Unyleya
EDUCACIONAL

APRENDIZAGEM POR REFORÇO PASSIVA

APRENDIZAGEM POR REFORÇO PASSIVA

Conceito

Aprendizagem por reforço passiva é quando o ambiente passa de um estado para outro **sem intervenção do agente**, onde a cada transição o agente apenas percebe o estado e recebe reforço, atualizando sua representação do valor desse estado.

Faz uso de um conceito baseado em estados, num ambiente completamente observável. A política do agente é fixa, no estado *s* ele executa a ação (s).

A tarefa aprendizagem passiva é semelhante à de avaliação de política. A **diferença principal** é que o agente de aprendizagem passiva não conhece o modelo de transição $P(s' \mid s, a)$ descrito no algoritmo de iteração de política (visto na seção 2.1.4 da Unidade II), que especifica a probabilidade de alcançar o estado s' a partir do estado s e depois realizar a ação a; **ele também não conhece** a função de recompensa $R(s)$, que especifica a recompensa para cada estado (RUSSELL e NORVIG, 2010).

APRENDIZAGEM POR REFORÇO PASSIVA

Conceito

O agente executa um conjunto de experiências no ambiente usando sua política . Em cada experiência, o agente começa num estado inicial e experimenta uma sequência de transições de estados até alcançar um dos estados terminais. Suas percepções fornecem tanto o estado atual quanto a recompensa recebida nesse estado.

A seguinte equação descreve a aprendizagem por reforço passiva:

$$U^{\pi}(s) = \mathrm{E}\,\Sigma_{t=0}^{\infty}\,\gamma^{t}R(S_t)$$

APRENDIZAGEM POR REFORÇO PASSIVA

Objetivo

Utilizar as informações sobre recompensas para aprender a utilidade esperada associada a cada estado não terminal *s*. A utilidade é definida como a soma esperada de recompensas (descontadas) obtidas se a política π for seguida.

A recompensa $R(s)$ para o estado S_t (uma variável aleatória) é o estado alcançado no tempo t quando é executada a política π e $S_0 = s$, e a variável γ representa o fator de desconto.

Métodos: **Estimativa de Utilidade Direta**, **Programação Dinâmica Adaptativa** e **Diferença Temporal**.

APRENDIZAGEM POR REFORÇO PASSIVA

Estimativa de utilidade direta

Ainda estamos operando sob um ambiente estocástico, portanto, uma ação específica executada em um estado específico nem sempre leva ao mesmo estado seguinte.

Se quisermos aprender as utilidades desses estados sob uma política fixa, poderemos imaginar uma maneira bastante direta de fazê-lo:

- execute a política várias vezes;
- no final de cada execução, calcule a utilidade para cada estado na sequência (lembre-se, a utilidade de um estado é a soma de recompensas para esse estado e todos os estados subsequentes;
- atualize a utilidade média para cada um dos estados que observamos com nossos novos pontos de dados.

APRENDIZAGEM POR REFORÇO PASSIVA

Estimativa de utilidade direta

Ideia: utilidade de cada estado é a recompensa total que se espera a contar desse estado em diante (conhecido como recompensa a obter), e a cada teste propicia uma amostra dessa quantidade para cada estado visitado (RUSSELL e NORVIG, 2010).

Resultados menos prováveis acontecerão com **menos frequência**. Afetarão menos nossas estimativas, o que significa que não precisamos conhecer um modelo de transição para que isso funcione.

Acabará convergindo para as verdadeiras utilidades, mas é lento, porque não tira proveito do processo de decisão de Markov.

APRENDIZAGEM POR REFORÇO PASSIVA

Estimativa de utilidade direta

Lembrete: uma vez que calculamos a utilidade de um estado como recompensa, a utilidade de cada estado pode ser escrita estritamente em termos das utilidades de seus vizinhos imediatos.

Os valores de utilidade obedecem às equações de *Bellman* para uma política fixa:

$$U^{\pi}(S) = R(S) + \gamma \Sigma_{s'} P(S' \mid S, \pi(S)) U^{\pi}(S')$$

Pode-se visualizar a estimativa de utilidade direta como a busca em um espaço de hipóteses U muito maior do que precisa ser no sentido de incluir muitas funções que violam as equações de *Bellman* (RUSSELL e NORVIG, 2010).

APRENDIZAGEM POR REFORÇO PASSIVA

Referências

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

Obrigada!

hulianeufrn@gmail.com

Unyleya
EDUCACIONAL

EXEMPLO DE APRENDIZAGEM POR REFORÇO PASSIVA

EXEMPLO DE APRENDIZAGEM POR REFORÇO PASSIVA

Exemplo – Ambiente 4 × 3

Para manter a simplicidade, começaremos com o caso de um agente de aprendizagem passiva que utiliza uma representação baseada em estados em um ambiente completamente observável.

