مدل سازی جدیدترین الگوریتم های یادگیری تقویتی در فرایندهای تصمیم گیری غیر متعارف مارکوف

دانشجو: محمد بیات

استاد راهنما: دکتر خونساری

دانشکده مهندسی برق و کامپیوتر، دانشگاه تهران

چکیده

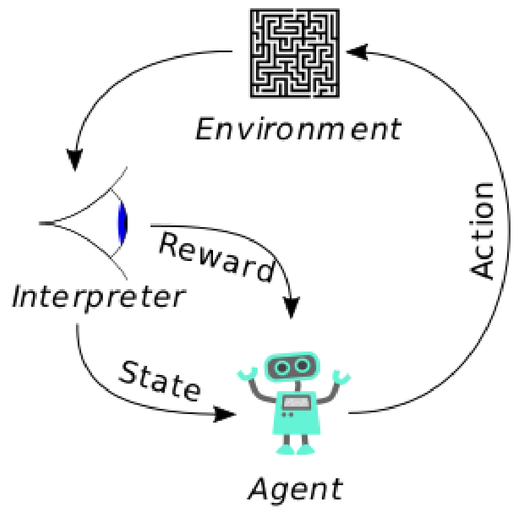
امروزه شرکت‌ها و دولت‌های بسیاری سرمایه‌گذاری زیادی روی هوش مصنوعی انجام داده‌اند تا به صورت بهینه اهداف خود را محقق کنند. یکی از جنبه‌های مهم عامل‌های هوشمند توانایی یادگیری در محیط است تا بتواند با استفاده از تجربیات گذشته تصمیم بهتری در آینده بگیرند. به همین منظور، یادگیری تقویتی با تعامل با محیط بیرونی سعی در بهبود رفتار خود دارد و این رویکرد یادگیری تقویتی را یکی از گرایش‌های مهم هوش مصنوعی قرار داده است. در این پروژه، به پیاده‌سازی و مقایسه‌ی جدیدترین الگوریتم‌های یادگیری تقویتی می‌پردازیم، که محیط آن‌ها را به کمک فرآیندهای تصمیم‌گیری مارکوف می‌توان مدلسازی کرد. در اینجا یادگیرنده فقط به یک رشته از مشاهدات دسترسی دارد. در این پروژه معیارهای کارایی مثل همگرایی، پیچیدگی نمونه و پشیمانی را اندازه‌گیری می‌کنیم. علاوه بر آن نگاهی بر تاثیر ویژگی‌های فرآیند تصمیم‌گیری غیرمتعارف مارکوف (مانند قطر) در مقایسه‌مان می‌اندازیم. در نهایت الگوریتم‌ها را از نظر مجموع خطا با یکدیگر مقایسه می‌کنیم.

ساختار

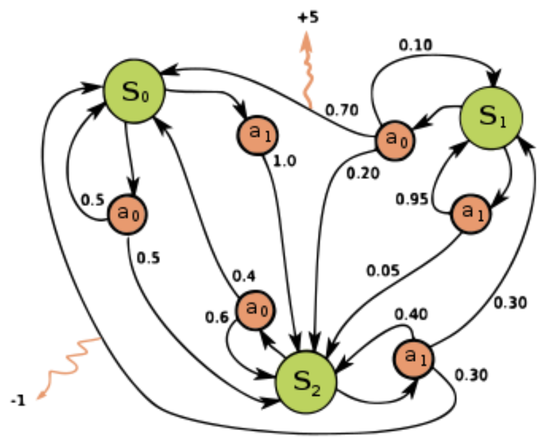
مسائل یادگیری تقویتی معمولا به کمک فرآیند تصمیم‌گیری مارکوف مدل‌سازی می‌شوند. هر مسئله‌ی فرآیند تصمیم‌گیری مارکوف به کمک حالات، عمل‌ها، ماتریس پاداش و ماتریس احتمالات مدل می‌شود.

در این پروژه الگوریتم‌های مختلف همان به کمک کلاس عامل و خود فرآیند مارکوف به کمک یک کلاس به نام محیط در کد مدل شده‌اند. به کمک این مدل می‌توان محیط‌های مختلف را با الگوریتم‌های مختلف تست کرد و به نوعی زیرساختی برای توسعه و تست سریع الکوریتم‌ها در اختیار داریم.

الکوریتم‌های پیاده‌سازی شده عبارت‌اند از: UCRL2 ، Epsilon Greedy و انواع مختلف الگوریتم‌های PSRL

