند سرعت آن افزایش می‌یابد. افزایش سرعت باعث کاهش نیروی کششی و افزایش مقاومت می‌شود. پس از مدتی نیروهای کششی و مقاومتی برابر می‌شوند و متعاقباً نیروی شتاب (F_a) صفر می‌شود. پس از آن، قطار با سرعت ثابتی در شیب حرکت می‌کند. با ایجاد نیروی مقاومت برابر با نیروی کششی بر اساس (٦) می‌توان مقدار سرعت ثابت را بدست آورد.

$$\begin{aligned} &F_t = R_g + R_t \\ &\frac{270N}{v} n_L \eta = m(q_0 + q_1 v + q_2 v^2 + g) \end{aligned} \qquad (٦)$$

برای تعیین فاصله و زمان کاهش سرعت، نیروی کششی از معادله حرکت حذف می‌شود در حالی که نیروی مقاوم سرعت به آن اضافه می‌شود. از آنجایی که شتاب (a) برابر با نرخ تغییرات سرعت است، (۷) حرکت قطار را در مراحل کاهش سرعت نشان می‌دهد.

$$m\frac{dv}{dt} = -R_b - R_t(v) - mg\theta \qquad (۷)$$

اگر سرعت اولیه مشخص باشد، سرعت تابعی از زمان محاسبه می‌شود. از آنجایی که نیروی کاهش سرعت ثابت ($mg\beta$) و مستقل از سرعت است، معادله حرکت را می‌توان به صورت (۸) ارائه کرد.

$$\int_V^0 \frac{mdv}{mg(\beta+\theta) + R_t(v)} = -\int_0^T dt\,. \qquad (۸)$$

با استفاده از معادله دیویس، زمان کاهش سرعت با سرعت اولیه v را می توان با حل (۹) به دست آورد.

$$T(v) = \int_0^v \frac{du}{au^2 + bu + c} \qquad (۹)$$

که در آن $c = q_0/m + g(\beta+\theta)$ و $a = q_2/m$ هستند.

$$a=\frac{dv}{dt}=\frac{dv}{dx}\frac{dx}{dt}=v\frac{dv}{dx} \quad (۱۰)$$

با قرار دادن (۱۰) در معادله حرکت، (۱۱) به دست می‌آید.

$$mv\frac{dv}{dx}=-R_b-R_t(v)-mg\theta \quad (۱۱)$$

که رابطه بین سرعت و مسافت طی شده را بیان می‌کند. با جداسازی متغیرها، (۱۱) به (۱۲) تبدیل می‌شود.

$$\int_{V}^{0}\frac{mvdv}{mg(\beta+\theta)+R_t(v)}=-\int_{0}^{x}dx \quad (۱۲)$$

بنابراین، فاصله کاهش سرعت مورد نیاز x زمانی که سرعت اولیه v است، با حل (۱۳) به دست می‌آید.

$$x(v)=\int_{0}^{v}\frac{udu}{au^2+bu+c} \quad (۱۳)$$

$$a=q_2/m,\ b=q_1/m,$$
$$c=q_0/m+g(\beta+\theta)$$

٤- روش کنترل بهینه پیشنهادی

مشکلات بهینه‌سازی در دنیای امروز جایگاه قابل توجهی دارد. این مشکلات در زندگی روزمره رخ می‌دهد و حل آنها به غلبه بر عوارض متعدد اقتصادی و اجتماعی کمک می‌کند. با این حال، این مشکلات را نمی‌توان در اشکال کاربردی و عملی با استفاده از روش‌های بهینه‌سازی دقیق حل کرد و معمولاً نمی‌توان به راه‌حل‌های بهینه در محدودیت‌های زمانی قابل قبول دست یافت. در نتیجه، پژوهش‌ها به دنبال الگوریتم‌هایی می‌گردند که روش‌های مناسبی را برای یافتن پاسخ‌های باکیفیت و ارائه پاسخ‌های آن‌ها در محدوده‌های زمانی پذیرفته شده در خود جای دهند. برای مسائلی که فضاهای جستجوی بزرگی را در خود جای داده و باید در زمان واقعی حل شوند، روش‌های هوش مصنوعی اکتشافی می‌توانند با استفاده از ترکیبی از الگوریتم‌های آنلاین و آفلاین به دستیابی به پاسخ‌های بهتر در زمان‌های کوتاه‌تر کمک کنند. در این قسمت از این مقاله ابتدا الگوریتم RL معرفی شده و سپس کاربردهای آن در کنترل بهینه قطار و ایجاد پروفیل سرعت قابل قبول توضیح داده شده است.

٤-١- روش یادگیری تقویتی

هدف اصلی یادگیری یافتن روشی برای عملکرد حالت‌های مختلف است که در مقایسه با روش‌های دیگر که معیارهای خاصی در ذهن دارند، بهترین پاسخ را دارد. یادگیری زمانی اتفاق می‌افتد که عامل یادگیری بر اساس تجربیاتی که به دست آورده است به روشی متفاوت و احتمالاً بهتر از قبل عمل کند. مسئله یادگیری که فقط بر اساس داده‌های پاسخ محیط است، یک مشکل یادگیری تقویتی است. به طور کلی، یادگیری تقویتی به این معناست که استراتژی بر اساس شناخت محیط، تعامل با فضای اطراف و همچنین بر اساس پاداش و زیان انتخاب شود. اقدام بر اساس استراتژی انتخاب شده در بلندمدت، به پاسخ مطلوب منتهی می‌شود (Gosavi, 2009).

٤-١-١- عامل و محیط

تصمیم گیرنده یا یادگیرنده «عامل» نامیده می‌شود. هر چیزی که با عامل تعامل داشته باشد (در واقع هر آنچه در فضای خارج از عامل وجود دارد)، محیط نامیده می‌شود. فعل و انفعالات عامل-محیط به طور مداوم رخ می‌دهد. عامل تصمیم می‌گیرد و بر اساس آن عمل می‌کند و محیط بر اساس تصمیم خود به عامل پاداش می‌دهد و سپس به حالت جدیدی منتقل می‌شود.