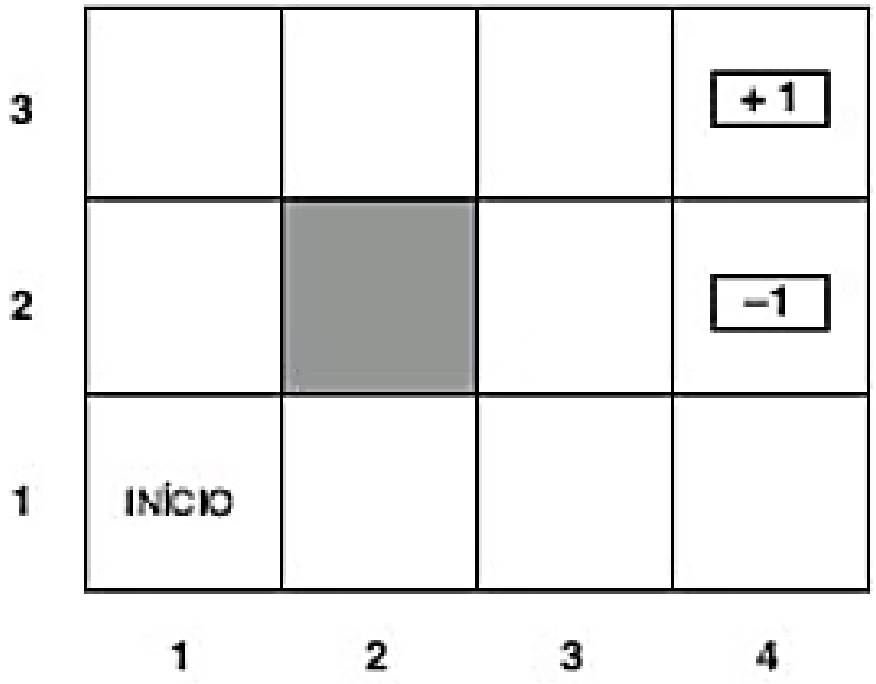
Na aprendizagem passiva, a política π do agente é fixa: no estado s, ele sempre executa a ação $\pi(s)$.

Meta: simplesmente aprender o quanto a política é boa, ou seja, aprender a função utilidade $U^{\pi}(s)$.

EXEMPLO DE APRENDIZAGEM POR REFORÇO PASSIVA

Exemplo – Ambiente 4 × 3

Figura 1: Um ambiente simples de 4 × 3 que apresenta ao agente um problema de decisão sequencial.

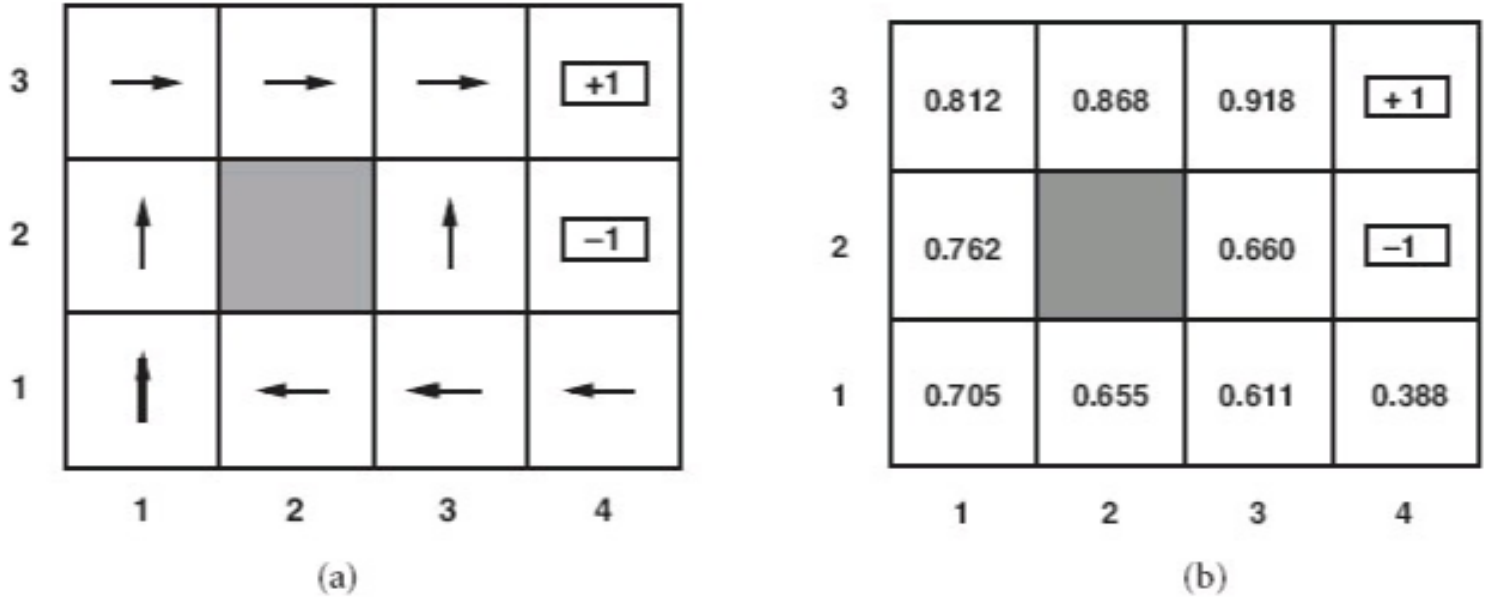


Fonte: Russell e Norvig (2010)

EXEMPLO DE APRENDIZAGEM POR REFORÇO PASSIVA

Exemplo – Ambiente 4 × 3

Figura 2: (a) Uma política π para o mundo 4 × 3; essa política é ótima com recompensas iguais a R(s) = –0,04 nos estados não terminais e sem desconto. (b) Utilidades dos estados no mundo 4 × 3, dada a política π.



