Aprendizagem por Reforço Clássica e Conexionista: análise de aplicações

Thais L. Silva, Maury M. Gouvêa Jr.

Instituto Politécnico – Pontifícia Universidade Católica de Minas Gerais
Av. Dom José Gaspar, 500 – 31.535-901 – Belo Horizonte – MG – Brazil

thaisl.silva@hotmail.com, maury@pucminas.br

Resumo. *A aprendizagem por reforço é um paradigma do tipo não supervisionado, onde o agente aprende sem um professor – sem exemplos rotulados. Assim, pode-se aprender em tempo real, pela experiência de executar uma ação e avaliar seu sucesso. Neste artigo apresenta-se dois exemplos de aprendizagem por reforço: um baseado no modelo clássico, utilizando o algoritmo* Q-Learning*, o outro baseado em sistema conexionista, utilizando uma rede neural artificial. No último caso, a aplicação é voltada à Tecnologia Assistiva. Em ambos os exemplos, mostra-se que o agente aprendeu com suas próprias ações, melhorando seu desempenho com o tempo.*

1. Introdução

Dentre os paradigmas de aprendizagem de máquina, a aprendizagem por reforço (AR) se destaca na robótica por não necessitar de exemplos de treinamento, isto é, o agente aprende em tempo real, com seus próprios erros, interagindo com seu meio. Essa característica da AR ganha destaque em situações em que não se conhece em detalhes o ambiente no qual o agente está inserido, sem qualquer histórico de comportamento a ser seguido. Alguns exemplos desses ambientes são as situações de desmoronamento, grande profundidades oceânicas e, até mesmo, exploração espacial.

A aprendizagem por reforço pode ser clássica ou conexionista. O modelo clássico usa algoritmos, como o Q-Learning [Sutton e Barto 1991]. A aprendizagem por reforço conexionista é baseada em redes neurais artificiais [Haykin 2001]. Na aprendizagem por reforço, o agente executa uma ação que é avaliada por um crítico, que dá ao agente um sinal de reforço positivo ou negativo que é utilizado para se auto ajustar e, assim, aperfeiçoar suas ações. Apesar de haver uma pré-programação, os algoritmo ou modelos analíticos possuem parâmetros cujos valores ótimos são muito difíceis de ajustar previamente, sem que situações reais aconteçam. Assim, uma possibilidade promissora de se ajustar esses parâmetros é a AR, que usará as ações pré-definidas para avaliar o desempenho dos parâmetros do sistema que sofrerão adaptações em função das avaliações do crítico.

2. Aprendizagem por Reforço

A aprendizagem por reforço (AR) é um paradigma que não necessita de histórico de ocorrências, de exemplos de comportamento ou padrões. Por isso, a sua aplicação é voltada a tarefas que necessitam de adaptação em tempo real, como exploração de ambientes desconhecidos ou dinâmicos. A aprendizagem por reforço pode ser clássica [Sutton e Barto 1991] ou conexionista neurais [Hertz et al. 1991]. Em qualquer dos métodos, clássico ou conexionista, na aprendizagem por reforço o agente interage com o ambiente, recebendo um sinal de reforço que varia conforme seu desempenho nas ações executadas. Com ações boas ou ruins, o sinal de reforço é usado pelo agente para se adaptar ao meio.

2.1. AR Clássica

A aprendizagem por reforço permite a um determinado agente aprender interagindo com o ambiente sem a presença de um tutor. Dessa forma, o agente decide qual ação tomar buscando uma política ótima por meio de recompensas. A Figura 1 mostra que em cada passo da interação o **agente de aprendizagem** observa o estado, no instante *k*, do **ambiente**, *s*(*k*), e escolhe uma determinada ação, *a*(*k*). O agente realizada essa ação, modifica-se o estado do ambiente, *s*(*k*+1), que retorna ao agente, por meio de um **crítico**, um sinal de reforço, *r*(*s*, *a*), que pode ser uma recompensa ou penalização.

