حل مسئله راهرو مارپیچ با استفاده از روش‌های تفاضل موقتی

مهران احمدپور

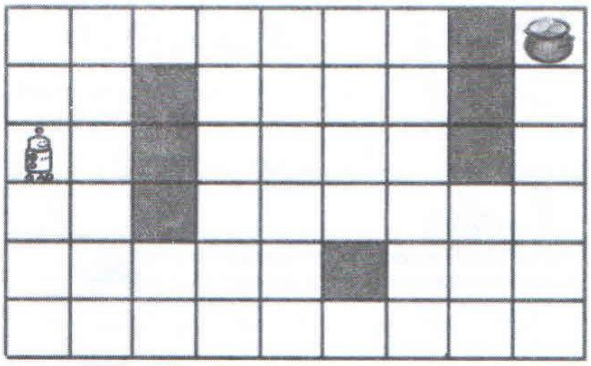
۹۸۲۰۴۹۸۵

چکیده

در این تمرین، با استفاده از روش‌های یادگیری تفاضل موقتی از جمله TD، $SARSA$ و $Q\_Learning$ تلاش شده است تا مسیر درست برای رسیدن به مقصد در راهرو مارپیچ محاسبه گردد. در انتها منحنی‌های تعداد برخورد با موانع و پاداش دریافتی در هر اپیزود برای الگوریتم‌های $SARSA$ و $Q\_Learning$ با یکدیگر مقایسه شده اند.

۱- تعریف مسئله

در تمرین حاضر هدف آن است که عامل با استفاده از روش‌های یادگیری تفاضل موقتی مسیر بهینه برای رسیدن به سلول هدف را بیابد فضای در نظر گرفته شده برای مسئله مورد نظر در شکل-۱ مشاهده می‌شود.



شکل-۱: محیط فرض شده در مسئله

در این مسئله برخورد با موانع و یا خروج از فضای مسئله پاداش ۱- و رسیدن به سلول هدف پاداش ۱ را برای عامل به همراه دارد. با فرض فاکتور تخفیف و نرخ یادگیری ۰.۹ و اپسیلون ۰.۱ به طور کاهشی برای ۲۰۰ اپیزود الگوریتم‌های TD، $SARSA$ و $Q\_Learning$ به کار گرفته شده است. توزیع تابع ارزش به دست آمده بعد از ۲۰۰ مرحله بکارگیری الگوریتم TD در بخش بعد ارائه شده و همچنین سیاست بهینه، منحنی تعداد برخورد با مانع و دریافت پاداش مسیر و پارامتر متوسط قدم ها در دو الگوریتم $SARSA$ و $Q\_Learning$ با یکدیگر مقایسه شده است.

۲- نتایج:

در این بخش نتایج بدست آمده از بکارگیری هریک از الگوریتم‌ها ارائه و بررسی شده است.

۲-۱) TD

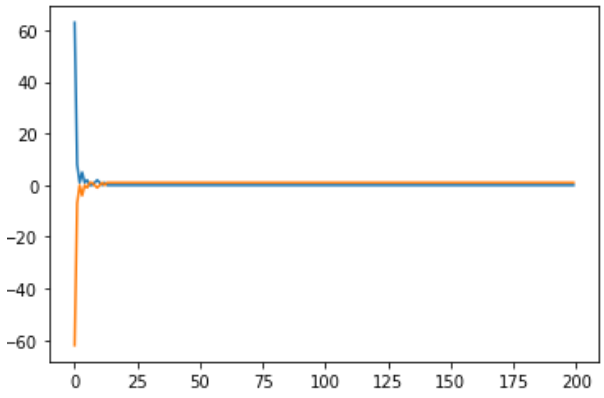
نتایج بدست آمده با بهبود تابع ارزش حالت در این الگوریتم در اپیزود ۲۰۰ در پیوست-۱ قابل مشاهده می‌باشد. از ماتریس محاسبه شده برای ارزش حالت سلول‌ها نمی‌توان سیاست بهینه را دریافت کرد. گرچه در بسیاری از موارد ارزش سلول‌ها به درستی تخمین زده شده است اما مسیر مشخصی را برای رسیدن به سلول هدف نمی‌توان استخراج نمود. باید توجه داشت، از آنجایی که بهبود سیاست در الگوریتم TD صورت نمی‌گیرد و هدف تنها محاسبه ارزش حالت‌ها می‌باشد کاهش اپسیلون می‌تواند زمان مورد نیاز برای به کارگیری الگوریتم را به شدت افزایش دهد.