Fonte: Russell e Norvig (2010)

EXEMPLO DE APRENDIZAGEM POR REFORÇO PASSIVA

Exemplo – Ambiente 4 × 3

O agente executa um conjunto de **experiências** no ambiente usando sua política π. Em cada experiência, o agente começa no estado (1,1) e experimenta uma sequência de transições de estados até alcançar um dos estados terminais, (4,2) ou (4,3). Suas **percepções** fornecem tanto o **estado atua**l quanto a **recompensa recebida** nesse estado. Experimentos típicos seriam:

$$(1,1)_{-0,04} \rightsquigarrow (1,2)_{-0,04} \rightsquigarrow (1,3)_{-0,04} \rightsquigarrow (1,2)_{-0,04} \rightsquigarrow (1,3)_{-0,04} \rightsquigarrow (2,3)_{-0,04} \rightsquigarrow (3,3)_{-0,04} \rightsquigarrow (4,3)_{+1}$$
$$(1,1)_{-0,04} \rightsquigarrow (1,2)_{-0,04} \rightsquigarrow (1,3)_{-0,04} \rightsquigarrow (2,3)_{-0,04} \rightsquigarrow (3,3)_{-0,04} \rightsquigarrow (3,2)_{-0,04} \rightsquigarrow (3,3)_{-0,04} \rightsquigarrow (4,3)_{+1}$$
$$(1,1)_{-0,04} \rightsquigarrow (2,1)_{-0,04} \rightsquigarrow (3,1)_{-0,04} \rightsquigarrow (3,2)_{-0,04} \rightsquigarrow (4,2)_{-1}.$$

Observe que cada percepção de estado tem como subscrito a recompensa recebida. O objetivo é utilizar as informações sobre recompensas para aprender a utilidade esperada $U^{\pi}(s)$ associada a cada estado não terminal s.

EXEMPLO DE APRENDIZAGEM POR REFORÇO PASSIVA

Exemplo – Ambiente 4 × 3

A utilidade é definida como a soma esperada de recompensas(descontadas) obtidas se a política π é seguida.

$$U^{\pi}(s) = \mathrm{E}\,\Sigma_{t=0}^{\infty}\,\gamma^{t}R(S_t)$$

Onde R(s) é a recompensa para o estado, S_t(uma variável aleatória) é o estado alcançado no tempo t quando é executada a política π e S_0 = s. Incluiremos um fator de desconto γ em todas as nossas equações, mas, para o mundo 4 × 3, definiremos γ = 1.

EXEMPLO DE APRENDIZAGEM POR REFORÇO PASSIVA

Referências

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

Obrigada!

hulianeufrn@gmail.com

Unyleya
EDUCACIONAL

APRENDIZAGEM POR DIFERENÇA TEMPORAL

APRENDIZAGEM POR DIFERENÇA TEMPORAL

Algoritmos de aprendizado por diferença temporal – DT (SUTTON, 1984)

Aprendem **novas estimativas** do valor com **base** em **outras estimativas**.

Não exige um modelo exato do sistema. Procura estimar valores de utilidade para cada estado do ambiente por recompensas oriundas das transições e de valores de estados sucessivos.

A **aprendizagem** ocorre **diretamente** a partir da **experiência**. Não é necessário um modelo completo do ambiente. Atualizar as estimativas da função valor a partir de outras estimativas já aprendidas em estados sucessivos. Não é preciso alcançar o estado final de um episódio antes da sua atualização.

Avaliação de uma política é encarada como um problema de predição. Estima-se a função valor-estado sob a política.

APRENDIZAGEM POR DIFERENÇA TEMPORAL

Vantagens

Em relação aos métodos apresentados anteriormente:

- não exige o modelo MDP do ambiente (não exige conhecimento prévio do modelo de transição do ambiente);
- pode ser implementado de forma totalmente incremental para aplicações *on-line* (a sua atualização considera apenas o estado seguinte);
- tem garantida a convergência assintótica para a resposta correta (embora as atualizações da função valor não sejam obtidas a partir dos dados reais, mas de valores aproximados);
- os métodos DT são mais rápidos na sua convergência para tarefas estocásticas (TSITSIKLIS, 1994).

APRENDIZAGEM POR DIFERENÇA TEMPORAL

Transições

Usa as transições obtidas para ajustar os valores dos estados observados, de forma a esses concordarem com as equações de restrições (que definem o ambiente).

Numa transição de estado $s \rightarrow s'$, é aplicada a seguinte atualização à utilidade:

$$U^{\pi}(s) = U^{\pi}(s) + \alpha(R(s) + \gamma U^{\pi}(s') - U^{\pi}(s))$$

A taxa de aprendizagem α (assume valores no intervalo [0, 1]) determina a velocidade com que o agente assimila a informação, apresentando-se menor à medida que $\alpha \rightarrow 0$.

Se α não for um parâmetro fixo, então $U^{*}(s)$ convergirá para o valor correto. A atualização somente envolve o sucessor observado s' (e não todos os estados seguintes possíveis), tendo implícito um cálculo pouco exigente em termos computacionais.

APRENDIZAGEM POR DIFERENÇA TEMPORAL

Algoritmo AGENTE-DT-PASSIVO (RUSSELL e NORVIG, 2010)

Figura 1: Agente de aprendizagem por reforço passivo baseado em DT.

função AGENTE-DT-PASSIVO (*percepção*) **retorna** uma ação

entradas: *percepção*, um percepção indicando o estado atual s' e o sinal de recompensa r'

variáveis estáticas: π, uma política fixa

U, uma tabela de utilidades, inicialmente vazia

N_s, uma tabela de frequências para estados, inicialmente zero

s, a, r estado, ação e recompensas anteriores, inicialmente nulos

se s' é novo **então faça** $U[s'] \leftarrow r'$

se s é não nulo **então faça**

incrementar $N_s[s]$

$U[s] \leftarrow U[s] + a(N_s[s])\ (r + \gamma\ U[s'] - U[s])$

se *TERMINAL?* $[s']$ **então** $s, a, r \leftarrow$ nulo **senão** $s, a, r \leftarrow s', p[s'], r'$

retornar a

Fonte: Russell e Norvig (2010)

