Unyleya
EDUCACIONAL

APRENDIZAGEM POR REFORÇO INVERSO DE TRAGETÓRIAS AMOSTRADAS

APRENDIZAGEM POR REFORÇO INVERSO DE TRAGETÓRIAS AMOSTRADAS

Russell e Norvig (2010) descreveram algoritmos de IRL para casos em que, em vez de uma política ótima total, podemos apenas amostrar trajetórias, a partir de uma política ótima.

São conhecidos os estados, ações e recompensas geradas por uma política para um número finito de episódios, mas não a política em si.

Esta situação é mais comum em casos aplicados, especialmente aqueles que lidam com dados de humanos especialistas

Nesta formulação do problema, substitui-se o vetor de recompensa que é usado para espaços de estados finitos com uma aproximação linear da função de recompensa, que usa um conjunto de funções para obter vetores de recursos com valor real (s).

APRENDIZAGEM POR REFORÇO INVERSO DE TRAGETÓRIAS AMOSTRADAS

Recursos capturam informações importantes de um espaço de estados de alta dimensão (por exemplo, ao invés de armazenar a localização de um carro durante cada etapa de tempo, podemos armazenar sua velocidade média como um recurso).

Para cada recurso $\phi_i(s)$ e peso a_i temos:

$$R(s) = \alpha_1\phi_1(s) + \alpha_2\phi_2(s) + \ldots + \alpha_d\phi_d(s)$$

Onde o objetivo é encontrar os valores mais adequados para cada peso característico a_i.

APRENDIZAGEM POR REFORÇO INVERSO DE TRAGETÓRIAS AMOSTRADAS

A ideia por trás do IRL com trajetórias amostradas é melhorar de forma iterativa uma função de recompensa comparando o valor da política especializada aproximadamente ótima com um conjunto de políticas k geradas.

O algoritmo é inicializado gerando pesos aleatoriamente para a função de recompensa estimada e inicializando o conjunto de políticas candidatas com uma política gerada aleatoriamente.

APRENDIZAGEM POR REFORÇO INVERSO DE TRAGETÓRIAS AMOSTRADAS

Principais passos para o algoritmo são:

1.Estimar o valor da política ótima para o estado inicial $v^{\pi}(s_0)$, bem como o valor de cada política gerada $v^{\pi_i}(s_0)$tomando a recompensa acumulada média de muitos ensaios aleatoriamente amostrados;

2. Gerar uma estimativa da função de recompensa R resolvendo um problema de programação linear. Especificamente, define-se para maximizar a diferença entre a política ótima e cada um das outras k políticas gerada;

3. após um grande número de iterações, finalizar o algoritmo nessa etapa;

APRENDIZAGEM POR REFORÇO INVERSO DE TRAGETÓRIAS AMOSTRADAS

Principais passos para o algoritmo são:

4. Caso contrário, usar um algoritmo de aprendizagem por reforço padrão para encontrar a política ideal para R. Essa política pode ser diferente da política ótima dada, já que nossa função de recompensa estimada não é necessariamente idêntica à função de recompensa estimada que estamos procurando;

5. Adicionar a política recém-gerada ao conjunto de políticas k candidatas e repita o procedimento.

APRENDIZAGEM POR REFORÇO INVERSO DE TRAGETÓRIAS AMOSTRADAS

Referências

RUSSELL, S.; NORVIG, P. **Artificial Intelligence:** A Modern Approach. 3.ed. New Jersey: Pearson Education, 2010.

Obrigada!

hulianeufrn@gmail.com