درس روش پژوهش و ارائه

بازی گو با یادگیری تقویتی

نویسنده:

امین خانی aminkhani@std.kashanu.ac.ir

فروردین ۱۴۰۲

فهرست مطالب

۱.۰ مقدمه

یادگیری ماشین [1] پدیده‌ای نیست که اخیرا به وجود آمده باشد. در واقع شبکه‌های عصبی برای اولین بار به عنوان یک مفهوم در یک مقاله تحقیقاتی در سال ۱۹۴۳ معرفی شد. اگرچه در روزهای ابتدایی پیشرفت در یادگیری ماشین به دلیل هزینه بالای محاسبات تا حدی کند بود که این حوزه را فقط در دسترس موسسات دانشگاهی بزرگ یا شرکت‌های چند ملیتی قرار می داد. همچنین این مسئله وجود داشت که تهیه داده‌های مورد نیاز برای آموزش سیستم‌ها بسیار دشوار بود. اما امروزه با ظهور اینترنت بسیاری از مشکلات بر سر راه هوش مصنوعی و یادگیری ماشین از بین رفته و با پیشرفت سخت افزارها هزینه‌های تهیه سخت افزارهای مورد نیاز برای پیاده سازی هوش مصنوعی و یادگیری ماشین کاهش یافت که همین مسئله باعث پیشرفت بیشتر این حوزه در سال‌های اخیر شد.
یکی از روش های یادگیری ماشین، یادگیری تقویتی [2] است که در آن، عامل یادگیری پس از هر اقدام، بازخوردی به صورت پاداش و یا جریمه دریافت می کند. این روش اغلب در بازی‌ها (از جمله بازی‌های آتاری و ماریو) به‌کار گرفته می‌شود و عملکرد آن در سطح انسان و حتی گاهی فراتر از توانایی ما است.

۲.۰ انواع یادگیری

یادگیری تقویتی را می‌توان شاخه‌ای مجزا در یادگیری ماشین در نظر گرفت، هرچند شباهت‌هایی هم با سایر روش‌های یادگیری ماشین دارد. برای دریافتن این شباهت‌ها و تفاوت‌ها بهتر است نیم نگاهی به سایر روش‌ها داشته باشیم.

۱.۲.۰ یادگیری با ناظر (Supervised Learning)

در یادگیری با ناظر [1]، الگوریتم‌ها با استفاده از یک سری داده برچسب‌دار(label) آموزش داده می‌شوند. این الگوریتم‌ها فقط ویژگی‌هایی را یاد می‌گیرند که در دیتاست مشخص شده است و به آنها هدف یا target گفته می‌شود. در واقع «هدف» در این نوع یادگیری کاملا تعریف شده است و نمونه‌های از داده و پاسخ درست در اختیار مدل قرار می‌گیرد تا با استفاده از آنها بتواند هر داده‌ی جدیدی را که می‌بیند برچسب بزند. یکی از رایج‌ترین کاربردهای یادگیری نظارتی، مدل‌های تشخیص تصویر است. این مدل‌ها یک مجموعه عکس برچسب‌دار دریافت می‌کنند و یاد می‌گیرند بین ویژگی‌های متداول آنها تمایز قائل شوند. به عنوان مثال با دریافت عکس‌هایی از صورت انسان‌ها، می‌توانند اجزای صورت را تشخیص دهند. یا بین دو یا چند حیوان تمایز قائل شوند.

۲.۲.۰ یادگیری بدون ناظر (Unsupervised Learning)

در یادگیری بدون ناظر [3]، فقط داده‌های بدون برچسب در اختیار الگوریتم قرار داده می‌شود. این الگوریتم‌ها بدون اینکه مستقیم به آنها گفته شده باشد دنبال چه ویژگی‌ها بگردند، براساس مشاهده های خودشان آموزش می‌بینند. نمونه‌ای از کاربرد این نوع یادگیری، خوشه‌بندی مشتری‌ها است.

۳.۲.۰ یادگیری نیمه نظارتی (Semi supervised Learning)

این روش [4]، روشی بینابینی است. توسعه‌دهندگان، یک مجموعه نسبتا کوچک از داده‌های برچسب‌دار و یک مجموعه بزرگ‌تر از داده بدون برچسب آماده می‌کنند. سپس از مدل خواسته می‌شود، براساس چیزی که از داده‌های برچسب‌دار یاد می‌گیرد، درمورد داده‌های بدون برچسب هم پیش‌بینی انجام دهد و در نهایت داده‌های بدون برچسب و برچسب‌دار را به عنوان یک مجموعه داده کل درنظر بگیرد و نتیجه‌گیری نهایی را انجام دهد.