APRENDIZAGEM POR DIFERENÇA TEMPORAL

Estratégias para aprendizagem por reforço utilizando DT (KAELBLING; LITTMAN; MOORE, 1996)

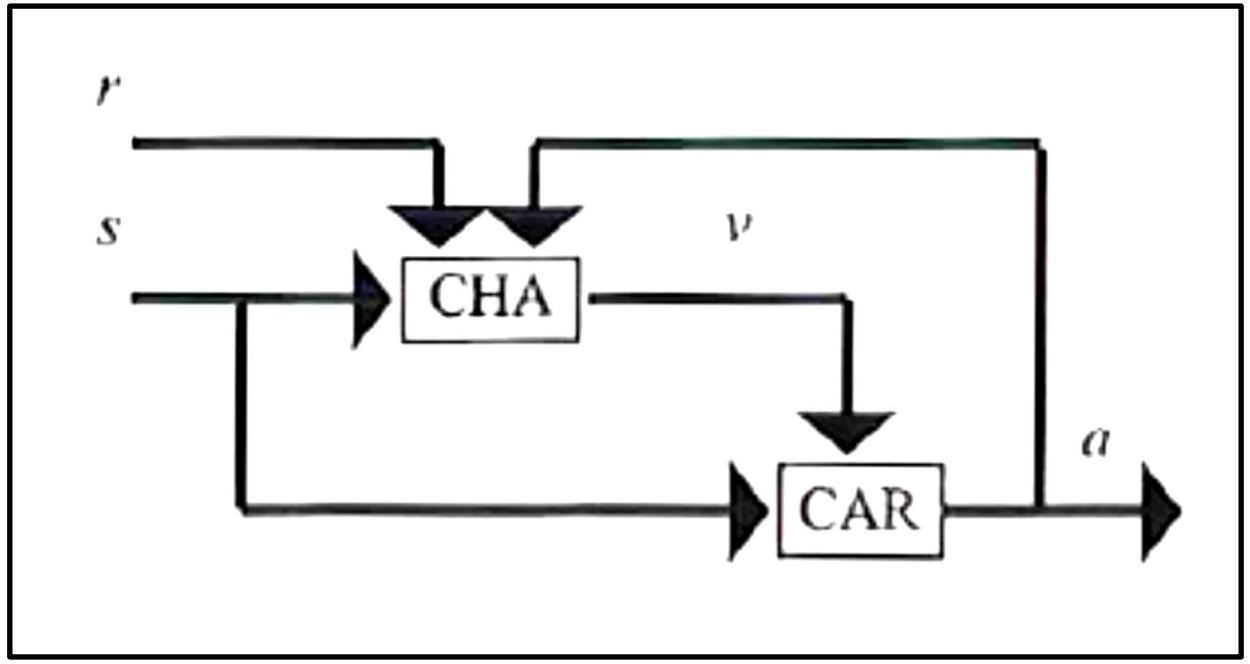
Crítico Heurístico Adaptativo (CHA) (BARTO, SUTTON e ANDERSON, 1983): consiste de dois componentes, um **crítico adaptativo** (CHA) e um de **aprendizagem por reforço** (CAR).

O componente de aprendizado por reforço busca agir de forma a maximizar o valor heurístico *v*, que é calculado pelo crítico. O crítico usa o sinal real externo de reforço *r* para aprender a avaliar os estados *s*. Na maioria das implementações, os componentes CHA e CAR operam simultaneamente e apenas para essas implementações há garantia, sob condições adequadas, de convergência para uma política ótima.

APRENDIZAGEM POR DIFERENÇA TEMPORAL

Algoritmo AGENTE-PDA-PASSIVO (RUSSELL e NORVIG, 2010)

Figura 1: Arquitetura para o crítico heurístico adaptativo.



Fonte: Kaelbling, Littman e Moore (1996).)

APRENDIZAGEM POR DIFERENÇA TEMPORAL

Estratégias para aprendizagem por reforço utilizando DT (KAELBLING; LITTMAN; MOORE, 1996)

***Q-Learning*:** o algoritmo de aprendizagem por reforço mais estudado para os propósitos de controle, possuindo provas de convergência.

Q-Learning estima a avaliação do par estado-ação diretamente a partir de sinais externos de reforço utilizando uma política gulosa.

Veremos esse algoritmo em detalhes na Unidade IV.

APRENDIZAGEM POR DIFERENÇA TEMPORAL

Referências

BARTO, A. G; SUTTON, R. S.; ANDERSON, C. W. Neuronlike adaptative elements that can solve difficult learning control problems. **IEEE Transactions on Systems, Man and Cybernetics**, V. 3, N. 5, pp: 834-846, 1983.

KAELBLING, L. P; LITTMAN, M. L.; MOORE, A. W. Reinforcement learning: a survey. **Journal of Artificial Intelligence Research**, V. 4, pp: 237-285, 1996.

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

SUTTON, R. S. **Temporal credit assignment in reinforcement learning**. Tese de doutorado, University of Massachusetts Amherst, 1984.

TSITSIKLIS, J. N. **Asynchronous stochastic approximation and Q-learning**. Boston: Kluwer Academic Publishers, 1994.

Obrigada!

hulianeufrn@gmail.com

Unyleya
EDUCACIONAL

COMPARATIVO ENTRE AS ABORDAGENS PDA E DT

COMPARATIVO ENTRE AS ABORDAGENS PDA E DT

Abordagem PDA x DT

Ambas estão na realidade estreitamente relacionadas. Tentam fazer ajustes locais para as estimativas de utilidade, a fim de fazer cada estado “concordar” com seus sucessores.