٤-١-٢- استراتژی عامل

خط مشی یا استراتژی عامل، π، یک تابع احتمال است که شانس انتخاب هر عمل را در هر حالت با توجه به مرحله زمانی تعیین می‌کند. خط مشی مهم‌ترین پاسخی است که در RL یا هر مشکل یادگیری دیگری یافت می‌شود. برای مثال $\pi_t(s,\ a)=p$ یعنی اگر در t عامل در s باشد، عمل a را با احتمال p انتخاب می‌کند (Sutton et al., 1998).

٤-١-٣- اهداف و پاداش‌ها

در یادگیری تقویتی، هدف عامل در قالب یک سیگنال پاداش دریافتی از محیط بیان می‌شود. در هر مرحله زمانی، پاداش به صورت یک عدد ساده بیان می‌شود. به عبارت ساده‌تر، هدف عامل به حداکثر رساندن تعداد کل پاداش‌ها است. مهم است که در نظر داشته باشید که هدف، به حداکثر رساندن پاداش‌های کسب شده در کل زمان است و نه در هر مرحله. همچنین توجه به این نکته ضروری است که پاداش به نماینده به گونه‌ای داده می‌شود که عامل با به حداکثر رساندن پاداش، اهداف تعریف شده را برآورده کند. عامل نباید آموزش ببیند که

چگونه اهداف را برآورده کند. در واقع، سیگنال پاداش کانال ارتباطی با عامل است که به آن اطلاع می‌دهد که چه پاداشی باید دریافت کند.

٤-١-٤- فرایند تصمیم‌گیری مارکوف

یک مسئله RL که دارای ویژگی‌های مارکوف است، فرآیند تصمیم‌گیری مارکوف (MDP) نامیده می‌شود. اگر فضای حالت و اقدامات ممکن محدود باشد، فرآیند تصمیم‌گیری یک MDP محدود است. یک MDP با مجموعه‌ای از حالت‌ها، اقدامات ممکن و احتمال انتقال تعیین می‌شود. (١٤) احتمال انتقال را نشان می‌دهد که احتمال رفتن به حالت $s_{t+1}=s'$ است اگر در زمان t و حالت $s_t=s$ عمل $a_t=a$ امتحان شود.

$$P_{ss'}^{a}=\Pr\{s_{t+1}=s'\mid s_t=s,a_t=a\} \qquad (١٤)$$

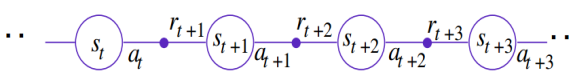
به همین ترتیب، اگر عمل و حالت فعلی و همچنین وضعیت بعدی مشخص باشد، مقدار مورد انتظار پاداش در مرحله بعد از طریق (١٥) به دست می‌آید.

$$R_{ss'}^{a}=E\{r_{t+1}\mid s'=s,a_t=a,s_{t+1}=s'\} \qquad (١٥)$$

معادلات (١٤) و (١٥) تمام ویژگی‌های مهم پویایی محیط در MDP را تعیین می‌کنند.

٤-١-٥- توابع مقدار حالت

تقریباً همه الگوریتم‌های RL مبتنی بر تقریب توابع مقدار حالت هستند. مقدار state می‌تواند به عنوان عاملی برای تعیین مناسب بودن حالت استفاده شود. می‌توان گفت که مقدار حالت، مقدار مورد انتظار پاداش کلی است که عامل با شروع از حالت $s_t=s$ و پیروی از خط مشی π کسب می‌کند. مجموعه‌ای از حالت‌ها، اقدامات و پاداش‌های متوالی را مانند شکل (٢) در نظر بگیرید.



شکل ٢. وضعیت‌ها، اقدامات و پاداش‌های متوالی

مقداری که باید برای حالت S در نظر گرفت به صورت (١٦) است.

$$V(s)\leftarrow E\{r_1+\gamma r^2+\gamma r^3+\cdots\mid s_0=s\} \qquad (١٦)$$
$$V(s)=E\{z_1\mid s_0=s\}$$

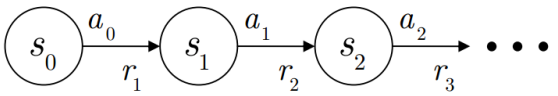
اگر عامل از خط مشی π پیروی کند، مقدار S به صورت (١٧) است.

$$V^{\pi}(s)=E^{\pi}\{z_1\mid s_0=s\} \qquad (١٧)$$

محاسبه مقدار واقعی همه حالت‌ها هنگام پیروی از خط مشی π، ارزیابی سیاست نامیده می‌شود و برای یادگیری صحیح ضروری است.

٤-١-٦- تابع ارزش عمل-حالت

تابع دیگری که در این مقاله تعریف شده است $Q_\pi(s,a)$ است که پاداش مورد انتظاری را که عامل دریافت می‌کند با شروع از حالت $s_t=s$، انجام عمل $a_t=a$ و پیروی از خط مشی π تعریف می‌کند. مجموعه‌ای از حالت‌ها، اقدامات و پاداش‌های متوالی را همانطور که در شکل (٣) نشان داده شده است در نظر بگیرید.



شکل ٣. وضعیت‌ها، اقدامات و پاداش‌های متوالی

مقداری که برای حالت عمل (s,a) باید در نظر گرفته شود برابر با (١٨) است.

$$Q(s,a)\leftarrow E\{r_1+\gamma r_2+\gamma^2 r_3+\cdots\mid s_0=s,a_0=a\}$$
$$=E\{z_1\mid s_0=s,a_0=a\} \qquad (١٨)$$

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

که در آن Q_π تابع مقدار عمل برای خط مشی π است. توابع مقدار حالت و مقدار عمل را می‌توان به صورت تجربی به دست آورد.