۴.۲.۰ یادگیری تقویتی (Reinforcement Learning)

رویکرد یادگیری تقویتی کاملا متفاوت است. در این روش، یک عامل در محیط قرار می‌گیرد تا با آزمون و خطا یاد بگیرد کدام کارها مفید و کدام کارها غیرمفید هستند و در نهایت به یک هدف مشخص برسد. از این جهت که درمورد یادگیری تقویتی هم هدف مشخصی از یادگیری وجود دارد، می‌توان آن را شبیه یادگیری با ناظر دانست. اما وقتی که اهداف و پاداش‌ها مشخص شدند، الگوریتم به صورت مستقل عمل می‌کند و نسبت به یادگیری با ناظر تصمیمات آزادانه‌تری می‌گیرد. به همین علت است که برخی یادگیری تقویتی را در دسته نیمه نظارتی جای می‌دهند. اما با توجه به آنچه گفته شد، منطقی‌تر این است که یادگیری تقویتی را به عنوان یک دسته جدا در یادگیری ماشین در نظر گرفت.

۳.۰ یادگیری تقویتی

یادگیری تقویتی تلاش می‌کند کارهایی انجام دهد تا دستاورد موقعیت‌های خاص به حداکثر برسد. هدف الگوریتم‌های یادگیری تقویتی این است که بهترین کاری که می‌شود در یک موقعیت خاص انجام داد را پیدا کنند. این نوع یادگیری ماشین می‌تواند یاد بگیرد حتی در محیط‌های پیچیده و غیر مطمئن هم فرآیند یادگیری را انجام دهد و به اهدافش دست پیدا کند. این سیستم، درست مانند مغز انسان برای انتخاب‌های خوب پاداش می‌گیرد، برای انتخاب‌های بد جریمه می‌شود و از هر انتخاب یاد می‌گیرد.
ساده‌ترین مدل ذهنی که می‌تواند به درک یادگیری تقویتی کمک کند یک بازی است. جالب است بدانید الگوریتم‌های یادگیری تقویتی در بازی‌ها نقش برجسته‌ای دارند. در یک بازی معمولی شما عناصر زیر را دارید:

- یک عامل (بازیکن - **Agent**) که کارهای مختلفی انجام می‌دهد.
- کارهایی که عامل باید انجام دهد (حرکت در فضا به بالا، خرید یک وسیله یا هرچیز دیگری - **Action**).
- پاداش عامل (سکه، از بین رفتن دشمن و... - **Rewards**).
- محیطی که عامل در آن قرار دارد (یک اتاق، یک نقشه و... - **Environments**).
- هدفی برای عامل که به با دست یابی به آن به بیشترین پاداش ممکن می‌رسد (**Goal**).

همین عناصر دقیقا سازندگان یادگیری تقویتی هم هستند (شاید یادگیری ماشین در حقیقت یک بازی است.) در یادگیری تقویتی ما یک عامل را در به‌صورت مرحله به مرحله در یک محیط راهنمایی می‌کنیم و اگر کارش را در هر مرحله درست انجام دهد به او پاداش می‌دهیم.

۴.۰ کاربردهای یادگیری تقویتی

یادگیری تقویتی در حوزه ها و جاهای گوناگون ممکن است به کار گرفته شود، مانند:

۱.۴.۰ کنترل چراغ‌های راهنمایی

نویسندگان مقاله «Reinforcement learning-based multi-agent system for network traffic signal control» [5] تلاش کردند تا سیستمی برای کنترل چراغ‌های راهنمایی طراحی نمایند که مسئله ترافیک سنگین خیابان‌ها را حل کند. این الگوریتم تنها در محیط شبیه‌سازی‌شده و غیرواقعی آزمایش شد، اما نتایج آن بسیار بهتر از روش سنتی کنترل ترافیک بود و بدین ترتیب، کاربردهای بالقوه الگوریتم‌های چند عاملی یادگیری تقویتی در حوزه طراحی سیستم‌های کنترل ترافیک را برای همه آشکار کرد.