Uma diferença é que DT ajusta um estado para concordar com seu sucessor observado na equação abaixo:

$$U^{\pi}(s) = U^{\pi}(s) + \alpha(R(s) + \gamma U^{\pi}(s') - U^{\pi}(s))$$

Enquanto PDA ajusta o estado para concordar com todos os sucessores que poderiam ocorrer, ponderados por suas probabilidades, conforme equação abaixo:

$$U^{\pi}(S) = R(S) + \gamma \Sigma_{s'} P(S' \mid S, \pi(S)) U^{\pi}(S')$$

COMPARATIVO ENTRE AS ABORDAGENS PDA E DT

Abordagem PDA x DT

Uma diferença mais importante é que, enquanto DT faz um único ajuste por transição observada, PDA faz tantos quantos necessita para restaurar a consistência entre as estimativas de utilidade U e o modelo de ambiente P.

Embora a transição observada faça apenas uma mudança local em P, seus efeitos talvez tenham de ser propagados ao longo de U. Desse modo, DT pode ser visualizada como uma primeira aproximação, crua mas eficiente, para PDA.

COMPARATIVO ENTRE AS ABORDAGENS PDA E DT

Abordagem PDA x DT

É possível estender a abordagem de DT para usar um modelo de ambiente que gere várias pseudoexperiências — transições que o agente de DT talvez imagine que poderiam acontecer, dado seu modelo atual.

Para cada transição observada, o agente de DT pode gerar grande número de transições imaginárias. Desse modo, as estimativas de utilidade resultantes se aproximarão cada vez mais das estimativas de PDA — é claro, a um preço de um tempo maior de computação.

COMPARATIVO ENTRE AS ABORDAGENS PDA E DT

Abordagem PDA x DT

De modo semelhante, podemos gerar versões mais eficientes de PDA pela aproximação direta dos algoritmos de iteração de valor ou iteração de política. Mesmo que o algoritmo de iteração de valor seja eficiente, ele é intratável se tivermos, digamos, 10100 estados.

No entanto, muitos dos ajustes necessários para os valores de estado em cada iteração serão extremamente pequenos.

Uma abordagem possível para gerar respostas de qualidade razoável com rapidez é **limitar o número de ajustes feitos depois de cada transição observada**. Também poderíamos utilizar uma heurística para ordenar os ajustes possíveis de modo a executar apenas os mais significativos.

COMPARATIVO ENTRE AS ABORDAGENS PDA E DT

Abordagem PDA x DT

A heurística de **varredura priorizada** prefere fazer ajustes em estados cujos **prováveis** sucessores acabaram de sofrer um **grande** ajuste em suas próprias estimativas de utilidade.

Usando heurísticas como essa, os algoritmos de PDA aproximada em geral podem aprender quase tão rápido quanto a PDA completa, em termos do número de sequências de treinamento, mas podem ser várias ordens de magnitude mais eficientes em termos de computação.

Isso lhes permite manipular espaços de estados que são muito maiores para a PDA completa. Os algoritmos de PDA aproximada têm uma vantagem adicional: nas fases iniciais de aprendizagem de um novo ambiente, o modelo de ambiente P com frequência estará longe de ser correto e, assim, haverá pouca razão para calcular uma função utilidade exata que corresponda a esse modelo

Obrigada!

hulianeufrn@gmail.com

Unyleya
EDUCACIONAL

APRENDIZAGEM POR REFORÇO ATIVA

APRENDIZAGEM POR REFORÇO ATIVA

Conceito

O **agente** tenta **localizar** uma **boa política de ação**. A cada transição, o agente compreende o estado e obtém o reforço, atualizando sua interpretação do valor nesse estado e escolhendo uma ação a ser executada que mudará o estado do ambiente.

Um agente deve escolher quais ações deve executar. **Diferentemente** da **aprendizagem passiva** onde é a política fixa que estabelece sua conduta. Torna um agente de reforço ativo mais poderoso, mas adiciona complicações quando tentamos formalizar seu processo em um algoritmo.

Não só é preciso escolher a ação em cada iteração, mas também é necessário garantir que nossas escolhas garantam que possamos encontrar uma política ótima.

APRENDIZAGEM POR REFORÇO ATIVA

Conceito

O agente precisa aprender um modelo completo com probabilidades de resultados para todas as ações, em vez de aprender apenas o modelo para a política fixa.

O mecanismo de aprendizagem do algoritmo AGENTE-PDA-PASSIVO funciona bem para isso, desde que se leve em conta o fato de que o agente tem uma escolha de ações.

As utilidades que o agente precisa aprender são definidas pela política ótima, obedecendo-as à equação de *Bellman* (RUSSELL e NORVIG, 2010).

APRENDIZAGEM POR REFORÇO ATIVA

Cenário

Considerando esta equação no contexto de tentar aprender uma política ótima, pode fazer sentido escolher a melhor ação que conhecemos cada vez, e usar o resultado dessa ação para atualizar a utilidade do estado s.

Exemplo: se sua refeição favorita desde criança é frango e macarrão com queijo, então comer esta refeição quando você está com fome é a escolha que maximiza sua recompensa (ela é sua comida favorita!). Porém, se você sempre escolhe esta ação (que maximiza a sua recompensa), não tem a chance de descobrir melhores opções de comida, caso elas existam. Então, para encontrar uma política ótima, temos que arriscar alguns resultados ruins pela possibilidade de que algumas ações possam levar a melhores resultados do que já conhecemos. Isso nos dá o problema de como equilibrar a **exploitação** (fazendo escolhas que sabemos que vai nos ajudar) contra a **exploração** (descobrir quão valiosas são as escolhas que ainda não fizemos).