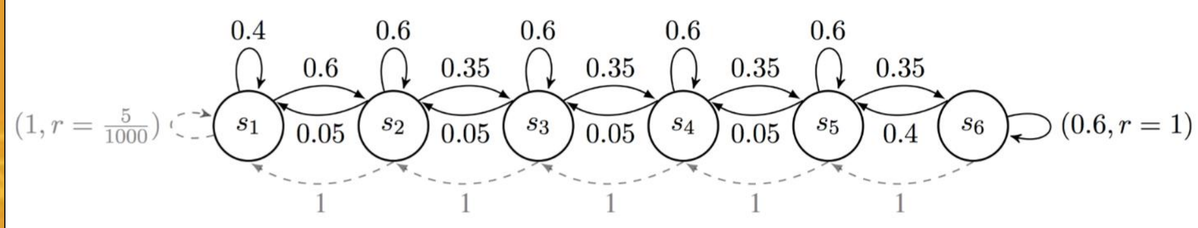


نمونه‌ای از فرآیند تصمیم‌گیری مارکوف



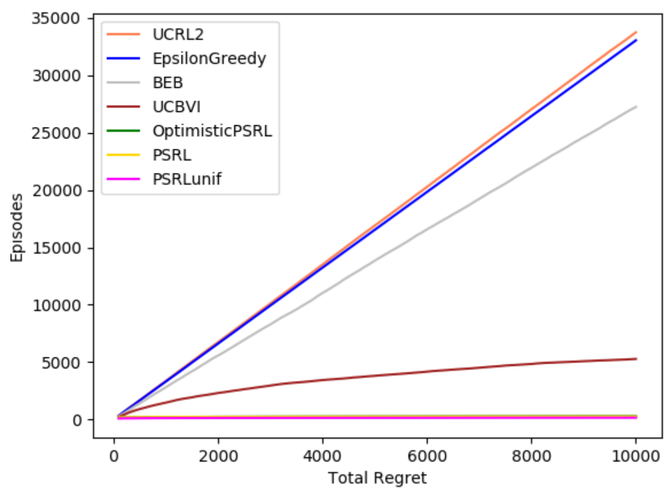
مدلی ساده از مسئله‌ی یادگیری تقویتی

محیط تست این پروژه نیز محیط استاندارد River Swim است که دارای ۶ حالت بوده و نسبتا قطر طولانی‌ای دارد.

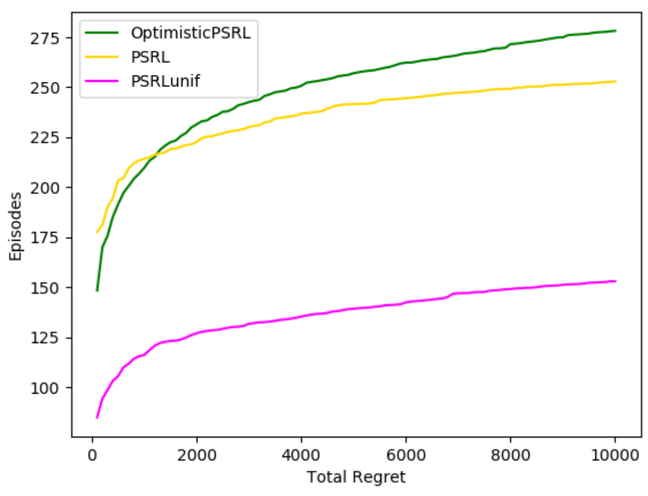


نتایج

با اجرای الگوریتم‌های مختلف بر روی محیط River Swim به این نتیجه رسیدیم که الگوریتم‌های مبتنی بر روش‌های بیزی (خانواده‌ی PSRL) از سایر الگوریتم‌ها مجموع خطای کمتری دارند و به سرعت خطای آن‌ها کاهش داشته و فرم لگاریتمی به خود می‌گیرد.



در شکل زیر به صورت جزئی‌تر الگوریتم‌های PSRL را بررسی کرده‌ایم.



جمع بندی

با توجه به پیشرفت روزافزون یادگیری ماشین و نیاز به تست الگوریتم‌های مربوط به آن، نیاز به چارچوبی برای پیاده‌سازی، شبیه‌سازی الگوریتم‌ها در یک محیط و مقایسه‌ی آن‌ها در قالب نمودار احساس می‌شود. در این پروژه سعی شده که یک چارچوب برای برطرف‌کردن نیازهای بالا توسعه داده شود. این راهکار با جدا کردن محیط که در واقع مسئله‌ی فرآیند تصمیم‌گیری مارکوف را مدل‌سازی می‌کند و جدا کردن عامل‌ها که در واقع همان الگوریتم‌های مورد نظر ما هستند، صورت گرفته است. برای تست یک محیط و یک الگوریتم جدید کافی است در قدم اول یک محیط جدید با پارامترهای مارکوف آن مسئله بسازیم و در قدم دوم کلاس عامل را گسترش داده و توابع مربوط به بروزرسانی سیاست را پیاده‌سازی کنیم. در نهایت به کمک توابع مربوط که محیط و عامل را می‌گیرد و شبیه‌سازی را انجام می‌دهد، نتایج را در غالب فایل CSV داریم. در مرحله‌ی بعد می‌توانیم خروجی این الگوریتم را با الگوریتم‌های دیگر در غالب نمودار با یکدیگر مقایسه کنیم.

مراجع اصلی

1. **R. Fruit, M. Pirotta, A. Lazaric, R. Ortner, “Efficient Bias-Span-Constrained Exploration-Exploitation in Reinforcement Learning”, 2018. [Online]. Available: https://arxiv.org/abs/1802.04020**
2. **A. Tolver, “An Introduction to Markov Chain”, 2016, University of Copenhagen, Department of Mathematical Sciences**
3. **P. L. Bartlett, A. Tewari, “REGAL: A Regularization based Algorithm for Reinforcement Learning in Weakly Communicating MDPs”, 2012. [Online]. Available: https://arxiv.org/abs/1205.2661**