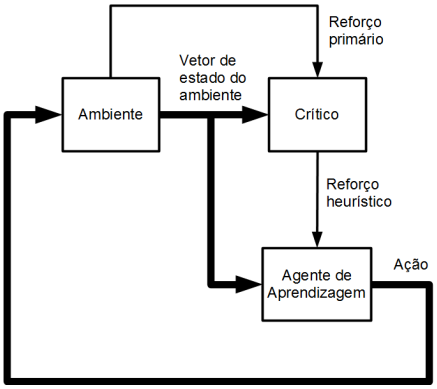

Figure 13. Ciclo percepção-ação da AR

Por meio de tentativa e erro o agente aprende em relação ao ambiente, criando um mapa estado-ação. O objetivo do aprendizado é definir qual ação tomar a cada iteração que maximize o valor da recompensa [Sutton e Barto 1991].

2.2. AR Conexionista

A aprendizagem por reforço conexionista é baseada em redes neurais. O agente, na Figura 1, é constituído de uma rede neural que aprende em tempo real [Hertz et al. 1991]. O agente percebe o ambiente e toma uma decisão. O crítico analisa a ação e a considera como a desejada ou não dependendo da qualidade da ação. O sistema de aprendizagem utiliza um algoritmo de aprendizagem supervisionada para efetuar os ajustes dos pesos da rede neural.

Na AR Conexionista, em vez de um conjunto de exemplos de treinamento, existe apenas o sinal de reforço, *r*, que pode ser 1, para um sinal de reforço positivo, ou –1, para um sinal de reforço negativo. Assim, os exemplos de treinamento são construídos em tempo real, pela regra

$$d_j = \begin{cases} S_j & \text{se } r_j = 1 \\ -S_j & \text{se } r_j = -1 \end{cases} \qquad (1)$$

sendo d_j a *j*-ésima saída desejada fruto da ação do agente analisada pelo crítico. Essas regras sugerem que a rede neural será mais propensa a executar ações que foram recompensadas, e vice-versa.

Para construir as regras de treinamento, compara-se d_j com o valor médio da saída, $\langle S_j \rangle$, isto é, $\breve{\mathbf{u}}_j = d_j - \langle S_j \rangle$. Os pesos da rede neural serão atualizados, como segue

$$\varphi w_{ki} = \eta\, r_j\, \breve{\mathbf{u}}_j\, y_j \qquad (2)$$

sendo φw_{ki} o *i*-ésimo peso do neurônio *k* e η a taxa de aprendizagem.

3. Aplicações de Aprendizagem por Reforço Conexionista

Esta seção apresenta dois exemplos de aprendizagem por reforço, a primeira conexionista e a segunda clássica, usando o método Q-Learning. As próximas subseções apresentam os dois métodos de aprendizagem.

3.1. Aplicações de AR Conexionista

Este exemplo de AR Conexionista apresenta uma bengala eletrônica para deficientes visuais que detecta obstáculos remotamente, sem contato físico, emitindo uma vibração que alerta o usuário. Um sistema de aprendizagem *on-line* adapta o sistema de alerta ao padrão de comportamento do usuário. Um método de aprendizagem conexionista adapta a distância de alerta em função da velocidade do usuário. De uma forma suave, a distância de alerta diminui para usuário mais lentos ou aumenta para usuários mais rápidos.

Uma rede neural artificial com uma saída executa a ação de diminuir ou aumentar a distância de alerta conforme a velocidade do usuário e as sucessivas detecções de obstáculos. Os resultados experimentais mostraram que o módulo de aprendizagem respondeu de forma esperada e satisfatória aos testes realizados, como mostra a Figura 2.