APRENDIZAGEM POR REFORÇO ATIVA

Aprendizagem ativa x aprendizagem passiva

A grande diferença dessas aprendizagens é a **exploração**, em que um agente deve experimentar tanto quanto possível do seu ambiente, com o objetivo de aprender a se comportar nele.

APRENDIZAGEM POR REFORÇO ATIVA

Referências

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

Obrigada!

hulianeufrn@gmail.com

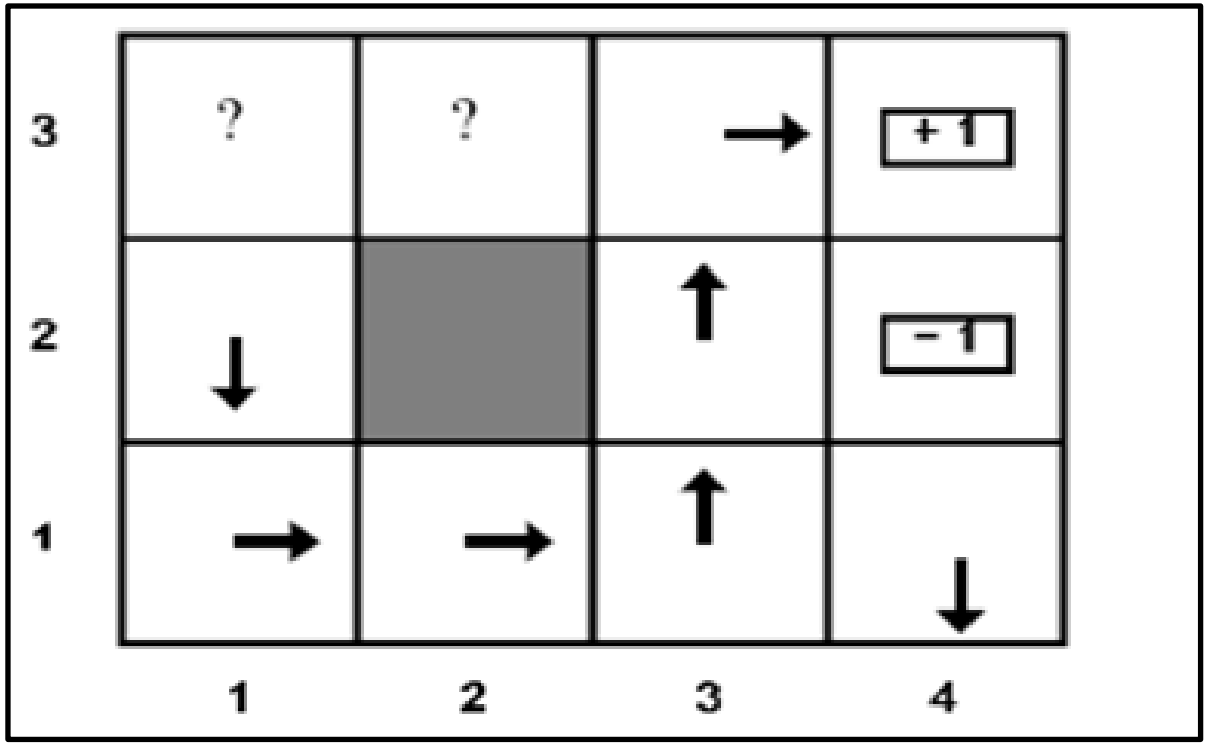
Unyleya
EDUCACIONAL

EXPLORAÇÃO

EXPLORAÇÃO

Um agente de PDA e o modo como ele deve ser modificado para decidir quais ações executar

Figura 1: Política não ótima para a qual o agente guloso converge nessa sequência.



Fonte: Russell e Norvig (2010)

Um agente PDA que segue a recomendação da política ótima para o modelo aprendido em cada etapa.

Podemos verificar que o agente não aprende as utilidades verdadeiras ou a política ótima verdadeira. Ao contrário, ele encontra uma política que alcança a recompensa +1 ao longo da rota inferior.

EXPLORAÇÃO

Exploração na aprendizagem ativa

Após experiências com pequenas variações, o agente se fixa nessa política. Dessa forma, ele nunca aprende as utilidades dos outros estados e assim, nunca encontra a rota ótima via (1, 2), (1, 3) e (2, 3).

A esse agente é dado o nome de **agente guloso**. Raramente converge para a política ótima correspondente a esse ambiente. As vezes converge para políticas realmente ruins.

As ações fornecem recompensas de acordo com o modelo atual aprendido. Fazem bem mais que isso, elas também contribuem para a aprendizagem do modelo verdadeiro afetando as percepções que são recebidas.

Faz com que haja um melhoramento do modelo, fazendo com que o agente receba recompensas melhores no futuro.

EXPLORAÇÃO

Exploração na aprendizagem ativa

Para isso ocorrer: um agente deve assumir um compromisso entre **exploitação** (aproveitamento) para maximizar a sua recompensa (que reflete em suas estimativas de utilidades atuais) e **exploração**, a fim de maximizar seu bem-estar a longo prazo.

Usar apenas a **exploitação**, corre o risco de ficar paralisado num único lugar. A exploração sozinha para melhorar o conhecimento é inútil, se não colocar em prática o conhecimento.

Os agentes devem então, explorar no começo e aproveitar no final.