۲-۲) $SARSA$:

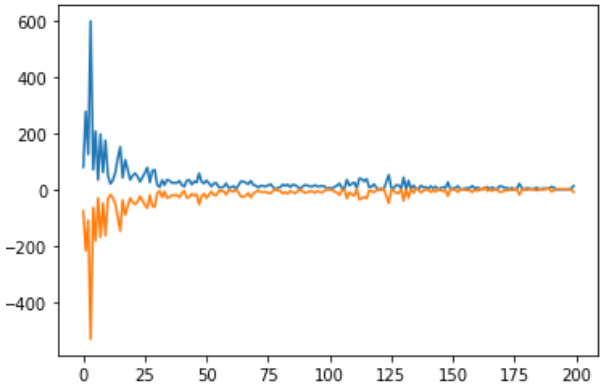
نتایج حاصل از به کارگیری این الگوریتم بعد از محاسبه سیاست بهینه ۱۴ حرکت را برای رسیدن به سلول هدف نشان می‌دهد.

تعداد برخورد با موانع و میزان پاداش دریافتی در حالت به کارگیری این الگوریتم به شدت وابسته به اپسیلون تعریف شده می‌باشد. در صورت شروع به کار الگوریتم با انتخاب عمل تصادفی (ε=۱) تعداد برخوردها با موانع در هر بار اپیزود می‌تواند به ۶۰۰ بار در اپیزودهای ابتدایی برسد در صورتی که این تعداد با استفاده از روش انتخاب عمل نرم (ε=۰.۱) در محدوده اعداد دورقمی خواهد بود.

شکل-۲ تاثیر این پارامتر را بر تعداد برخوردها و پاداش دریافتی نشان می‌دهد.



شکل۱-۲: (ε=۰.۱) و کاهشی برای الگوریتم $SARSA$



شکل۲-۲: (ε=۱) و کاهشی برای الگوریتم $SARSA$

همچنین سیاست بهینه محاسبه شده با بکارگیری این روش در برخی شرایط متفاوت بوده و وابسته به تعداد اپیزودها می‌باشد.

ماتریس زیر یکی از پاسخ های متفاوت و غیر بهینه را در اپیزود ۲۰۰ از بکارگیری روش $SARSA$ نشان میدهد. در این ماتریس عدد ۰ نشان دهنده حرکت به سمت بالا، ۱ به سمت پایین ۲ راست و ۳ حرکت به سمت چپ را نشان می‌دهند.

```
[2 2 2 1 2 1 3 0 0]
[2 0 0 2 1 1 3 0 0]
[0 0 0 1 2 2 1 0 0]
[0 1 0 0 2 0 2 2 0]
[0 0 3 0 0 0 1 1 0]
[0 3 0 3 3 3 2 0 0]
```

همچنین ماتریس زیر سیاست بهینه محاسبه شده توسط این روش را نشان می‌دهد.

```
[2 2 2 1 2 1 3 3 0]
[0 0 1 2 2 2 1 2 0]
[2 1 1 1 1 2 1 3 0]
[2 1 2 2 2 2 2 2 0]
[2 2 2 0 3 1 2 2 0]
[1 2 2 0 2 2 2 0 3]
```

مسیر طی شده توسط عامل در هر دو ماتریس بالا با رنگ آبی نشان داده شده است.