٤-١-٧- یادگیری Q

یادگیری Q یک نوع بدون مدل از یادگیری تقویتی بر اساس برنامه‌ریزی پویا تصادفی (DP) است. در یادگیری Q به جای

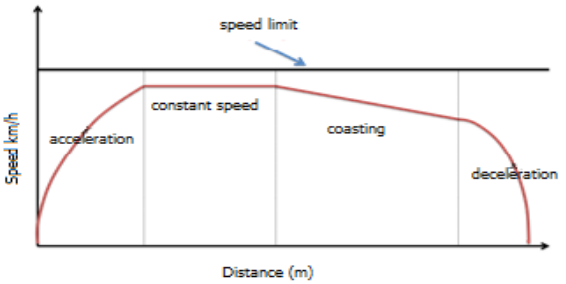
نگاشت از حالت‌ها به مقادیر حالت‌ها، نگاشت از جفت حالت–عمل به مقادیر Q انجام می‌شود. $Q\ (s,\ a)$ به هر جفت کنش–حالت اختصاص داده می‌شود. این شامل کل جوایز دریافتی هنگام شروع از s، انجام عمل a و پیروی از خط مشی موجود است. طبق نمودار جریان الگوریتم یادگیری Q، برای یادگیری یک تابع Q، می‌توان از جدولی در هر ورودی استفاده کرد که جفت $(s,\ a)$ به اضافه تقریبی است که یادگیرنده از مقادیر واقعی Q به دست آورده است. جدول با مقادیر تصادفی اولیه (معمولا صفر) پر شده است. عامل به صورت دوره‌ای وضعیت فعلی s را تشخیص می‌دهد و عمل a را انجام می‌دهد. سپس پاداش $r\ (s,\ a)$ را دریافت و $s'=\delta\ (s,\ a)$ ناشی از انجام عمل a را مشاهده می‌کند. مقادیر جدول بر اساس (۱۹) تغییر می‌کند.

$$Q(state, action) = R(state, action) + \gamma.\{\max[Q(next\ state, all\ actions)]\} \quad (۱۹)$$

از آنجایی که یک حالت در محیط به عنوان یک حالت جاذب در نظر گرفته می‌شود، که در این حالت وقتی عامل در آن حالت است حرکت آن متوقف می‌شود، یادگیری به صورت اپیزودیک انجام می‌شود. در یک قسمت عامل در یک محیط تصادفی قرار می‌گیرد و تا زمانی که به حالت جذبی برسد به تغییر مقدار Q ادامه می‌دهد. با تکرار این اپیزودها، مقادیر غیر صفر در کل جدول پخش شده و در نهایت به مقادیر بهینه همگرا می‌شوند.

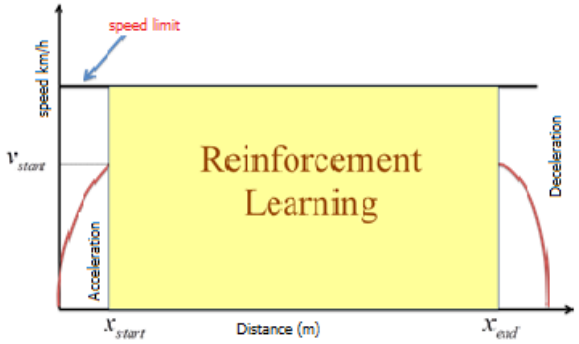
٤-۲- الگوریتم روش پیشنهادی

به منظور پیاده‌سازی الگوریتم پیشنهادی برای تولید پروفیل سرعت قطار و بهینه‌سازی مصرف انرژی، یک خط اختیاری با مشخصات سرعت ساده همانطور که در شکل (٤) نشان داده شده است، در نظر گرفته شده است.



شکل ٤. پروفیل سرعت در طول مسیر انتخابی

همانطور که در شکل (۵) نشان داده شده است، قبل از اعمال الگوریتم RL، قطار با حداکثر شتاب به محل x_{start} حرکت می‌کند. در نقطه x_{start} قطار دارای سرعت v_{start} است. بلافاصله پس از رسیدن به مکان x_{start}، الگوریتم RL شروع می‌شود و تا زمانی که قطار در نقطه X_{end} قرار می‌گیرد ادامه می‌یابد. سپس، بسته به مشخصات خط، می‌توان از کاهش سرعت یا حرکات خلاصی استفاده کرد و پس از رسیدن به نقطه کاهش سرعت، ترمزها را اعمال کرد. x_{start} نقطه‌ای از خط است که بر اساس شیب خط و فاصله بین دو ایستگاه قطار تعیین می‌شود.



شکل ۵. اعمال الگوریتم RL به مسیر انتخابی

به عبارت دیگر، قطار باید تا این نقطه با شتاب کافی حرکت کند تا حتی اگر قطار پس از عبور از این نقطه در حالت خلاص به حرکت خود ادامه دهد، قبل از رسیدن به ایستگاه پایانی توقف کامل نداشته باشد. X_{end} نقطه‌ای از خط است که در آن قطاری که با شتاب از ایستگاه شروع به ایستگاه بعدی حرکت می‌کند، می‌تواند با شروع به ترمز کردن در مکان مورد نظر در آن توقف کند. برای رعایت مقررات ایمنی، این نقطه باید دقیقا محاسبه شود. در غیر این صورت، قطار در ایستگاه هدف متوقف نخواهد شد. دلیل این امر این است که حداکثر نیروی ترمز برای هر قطار مشخص است و با وزن و سرعت اولیه قطار نسبت مستقیم دارد. بنابراین، اگر قطار با سرعتی بالاتر از سرعت اولیه خود شروع به ترمز کند، نمی‌تواند در ایستگاه مورد نظر متوقف شود. t_{start} زمان صرف شده برای رسیدن به نقطه x_{start} است و v_{start} سرعت قطار در زمان t_{start} است. پس از محاسبه x_{start}، v_{start}، t_{start} و x_{end} محدودیت‌های سرعت و زمان عملیات بهینه T برای پیمودن فاصله بین دو ایستگاه به طرح اعمال می‌شود. سپس مراحل زیر برای تولید پروفیل سرعت بهینه انجام می‌شود.