۲.۴.۰ رباتیک

بوستون داینامیکس (Boston Dynamics) یکی از شرکت‌های مطرح تولیدکننده ربات است که تاکنون توانسته ربات‌های بسیار کارآمد و مبتکرانه‌ای را برای استفاده در حوزه‌های لجستیک، صنعتی و امدادرسانی توسعه دهد.
به جرات می‌توان گفت بوستون داینامیکس یکی از پیشگامان عصر نوین فناوری محسوب می‌شود و فناوری‌ها و دستاوردهای نوین دنیای رباتیک، مهندسی مکانیک و هوش مصنوعی را به بهترین شکل ممکن به خدمت گرفته است. این شرکت در طول دهه اخیر ربات‌های شگفت‌انگیزی را تولید کرده که می‌توانند بهتر از برخی رقبای انسانی بپرند، بدوند و حتی حرکات موزون انجام دهند.

۳.۴.۰ پیکربندی سیستم وب

در هر سیستم وب بیش از ۱۰۰ پارامتر قابل پیکربندی وجود دارد. هماهنگ کردن این پارامترها نیازمند یک اپراتور ماهر و به‌کارگیری روش آزمون و خطا است. مقاله «Reinforcement Learning Approach to Online Web System Auto-configuration» [6] یکی از اولین تلاش‌ها در این زمینه است که نحوه پیکربندی مجدد پارامترها در سیستم‌های وب چند لایه در محیط‌ها پویای مبتنی بر ماشین مجازی را بررسی می‌کند.

۴.۴.۰ پیشنهادات شخصی‌سازی‌شده

کارهای پیشین در زمینه پیشنهاد اخبار با چالش‌هایی از جمله سرعت بالای تغییرات در پویایی اخبار، نارضایتی کاربران و نامناسب بودن معیارها مواجه شدند. فردی به نام گوانجی برای غلبه بر این مشکلات، در سیستم پیشنهاد اخبار خود از یادگیری تقویتی استفاده کرد و نتایج این کار را در مقاله‌ای با عنوان «DRN: A Deep Reinforcement Learning Framework for News Recommendation» [7] منتشر کرد.

۵.۴.۰ مزایده و تبلیغات

محققین گروه Alibaba مقاله‌ای با عنوان «Real-Time Bidding with Multi-Agent Reinforcement Learningin Display Advertising» [5] منتشر کردند و ادعا کردند که راه‌کار آن‌ها با عنوان مزایده چند عاملی توزیعیو مبتنی بر خوشه‌بندی (DCMAB) نتایج امیدوارکننده‌ای به دنبال داشته است و به همین دلیل، قصد دارند آن را به‌صورت زنده بر روی سامانه TaoBao محک بزنند.

۶.۴.۰ بازی ها

شناخته شدن الگوریتم‌های یادگیری تقویتی عمدتاً به دلیل کاربردهای گسترده آن در بازی‌ها و گاه عملکرد فرابشری این الگوریتم‌ها بوده است.
نام‌آشناترین الگوریتم‌ها در این حوزه AlphaGo [8] و AlphaGo Zero هستند. برای آموزش الگوریتم آلفاگو داده‌های بیشماری از روند بازی‌های انسانی جمع‌آوری و به آن داده شد. این الگوریتم با بهره‌گیری از تکنیک جست‌وجوی درختی مونت کارلو (MCTS) و شبکه ارزش تعبیه شده در شبکه سیاست خود توانست عملکردی فرابشری داشته باشد.
آلفاگو (AlphaGo) یک برنامه رایانه‌ای است که توسط DeepMind گوگل در لندن، برای بازی تخته‌ای گو (Go) توسعه یافته‌است. در اکتبر ۲۰۱۵، آلفاگو ــ اولین برنامهٔ رایانه‌ای گو بود که با غلبه بر بازیکن‌های حرفه‌ای بازی گو بدون دادن آوانس روی یک تخته کامل در سایز ۱۹ × ۱۹ انجام شد. و در مارچ ۲۰۱۶، بر لی سِدُل در پنج دوره بازی غلبه کرد، این اولین باری بود که یک برنامه رایانه‌ای گو بر یکی از ۹ــ دانهای حرفه‌ای بدون آوانس غلبه می‌کرد؛ اگرچه در چهارمین بازی به لی سدل باخت، اما لی، بازی آخر را درخواست داد و امتیاز آخر ۴ به ۱ بازی را به آلفاگو داد. الگوریتم آلفاگو از تکنیک جستجوی درختی مونت کارلو برای یافتن حرکات که مبتنی بر دانش قبلی یادگرفته از یادگیری ماشینی و مخصوصاً ازشبکه عصبی مصنوعی با یادگیری عمیق از هردوی انسان و اجرای رایانه است، استفاده می‌کند.