EXPLORAÇÃO

Problema de decisão de um agente

Pode ser visto como o problema de selecionar uma determinada ação quando ele está em um determinado estado.

Exemplo conhecido de um problema de decisão envolvendo um único agente é uma variação do problema do "**bandido de *k* braços**" ou *k-armed bandit* (SUTTON e BARTO, 1998), nomeado em analogia a uma hipotética máquina de caça níqueis com *k* alavancas.

Esse problema consiste em um agente que pode escolher dentre *k* opções de ação, cada uma com uma probabilidade diferente e imutável de ser a escolha correta.

Cada vez que o agente escolhe a opção correta, esta recebe um sinal de recompensa positivo, nos demais casos recebe um sinal de recompensa negativo.

EXPLORAÇÃO

Problema de decisão de um agente

Conforme o agente explora estas opções ao longo de múltiplas iterações, tende a adequar sua política de tomada de decisões para favorecer a opção com maior probabilidade de resultar em recompensa positiva.

Problemas como esse são extremamente difíceis de resolver com exatidão para se obter um método de exploração ótimo.

No entanto, eles apresentam um esquema razoável que eventualmente leva a um comportamento ótimo do agente.

EXPLORAÇÃO

Ações gulosas

Se mantivermos estimativas dos valores das ações, a qualquer momento haverá pelo menos uma ação cujo valor estimado seja maior.

Chamamos essas ações de **ações gulosas**. Quando selecionamos uma dessas ações, dizemos que estamos explorando seu conhecimento atual dos valores das ações. Ao contrário, selecionarmos uma ação não gulosa dizemos que estamos explorando, porque isso permite o aumento da estimativa de valor da ação não gulosa.

Exploitação é o certo a se fazer para maximizar a recompensa esperada em um único passo.

A exploração pode produzir a maior recompensa total a longo prazo.

EXPLORAÇÃO

Exemplo

Suponha que o valor de uma ação gulosa seja conhecido com certeza, enquanto várias outras ações são estimadas como quase tão boas, mas com uma incerteza substancial.

A incerteza é tal que pelo menos uma dessas ações provavelmente é melhor do que a ação gulosa, mas não se sabe qual.

Se tivermos muitos passos à frente para fazer seleções de ações, talvez seja melhor explorar as ações não gulosas e descobrir quais delas são melhores do que a ação gulosa.

A recompensa é menor a curto prazo, durante a exploração. Maior a longo prazo, porque depois de ter descoberto as melhores ações, podemos explorá-las mais vezes. Como não é possível realizar exploração e exploitação com uma única seleção de ação, geralmente nos referimos ao conflito entre exploração e exploitação (SUTTON e BARTO, 2018).

EXPLORAÇÃO

Função de exploração

Escolher se é melhor explorar ou exploitar depende, de maneira complexa, dos valores precisos das estimativas, das incertezas e do número de etapas restantes.

Existem muitos métodos sofisticados para equilibrar exploração e exploitação.

A estratégia mais fácil para atingir esse equilíbrio é simplesmente tomar uma ação aleatória algumas vezes.

Uma maneira um pouco sofisticada é definir uma **função de exploração** $f(u, n)$, dessa forma, influenciará no valor percebido das ações que determinam sua entrada.

Podemos apenas escolher a melhor ação; mas agora, os estados menos visitados parecerão melhores do que realmente são.

EXPLORAÇÃO

Função de exploração

A função de exploração $f(u, n)$ toma como entradas a utilidade atual conhecida de um estado (u) e o número de vezes que esse estado foi visitado (n). A função de exploração precisa estar aumentando em u e diminuindo em n. Como no exemplo:

$$f(u, n) = u, se\ n\ \geq N_e$$
$$R^+\ de\ outra\ forma$$

Onde N_e é o número de vezes que se deseja explorar um estado antes que esteja razoavelmente certo que este é o seu valor, e R^+ é uma estimativa otimista do valor dos estados explorados.

EXPLORAÇÃO

Referências

SUTTON, R. S.; BARTO, A. G**. Reinforcement learning:** an introduction, 2ªEd. MIT Press, Cambridge, Massachusetts, EUA, 2018.

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

Obrigada!