٤-٢-١- مشخصات قطار و مسیر

پارامترهای در نظر گرفته شده در شبیه‌سازی‌ها و برای قطار و مسیر در این مقاله در جدول (۱) نشان داده شده است (Nolte, 2003)، (Sheu et al., 2012).

جدول ۱. مشخصات قطار و مسیر

Train Length	L = 140m (Tehran Metro)
Train Mass	180-capacity (7 wagons 50 tons) Total Mass = 350 (tons)
Max. Acceleration	1 (m/s^2)
Max. Deceleration	1 (m/s^2)
Max. Speed	80 (km/h)
Tractive Effort	$F_t = \frac{270N}{V} n_1 \eta = \frac{3358\,kw}{V}$
Braking Force	$R_b = r_b G = 310700\ (N)$
Davis Force	$r_t = r_0 + r_1 v + r_2 v^2$ $r_0 = 0.03, r_1 = 3\times10^{-4}, r_2 = 6\times10^{-5}$
Resist. Force	$R_g = G\sin\theta \cong mg\theta, m = 350, g = 9.8$

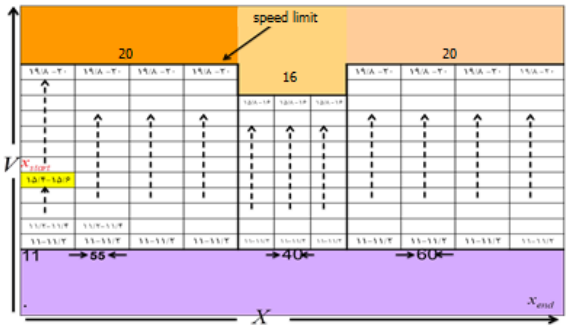
٤-٢-٢- تعیین مقادیر متغیرهای حالت

محور افقی شکل (۵) مکان قطار و محور عمودی سرعت آن را نشان می‌دهد. برای توصیف وضعیت قطار باید موقعیت و سرعت قطار در زمان t را تعیین کنیم. بنابراین، متغیرهای حالت قطار سرعت و مکان آنی قطار هستند. متغیرهای حالت به این شکل تعیین می‌شوند که خط بر اساس محدودیت‌های سرعت به چند بخش تقسیم می‌شود.

٤-٢-٣- متغیرهای حالت با پاداش‌های منفی

محدوده‌های ممنوعه، محدوده‌های داخل محدودیت سرعت هستند و رفتن به آن‌ها به این معنی است که قطار از سرعت مجاز تجاوز کرده است. از آنجایی که برای هر خط یک محدودیت سرعت بالا و پایین وجود دارد، چهار بخش محدودیت سرعت وجود دارد. برای هر کدام بخش‌های محدودیت سرعت یک حالت در نظر گرفته شده است.

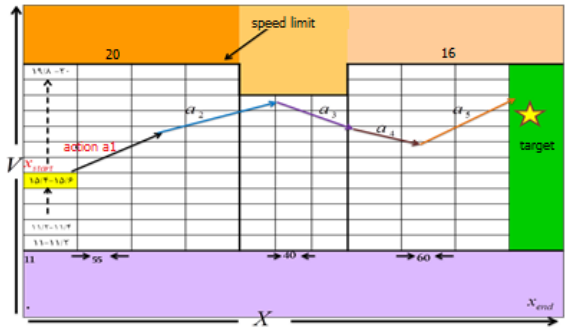
شکل (٦) این بخش‌ها که بر اساس طول مناطق محدودیت سرعت به مسافت‌های کوتاه تقسیم می‌شوند، را نشان می‌دهد.



شکل ٦. ساختار نحوه تعیین متغیرهای حالت

٤-٢-٤- اقدامات قطار

اقدامات فرض شده برای ایجاد پروفایل حرکت قطار بهینه به شرح زیر است. شتاب‌گیری در s_2، s_4 و s_6 و اقدام به خلاصی در s_5، s_7، s_8، s_{10} و s_{15} هستند. اقدامات با فلش در شکل (۷) نشان داده شده است که نشان دهنده توالی اقدامات انتخاب شده است که در آن عامل بدون دریافت پاداش به هدف رسیده است.



