حل ماز به کمک Q learning

محسن لیاقت ۶۱۰۳۹۸۱۶۳

۱۴ بهمن ۱۴۰۱

فهرست مطالب

۱ تعاریف

state : هر state شامل موقعیت مکانی agent و موقعیت پرچم های دیده نشده است.

actions : بالا، پایین، چپ، راست

rewards :

رفتن به goal state : + تعداد کل پرچم ها + ۱

رسیدن به پرچمی که دیده نشده : +۱

در غیر این صورت : -۱

goal state : موقعیت agent با هدف برابر باشد و هیچ پرچم دیده نشده ای وجود نداشته باشد.

۲ تعداد حالت ها

تعداد حالت ها حداکثر برابر است با:

$$2^{\text{number of flags}} * (\text{number of agent possible positions} - \text{number of flags}) + \text{number of flags} * 2^{\text{number of flags}-1}$$

$$= 2^{\text{number of flags}-1}(2 * \text{number of agent possible positions} - \text{number of flags})$$

زیرا این مقدار برابر است با کاردینالیتی ضرب دکارتی powerset موقعیت پرچم ها در مجموعه همه موقعیت های ممکن برای agent به جز حالت هایی که agent روی یک پرچم باشد و آن پرچم را ندیده باشد.
برای کم تر کردن تعداد حالت ها می توان پیش از اجرای الگوریتم تمام حالت هایی که برای agent قابل دسترس نیستند را شناسایی و حذف کرد. برای این کار نیز کافی است یک الگوریتم پیمایش گراف را اجرا کنیم و هر استیت که با شروع از استیت آغاز قابل دسترسی نبود را حذف کنیم.

۳ گراف حالت ها

باتوجه به نتیجه محاسبات در قسمت پیش می توان فهمید که تعداد استیت ها بسیار زیاد است چیزی در حدود ۹۰۰۰ حالت پس عملا نمایش گراف مربوط به آنها ممکن نمی باشد.

۴ بررسی مقادیر γ و α

با توجه به روش محاسبه Q که به صورت زیر است می توان فهمید که

- α درواقع تعیین کننده میزان یادگیری agent از دفعات قبلی اجرا است درواقع $\alpha = 1$ نمایانگر این است که agent از فقط از پیمایش کنونی خود یاد می گیرد و $\alpha = 0$ به معنای عدم یادگیری agent است. α مقادیر بین صفر و ۱ را اختیار می کند و هر چه به صفر نزدیک تر باشد یادگیری agent کمتر است.

- γ نیز مقادیر بین صفر و یک را اختیار می کند و مشخص کننده میزان توجه agent به بهترین آینده ممکن در هر state را نشان می دهد. درواقع هر چه γ به یک نزدیک تر باشد agent به بهترین آینده ممکن بیشتر توجه دارد.

توجه شود که با توجه به بالا بودن تعداد state ها نمی توان اپیزود ها را رسم کرد ولی Q table مربوط به ۹ حالت خواسته شده در صورت سوال ضمیمه شده است و سیاست حرکت از هر استیت با رنگ آبی مشخص شده است. (هر جدول حاصل ۱۰۰۰۰ اپیزود است.)

۵ یک مثال دیگر

به عنوان یک مثال دیگر ماز زیر را به همراه $\alpha = 0.5 = \gamma$ به برنامه دادیم Q table خروجی را در فایلی به نام another example.pdf ذخیره کردیم.

B	W	W	W	W	W	B	B	W	A
W	W	W	W	B	B	W	W	W	F
W	W	W	W	W	W	W	W	W	T
B	B	W	W	W	W	W	B	B	W
W	F	W	W	W	B	F	W	W	W
B	W	W	W	W	W	W	W	W	W
W	W	B	W	B	B	B	W	B	W
B	F	W	W	F	B	W	W	W	B
W	W	W	B	B	W	W	W	B	W
W	W	B	W	W	W	W	W	W	W