hulianeufrn@gmail.com

Unyleya
EDUCACIONAL

APRENDIZAGEM DE UMA FUNÇÃO DE AÇÃO-VALOR

APRENDIZAGEM DE UMA FUNÇÃO DE AÇÃO-VALOR

Função valor x função de ação-valor

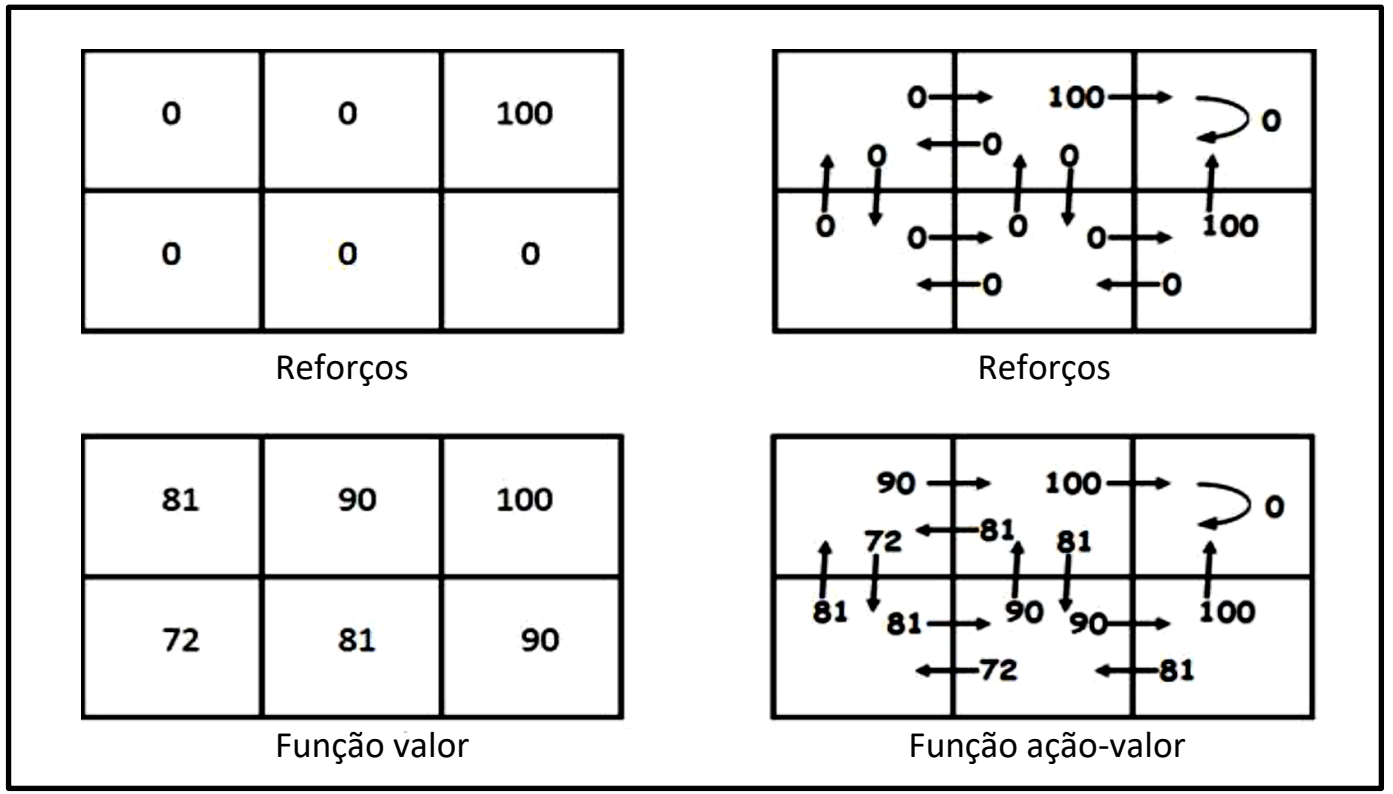
Função valor: associa o valor diretamente a estados do ambiente (visto na Unidade II). O valor de uma ação indiretamente definido como valor do estado aceitável apenas para agente com modelo efetivo do ambiente (prevendo o estado resultante da execução de cada ação). Aplicável apenas para agente com modelo efetivo do ambiente (para prever o estado resultado da execução de cada ação

Função de ação-valor: associa o valor diretamente a pares (estado, ação). Aplicável ao agente sem modelo efetivo do ambiente, quando não há nenhum modelo de ambiente acessível e com o modelo apenas perceptivo de ambiente inacessível.

APRENDIZAGEM DE UMA FUNÇÃO DE AÇÃO-VALOR

Algoritmo AGENTE-DT-PASSIVO (RUSSELL e NORVIG, 2010)

Figura 1: Comparação função-valor e função ação-valor.



Fonte: Robin (2002)

APRENDIZAGEM DE UMA FUNÇÃO DE AÇÃO-VALOR

Aprendizagem de uma função de ação-valor

Anteriormente, construímos um agente de PDA ativo, agora vamos considerar como construir um agente de aprendizagem ativo de diferença temporal (DT).

A principal diferença em relação ao caso passivo é que o agente não está mais equipado com uma política fixa.

Se for aprender uma função utilidade U, necessitará compreender um modelo apto a escolher uma ação que se baseia em U por meio da observação prévia de um passo (RUSSELL e NORVIG, 2010).

APRENDIZAGEM DE UMA FUNÇÃO DE AÇÃO-VALOR

Aprendizagem de uma função de ação-valor

O problema de aquisição de modelo para o agente DT é idêntico ao do agente de PDA, sendo a mesma equação de atualização do DT utilizada para este novo agente. Possível pela seguinte razão:

Considere que o agente realize um passo que comumente leve a um bom destino, mas devido ao não determinismo do ambiente, o agente acaba em um estado catastrófico.

A regra de atualização DT levará a circunstância tão a sério de forma que seu resultado se torna o resultado normal da ação.

O resultado improvável ocorrerá com pouca frequência em um grande conjunto de sequências de treinamento.

APRENDIZAGEM DE UMA FUNÇÃO DE AÇÃO-VALOR

Aprendizagem de uma função de ação-valor

A **longo prazo** seus **efeitos** serão **ponderados** de forma proporcional a sua probabilidade.

Mostra que o algoritmo de DT convergirá para os mesmos valores que a PDA, à medida que o número de sequências de treinamento tender a infinito (RUSSELL e NORVIG, 2010).

A **aplicação** do **método da diferença temporal** na **função-valor da ação permite** que o **ambiente** seja **desconhecido**, e leva a um dos métodos mais difundidos da aprendizagem por reforço, o *Q-learning* (será visto em detalhes na Unidade IV).

APRENDIZAGEM DE UMA FUNÇÃO DE AÇÃO-VALOR

Referências

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

ROBIN, J. **Aprendizagem por Reforço**, 2002. Disponível em: <www.cin.ufpe. br/~compint/aulas-IAS/ias/ias-021/rl.ppt>. Acesso em 05/09/2020.

Obrigada!

hulianeufrn@gmail.com