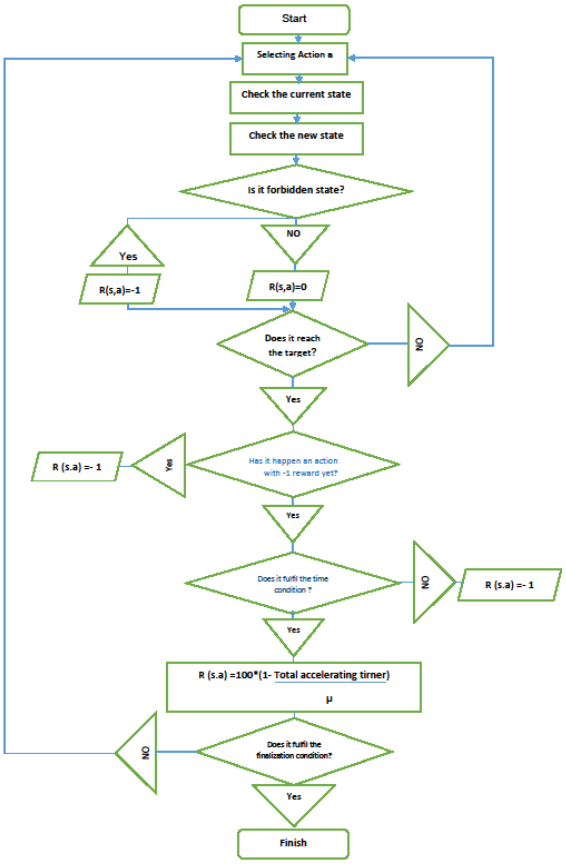
شکل ۷. ترتیب اقدامات اتخاذ شده

٤-٢-٥- ماتریس پاداش

مهم‌ترین بخش روش پیشنهادی ایجاد ماتریس پاداش است. برای این منظور از روش زیر استفاده می‌شود: ابتدا تمام عناصر ماتریس پاداش با بی‌نهایت پر می‌شوند. تعداد ستون‌ها و ردیف‌های ماتریس به ترتیب برابر با تعداد حالت‌ها و اقدامات است. از آنجایی که در هر ایستگاه قطار با سرعت صفر و حداکثر شتاب سفر خود را آغاز می‌کند که در آن متغیرهای حالت برابر با x_{start} و v_{start} است. پس از مشخص شدن حالت اول عامل، الگوریتم تعریف شده برای تولید ماتریس پاداش آغاز می‌شود. هرچه تکرار حلقه الگوریتم بیشتر باشد، عناصر بیشتری از ماتریس پاداش با پاداش‌های به دست آمده پر می‌شود. نحوه تخصیص پاداش به عوامل به شرح ادامه

است. عامل با انتخاب یک عمل از حالت فعلی خود به حالت جدید منتقل می‌شود. اگر حالت جدید یکی از حالت‌های ممنوعه نباشد، عامل مقدار صفر پاداش می‌گیرد. در غیر این صورت یک پاداش دریافت می‌کند. همچنین اگر عامل در یکی از حالت‌های ممنوعه باشد و هر یک از اقدامات تعریف شده را انتخاب کند، پاداشی معادل ۱- به عنصر مربوط به آن اقدام/حالت اختصاص می‌یابد. فرآیند انتخاب کنش، انتقال به حالت‌های جدید و دریافت پاداش تا زمانی که عامل به یکی از حالات جذب کننده برسد ادامه می‌یابد. این فرآیند از ابتدایی تا حالت‌ها، اپیزود نامیده می‌شود. ممکن است عامل در هر قسمت با چهار حالت زیر روبرو شود.

در حالت اول، عامل در یک قسمت حداقل یک بار به حالت ممنوعه منتقل شده است. در حالت دوم، عامل هرگز در یک قسمت به حالت ممنوعه منتقل نشده است. همچنین محدودیت زمانی را رعایت نکرده است. در حالت سوم، عامل هرگز در یک قسمت به حالت ممنوعه منتقل نشده است و از محدودیت‌های زمانی پیروی می‌کند، اما از نقطه X_{end} عبور کرده است. در حالت چهارم، عامل هرگز در یک قسمت به حالت ممنوعه منتقل نشده و تحت محدودیت زمانی تعریف شده به هدف رسیده است. در تمام حالت‌های ذکر شده، اگر عامل به حالت ممنوعه منتقل شود، پاداش ۱- دریافت می‌کند و اگر به حالت ممنوعه منتقل نشود، پاداش ۰ می‌شود. در غیر این صورت، اگر عامل به هیچ حالت ممنوعه منتقل نشود و به یکی از حالت‌های جذب تحت محدودیت زمانی تحمیل شده برسد، پاداش R را دریافت می‌کند. تمام مراحل الگوریتم ارائه شده در شکل (۸) باید چندین بار تکرار شوند (در شبیه‌سازی‌ها، ۲۰۰ بار فرض می‌شود)، به طوری که عامل تمام حالت‌های ممکن را تجربه کرده و نتایج آنها را به ماتریس پاداش اختصاص دهد. به همین ترتیب، تمام حالت‌هایی که امکان تجربه ۰ را بعد از تمام تکرارها دارند، به جای inf، منفی یک را به عناصر ماتریس مرتبط خود اختصاص می‌دهند. پس از تولید ماتریس پاداش، مرحله بعدی یادگیری Q و ارائه استراتژی بهینه عملیات قطار است.