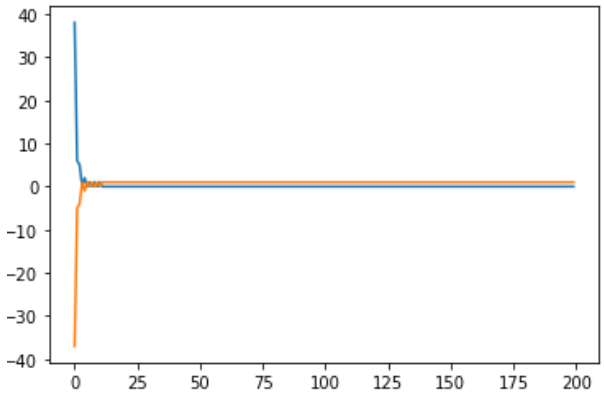
پارامتر متوسط تعداد قدم‌ها نیز وابسته به تعداد اپیزودهای تعریف شده می‌باشد با افزایش تعداد اپیزودها این عدد به سمت ۱۴ میل می‌کند. همچنین با هر بار بکارگیری الگوریتم این عدد با تغییراتی همراه است این درحالی است که برای ۲۰۰ اپیزود میانگین این عدد در محدوده ۵۰ تا ۷۰ قدم می‌باشد.

۳-۲) $Q_{Learning}$:

نتایج بدست آمده از بکارگیری این روش با سرعت بیشتر نسبت به روش ***SARSA*** همراه بوده است. همچنین پارامتر تعداد قدم‌ها برای این االگوریتم مقدار کمتری محاسبه شده است.

در این روش نیز تعداد برخوردها و میانگین پاداش دریافتی وابسته به مقدار دهی پارامتر اپسیلون می‌باشد. اما به طور کلی همگرایی سریع‌تری برای این پارامترها در این روش اتفاق می‌افتد.

شکل-۳ نمونه تعداد برخوردها و میانگین پاداش دریافتی محاسبه شده در صورت بکارگیری این روش را نشان می‌دهد.



شکل۳: (ε=۰.۱) و کاهشی برای الگوریتم $Q_{Learning}$

ماتربس زیر نیز مسیر بهینه محاسبه شده با بکارگیری این روش در اپیزود ۲۰۰ را نشان میدهد.

```
[2 2 2 1 2 1 3 3 0]
[0 1 3 2 1 2 1 1 0]
[1 1 2 0 1 2 1 0 0]
[2 1 3 2 2 2 2 2 0]
[2 2 2 0 0 2 0 0 0]
[2 0 2 2 2 2 2 0 0]
```

اگر چه تفاوت جزئی بین مسیر بهینه در دو الگوریتم ذکر شده دیده می‌شود اما تعداد گام‌های لازم در هر دو الگوریتم یکسان می‌باشد.

۳- منابع:

[1] Richard S. Sutton and Andrew G. Barto, Reinforcement Learning: An Introduction second edition, *The MIT Press Cambridge, Massachusett*s London, England, 2018, ISBN 9780262039246

پیوست-۱

```
[ 0.          0.          0.          0.          0.          0.          0.          0.          0.          0.          0.        ]
[ 0.         -0.28319869 -0.18332967 -0.23001257 -0.12581249 -0.25344964 -0.06069362 -0.07199138  0.          0.          0.        ]
[ 0.         -0.20175446 -0.21955955  0.         -0.099483   -0.09795122 -0.06696994 -0.24655515  0.          0.81623316  0.        ]
[ 0.         -0.1410339  -0.25794849  0.         -0.19192841 -0.12111436 -0.05954094 -0.08412311  0.          0.40602711  0.        ]
[ 0.         -0.10068021 -0.16373253  0.         -0.14903529 -0.05550763 -0.12882337 -0.10942048 -0.16623479  0.10521467  0.        ]
[ 0.         -0.2746358  -0.18064954 -0.19599788 -0.13634403 -0.1501261   0.         -0.240264   -0.23907951 -0.0864808   0.        ]
[ 0.         -0.10938179 -0.27908994 -0.14841519 -0.11573584 -0.25833726 -0.3500481  -0.09777192 -0.25183952 -0.2098383   0.        ]
[ 0.          0.          0.          0.          0.          0.          0.          0.          0.          0.          0.        ]
```

ماتریس توابع حالت با استفاده از روش TD