۵.۰ نتیجه گیری

دانشمندان به‌طور مداوم در حال توسعه الگوریتم‌های جدیدی هستند که به یادگیری تقویتی عمیق اجازه می‌دهد تا موفقیت بیشتری داشته باشد.

شکل‌های دیگر یادگیری تقویتی نیز در حال گسترش است، از جمله یادگیری تقویتی معکوس (inverse reinforcement learning) [9]، که در آن ماشین از مشاهده یک فرد متخصص یاد می‌گیرد. ماشین به جای تلاش برای یادگیری از تجربه‌ی خود، از تماشای دیگران یاد می‌گیرد. آن متخصص دیگر، معلم نیست، فقط کسی یا چیزی است که یک کار را اجرا می‌کند، نه اینکه آن کار را توضیح دهد.

یادگیری تقویتی مشروط به هدف (Goal-conditioned reinforcement) [10] مشکلات پیچیده یادگیری تقویتی را با استفاده از اهداف فرعی از بین می‌برد.

یادگیری تقویتی چندعاملی (Multi-agent reinforcement learning) [5] در حل مشکلات رباتیک، مخابرات و اقتصاد بسیار مفید است.

با وجود پتانسیل‌های یادگیری تقویتی، پیاده‌سازی آن می‌تواند دشوار باشد و به همین علت کاربردهای آن هنوز محدود مانده است. یکی از موانع پیاده‌سازی این نوع یادگیری ماشین، لزوم و تکیه این روش برای جستجو و کشف در محیط مورد نظر است.

به عنوان مثال، اگر بخواهیم رباتی را آموزش دهیم که در یک محیط فیزیکی حرکت کند، هر چه جلو می‌رود، با حالت‌های جدیدی مواجه می‌شود که در برابر آنها باید عمل‌های مختلفی را باید انجام دهد. در دنیای واقعی، کار آسانی نیست که به طور پیوسته، بهترین عمل‌ها انتخاب و انجام شوند؛ چرا که محیط پیوسته در حال تغییر است.

یکی دیگر از مشکلاتی که سر راه یادگیری تقویتی وجود دارد، مدت زمان و منابع محاسباتی‌ای است که لازم است تا اطمینان حاصل کنیم یادگیری به درستی انجام شده است. از طرفی، هرچه محیط آموزشی بزرگ‌تر باشد به زمان و منابع بیشتری برای فرآیند آموزش الگوریتم نیاز است.

مراجع

[۱] Mitchell, T.M., 2007. Machine learning (Vol. 1). New York: McGraw-hill.

[۲] Sutton, R.S. and Barto, A.G., 2018. Reinforcement learning: An introduction. MIT press.

[۳] Barlow, H.B., 1989. Unsupervised learning. Neural computation, 1(3), pp.295-311.

[۴] Zhu, X.J., 2005. Semi-supervised learning literature survey.

[۵] Buşoniu, L., Babuška, R. and De Schutter, B., 2010. Multi-agent reinforcement learning: An overview. Innovations in multi-agent systems and applications-1, pp.183-221.

[۶] Bu, X., Rao, J. and Xu, C.Z., 2009, June. A reinforcement learning approach to online web systems auto-configuration. In 2009 29th IEEE International Conference on Distributed Computing Systems (pp. 2-11). IEEE.

[۷] Zheng, G., Zhang, F., Zheng, Z., Xiang, Y., Yuan, N.J., Xie, X. and Li, Z., 2018, April. DRN: A deep reinforcement learning framework for news recommendation. In Proceedings of the 2018 world wide web conference (pp. 167-176).

[۸] Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M. and Dieleman, S., 2016. Mastering the game of Go with deep neural networks and tree search. nature, 529(7587), pp.484-489.

[۹] Ng, A.Y. and Russell, S., 2000, June. Algorithms for inverse reinforcement learning. In Icml (Vol. 1, p. 2).

[۱۰] Liu, M., Zhu, M. and Zhang, W., 2022. Goal-conditioned reinforcement learning: Problems and solutions. arXiv preprint arXiv:2201.08299.

[۱۱] Jin, J., Song, C., Li, H., Gai, K., Wang, J. and Zhang, W., 2018, October. Real-time bidding with multi-agent reinforcement learning in display advertising. In Proceedings of the 27th ACM international conference on information and knowledge management (pp. 2193-2201).