شکل ۸ الگوریتم ماتریس پاداش

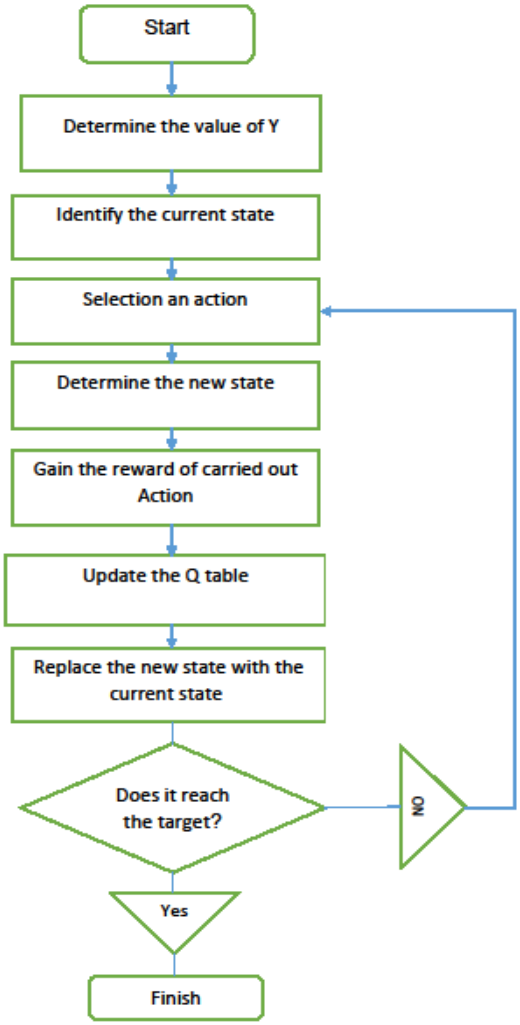
۴-۲-۶- الگوریتم یادگیری Q

مراحل الگوریتم Q به شرح زیر است. ابتدا باید γ و ماتریس پاداش را پیدا کنیم، سپس ماتریس Q را با تمام عناصر اولیه به صفر تولید کنیم، سپس مراحل زیر را برای هر قسمت تکرار کنیم. وضعیت فعلی را شناسایی کنیم، مراحل زیر را تا رسیدن به هدف تکرار کنیم. انتخاب یک عمل از تمام اقدامات ممکن، انتقال به حالت جدید در نتیجه اقدام انتخاب شده، دریافت پاداش مربوطه، محاسبه حداکثر مقدار Q برای اقدامات ممکن در حالت جدید، یافتن پاسخ محاسبه Q و جایگزینی وضعیت فعلی با حالت جدید. عامل با استفاده از الگوریتم فوق به اندازه کافی برای یادگیری تجربه می‌شود. هر قسمت برابر با یک مرحله یادگیری است. در هر مرحله یادگیری، عامل محیط را جستجو می‌کند و تا زمانی که به یکی از حالت‌های هدف برسد، پاداش دریافت می‌کند. این عمل توسط ماتریس پاداش ارائه می‌شود. هدف از یادگیری بهبود مغز عامل است که توسط ماتریس Q نشان داده شده است. یادگیری بیشتر منجر به ماتریس Q بهینه می‌شود. در راه‌حل پیشنهادی، هرچه یادگیری و بهینه بودن ماتریس Q بیشتر باشد، بازدهی پروفیل سرعت

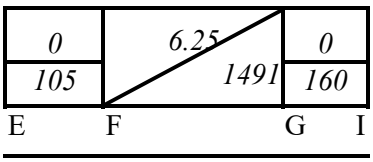
تحویل راننده قطار بیشتر خواهد بود. به طور کلی، این منجر به راندمان بیشتر و مصرف انرژی کمتر می‌شود. فلوچارت الگوریتم Q در شکل (۹) ارائه شده است. پارامتر گاما را می‌توان از بازه ۰ تا ۱ انتخاب کرد. اگر گامای انتخاب شده نزدیک به ۰ باشد، عامل تمایل به انتخاب پاداش‌های آنی دارد. با استفاده از ماتریس Q، عامل یک سری حالت از حالت اولیه تا هدف را ردیابی می‌کند. الگوریتم اقداماتی را پیدا می‌کند که برای هر حالت بیشترین تعداد پاداش را از ماتریس Q دریافت کرده‌اند. پس از اتمام مراحل یادگیری و جستجو، از الگوریتم نشان داده شده در شکل (۹) بایستی استفاده شود. بر اساس این الگوریتم، در مرحله ۱، حالت اولیه را با حالت فعلی جایگزین کنید. در مرحله ۲، اقدامی را با بالاترین پاداش از وضعیت فعلی انتخاب کنید. در مرحله ۳ حالت بعدی را با حالت فعلی جایگزین کنید و در مرحله ٤ مراحل ۲ و ۳ را تکرار کنید تا حالت فعلی با حالت هدف جایگزین شود.

ارائه نتایج

هر دو روش پیشنهادی و الگوریتم ژنتیک پس از پیاده‌سازی و کدگذاری در نرم‌افزار *Matlab* شبیه‌سازی شده‌اند. در تحلیل، دو مسیر نمونه از خط ۳ مترو تهران انتخاب شده است. همچنین مسیرهای ساده و پیچیده با محدودیت سرعت و شیب ملایم یا تند برای تحلیل روش پیشنهادی انتخاب شده‌اند تا تمامی جوانب مسئله به طور کامل بررسی شوند. اطلاعات مسیر در شکل (۱۰) نشان داده شده است. در مسیر ۱، طول ED، ۱۰۵ متر و شیب آن صفر است، طول FG، ۱٤۹۱ متر با شیب ٦,۲۵ در ۱۰۰۰، و طول GI، ۱٦۰ متر با شیب ۰ و طول EI، ۱٦٥٤ متر است.

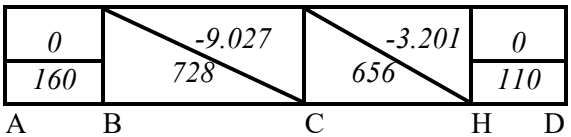


شکل ۹. الگوریتم یادگیری Q



شکل ۱۰. مشخصات مسیر اول

برای مسیر ۲، طول ایستگاه AB، ۱٦۰ متر با شیب ۰، طول BC، ۷۲۸ متر با شیب ۹,۰۲۷- در ۱۰۰۰، طول CH، ٦٥٦ متر با شیب ۳,۲۰۱- و طول ایستگاه HD، ۱۱۰ متر است. شکل (۱۱) مشخصات هندسی مسیر ۲ را نشان می‌دهد.



شکل ۱۱. مشخصات مسیر دوم

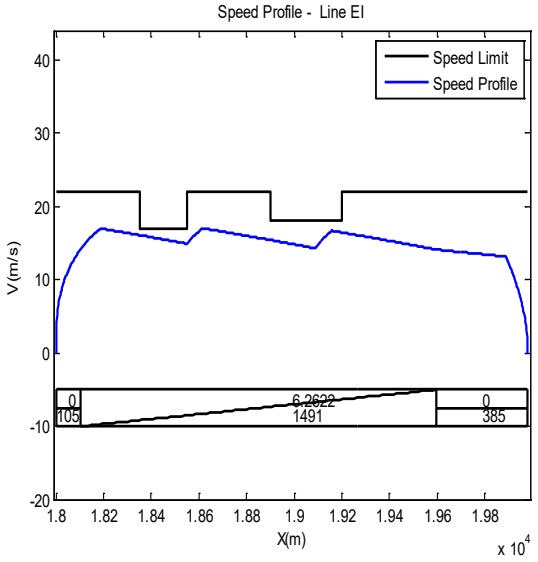
به منظور تحلیل دقیق و مقایسه‌ای روش پیشنهادی برای هر مسیر، الگوریتم ماتریس پاداش برای هر تکرار متفاوت اجرا شده است. x_{start} و x_{end} محاسبه شده برای مسیر ۱ به ترتیب ۱۰۹۷۸ و ۱۱۷۰۰ و برای مسیر ۲، ۱۸۲۲۰ و ۱۹۵۰۰ هستند.

اقدامات در نظر گرفته شده برای عامل در شبیه‌سازی مسیر ۱ و ۲ در جدول (۲) نشان داده شده است. در این جدول، عمل ۱ حرکت با حداکثر شتاب و عمل ۰ حرکت خلاص است.

جدول ۲. اقدامات در نظر گرفته شده برای عامل

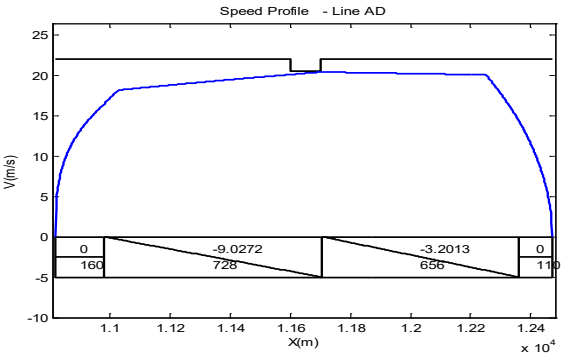
Action	Action Type	Action Time
1	1	3
2	1	3
3	1	4
4	1	5
5	1	4
6	1	5
7	0	5
8	0	7
9	0	8
10	0	10
11	0	7
12	0	15
13	0	10
14	0	10
15	0	5

شکل (۱۲) پروفیل سرعت قطار بهینه را پس از ۳۰۰۰ تکرار الگوریتم یادگیری Q برای مسیر اول نشان می‌دهد.



شکل ۱۲. نمودار سرعت – موقعیت برای مسیر اول

شکل (۱۳) پروفیل سرعت قطار بهینه را پس از ۳۰۰۰ تکرار الگوریتم یادگیری Q برای مسیر دوم نشان می‌دهد.



شکل ۱۳. نمودار سرعت – موقعیت برای مسیر دوم

شبیه‌سازی‌های ارائه شده نیز به منظور مقایسه با روش GA انجام شده و نتایج در جدول (۳) منعکس شده است.

جدول ۳. مقایسه نتایج RL و GA

Route Number		R1	R2
Desired Travel Time		112	117
Total Travel Time	GA	112	117
	RL, 200	32.11	39.11
Energy Consumption	GA	43.40	733.4
	RL, 200	5.418	54.45
	RL, 150	5.517	5223
	RL, 500	5.536	5142

در این جدول نتایج شبیه‌سازی GA با ۲۰۰ تکرار و شبیه‌سازی RL با تکرارهای مختلف ارائه شده است. اگرچه GA سرعت بیشتری نسبت به روش RL پیشنهادی دارد، اما روش دوم توانایی ارائه اکثر مراحل تولید پروفایل سرعت را در حالت آفلاین دارد. اگر ماتریس پاداش تعیین شود، زمان تولید پروفایل بهینه را می‌توان به چند ثانیه کاهش داد. از آنجایی که GA مبتنی بر انتخاب تصادفی است و پاسخ بهینه زمانی پیدا می‌شود که تکرارهای متعدد انجام شده باشد، استفاده از آن به دلیل محدودیت‌های زمانی در حالت آنلاین غیرممکن می‌شود.

۵- نتیجه‌گیری

در این مقاله یک روش کنترل بهینه برای تولید پروفیل سرعت برای کاهش مصرف انرژی قطار معرفی شد. الگوریتم جدیدی با استفاده از یادگیری تقویتی پیشنهاد شد و نتایج شبیه‌سازی نشان می‌دهد که روش پیشنهادی نه تنها در تولید سریع‌تر پروفایل آنلاین، بلکه در دقت نسبت به روش‌های مشابه برتری دارد. در روش پیشنهادی سیستم به گونه‌ای انتخاب شد که تمام حالت‌های قطار بررسی شوند و

کاستی‌های روش‌های مشابه مانند GA (که بر اساس انتخاب تصادفی کار می‌کند و نمی‌تواند پروفایل را با تکرارهای کمتر تولید کند)، مشخص شود. در پروفیل بهینه پیشنهادی، شتاب و خلاص به گونه‌ای انتخاب شدند که علاوه بر رعایت محدودیت‌های سرعت و زمان، مصرف انرژی به حداقل برسد. برای ادامه کار، پیشنهاد می‌شود برای تولید ماتریس پاداش، حالت‌ها در شرایط اولیه در نظر گرفته شوند. برای اینکه قطار در هر یک از حالت‌ها مسیر بهینه را در چند ثانیه طی کند، باید قطار با تکرارهای متعدد به دنبال مسیر بهینه در همه حالت‌ها باشد. در این حالت حتی اگر قطار در وسط مسیر متوقف شود، در عرض چند ثانیه یک پروفایل جدید تولید می‌شود و استراتژی‌های حرکت به صورت آنلاین به قطار تحویل داده می‌شود.

٦-مراجع

-Blanco-Castillo, M., Fernández-Rodríguez, A., Fernández-Cardador, A., & Cucala, A. P. (2022). Eco-driving in railway lines considering uncertainty associated with climatological conditions. *Sustainability*, *14*(14), 8645.

Ding, Y., Liu, H., Bai, Y., & Zhou, F. (2011). A Two-level Optimization Model and Algorithm for Energy-Efficient Urban Train Operation. *Journal of Transportation Systems Engineering and Information Technology, 11*(1), 96-101.

-Esveld, C. (2001). Modern railway track. *MRT-Productions*.

-Gosavi, A. (2009). Reinforcement learning: a tutorial survey and recent advances. *INFORMS Journal on Computing, 21*(2), 178-192.

-Howlett, P. (1996). Optimal strategies for the control of a train. *Automatica, 32*(4), 519-532.

-Howlett, P. (2000). The optimal control of a train. *Annals of Operations Research, 98*(1), 65-87.

-Howlett, PG, & Leizarowitz, A. (2001). Optimal strategies for vehicle control problems with finite control sets. *Dynamics of Continuous Discrete and Impulsive Systems Series B, 8*, 41-70.

-Howlett, PG, & Cheng, J. (1997). Optimal driving strategies for a train on a track with continuously varying gradient. *Journal of the Australian Mathematical Society-Series B, 38*(3), 388-410.

-Hu, H., Fu, Y., & Hu, C. (2010). PSO-based optimal operation strategy of energy saving control for train. *Paper presented at the Industrial Engineering and Engineering Management (IE&EM), 2010 IEEE 17Th International Conference.*

-Hwang, H.S. (1998). Control strategy for optimal compromise between trip time and energy consumption in a high-speed railway. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, 28*(6), 791-802.

-Kang, M.H. (2011). A GA-based Algorithm for Creating an Energy-Optimum Train Speed Trajectory. *The Journal of International Council on Electrical Engineering, 1*(2), 123-128.

-Khmelnitsky, E. (2000). On an optimal control problem of train operation. *Automatic Control, IEEE Transactions on, 45*(7), 1257-1266.

-Liao, J., Zhang, F., Zhang, S., Yang, G., & Gong, C. (2021). Energy-saving optimization strategy of multi-train metro timetable based on dual decision variables: A case study of Shanghai Metro line one. *Journal of Rail Transport Planning & Management*, *17*, 100234.

-Lu, Q., & Feng, X. (2011). Optimal control strategy for energy saving in trains under the fou-aspect fixed autoblock system. *Journal of Modern Transportation, 19*(2), 82-87.

-Milroy, I.P. (1980). *Aspects of automatic train control.* Loughborough University of Technology.
Monajem, M.S. (2007), Designing Metro and Railway Lines, *1st Edition, Angize Publications.*

-Monajem, M.S. & Bababeik, M. (2009). A study of train behavior in a track using movement simulation algorithms, *11th Railway Transportation Conference,* Nov 2009, *Iranian Association of Rail Transport Engineering, Tehran*.

-Nolte, R. (2003). Event Evaluation of Energy Efficiency Technologies for Rolling Stock and Train Operation of Railways-Final Report. *Studie im Auftrag des Internationalen Eisenbahnverbands*.

-Rochard, BP, & Schmid, F. (2000). A review of methods to measure and calculate train resistances. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit, 214*(4), 185-199.

-Sandidzadeh, M. A., Askarian, M., & Soleymaani, F. (2020). The Effect of Using the Tabu Search Algorithm on the Speed of Achieving the Optimal Train Speed Profile (in order to Reduce Energy Consumption). *Journal of Transportation Research*, *17*(4), 31-48.

-Sheu, J.W., & Lin, W.S. (2012). Energy-Saving Automatic Train Regulation Using Dual Heuristic Programming. *Vehicular Technology, IEEE Transactions on, 61*(4), 1503-1514.

-Sutton, R.S., & Barto, A.G. (1998). Reinforcement learning: An introduction. Vol. 1, *Cambridge Univ. Press.*

-Xu, R., Meng, J., Li, D., & Chen, X. (2023). Energy-Efficient Optimization Method of Urban Rail Train Based on Following Consistency. *Energies*, *16*(4).

-Yang, L., Li, K., Gao, Z., & Li, X. (2011). Optimizing trains movement on a railway network. *Omega*.

-Yeo, C., & Koseki, T. (2002). Optimization of Running Profile of Train by Dynamic Programing. *Paper presented at the National Convention of IEEE.*

-Zhu, Q., Su, S., Tang, T., Liu, W., Zhang, Z., & Tian, Q. (2022). An eco-driving algorithm for trains through distributing energy: A Q-Learning approach. *ISA Transactions, 122,* 24-37.

Proposing an Optimal Control Method for Energy Consumption Optimization in Railway Signalling Systems Using Reinforcement Learning

Mohammad Ali Sandidzadeh, Associate Professor, Faculty of Railway Engineering, Iran University of Science and Technology, Tehran, Iran.
Majid Azinfar, M.Sc., Grad., Faculty of Railway Engineering, Iran University of Science and Technology, Tehran, Iran.
Nasser Mozayani, Associate Professor, Faculty of Computer Engineering, Iran University of Science and Technology, Tehran, Iran.
Farzaad Soleymaani, Postdoctoral Researcher, Faculty of Railway Engineering, Iran University of Science and Technology, Tehran, Iran.

E-mail: sandidzadeh@iust.ac.ir

Received: November 2024- Accepted: February 2025

ABSTRACT
Nowadays, the optimization of energy consumption in public transportation systems is a serious issue. Since a large part of energy in transportation systems is consumed by subways, a new approach has been proposed for optimal control of a train to reduce energy consumption. The proposed model is based on the Reinforcement Learning algorithm. It is assumed that a train moves between two stations along a line with non-constant gradient, curve, and speed limits. Moreover, the train should complete its journey within a given time interval. The Reinforcement Learning of States, Actions, and Rewards are based on the selected Actions. In the proposed method, the train States are the velocity and position of the train, and the Action is acceleration or coasting motion. Unlike the former techniques, most stages of optimization in this method are offline and implemented only once for any route. Following the formation of the reward matrix, we could use this method in an online form and then the speed profile could be produced at a minimum time. The simulations of the proposed method are implemented in MATLAB and finally compared with those of the Genetic Algorithm.

Keywords: Train Speed Profile, Optimal Control, Energy Consumption Optimization, Reinforcement Learning Method, Railway Transportation System