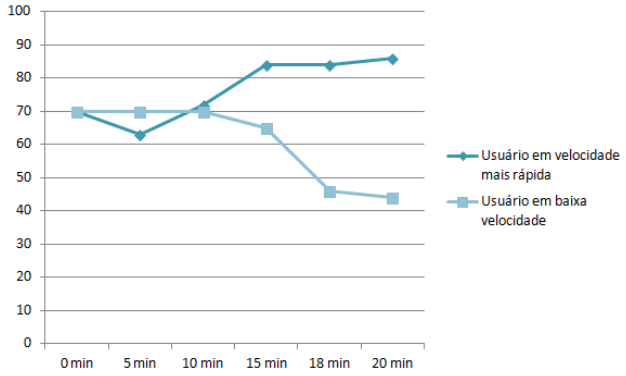

Figura 2. Distância de alerta média depois de 20 minutos de uso da BIN

3.2. Aplicações de AR Clássica

O exemplo de AR clássica, baseada no algoritmo *Q-Learning*, apresenta um robô humanoide que aprende a andar para fim didático. Os motores do humanoide podem assumir três velocidades: 0, 30 e 50 unidades de potência. Os motores são acionados de forma intercalada, por 1 segundo. As ações possíveis são não alterar a potência do motor, diminuir ou aumentar a potência. Quando o robô humanoide identifica um obstáculo, simulando uma presa, o robô começa a andar. O crítico avalia a ação em função do deslocamento do robô em relação ao alvo por um período de tempo. Se o robô se desloca até um limite x_1, a ação é punida; se o deslocamento é de x_1 até x_2, a ação é punida com menor severidade; finalmente, se o robô se desloca uma distância maior que x_2, a ação é recompensada. A ideia é que se a ação não for boa o suficiente, ela não produzirá um deslocamento do robô em relação ao alvo e, assim, será punida.

A tabela *Q-Learning* possui dimensão N_e x N_a, sendo N_e o número de estados e N_a o número de ações. No exemplo apresentado, há 2 motores assumindo 3 estados e cada motor pode executar 3 ações. Assim, $N_e = 3^2 = 9$ e $N_a = 3^2 = 9$. A Tabela 1 mostra a tabela *Q-Learning* antes de depois do treinamento *on-line*, isto é, durante o tempo em que o humanoide aprende a andar. Sabe-se previamente, pelo fim didático, que o $Q(s,a)$ ótimo é velocidades (50,50), pois são máximas e não desestabilizam o humanoide e ação (0,0), pois não alteram a velocidade ótima. Trata-se, portanto, da posição $Q(9,5)$. A Tabela 1, que tem seus valores iniciais aleatório no intervalo [0,1], tem sua posição $Q(9,5)$ maximizada em

relação aos demais valores depois de aproximadamente 20 minutos de treinamento por reforço, o que demonstra o sucesso da abordagem.

Table 1. Tabela *Q-Learning* antes e depois do treinamento

Matriz antes de aprender								
0,95	0,01	0,48	0,36	0,49	0,79	0,4	0,18	0,79
0,07	0,71	0,11	0,56	0,7	0,66	0,64	0,88	0,17
0,62	0,37	0,88	0,69	0,95	0,03	0,08	0,14	0,45
0,35	0,14	0,39	0,89	0,75	0,57	0,44	0,78	0,4
0,79	0,12	0,92	0,91	0,43	0,44	0,14	0,01	0,02
0,25	0,91	0,75	0,39	0,08	0,01	0,85	0,65	0,69
0,3	0,34	0,73	0,77	0,61	0,98	0,44	0,46	0,62
0,19	0,02	0,28	0,3	0,9	0,63	0,65	0,17	1
0,17	0,59	0,03	0,01	0,84	0,23	0,75	0,16	0,21
Matriz depois de aprender								
0,48	-0,7	0,42	-0,65	-0,57	0,023	-0,183	0,03	0,32
-0,01	0,46	-0,92	-0,66	0,37	0,6	0,55	0,07	-0,26
-0,69	0,3	0,83	0,22	0,42	-0,21	0,07	-0,2	-0,47
-0,67	-0,535	-0,48	0,37	0,01	0,14	-0,03	0,56	0,19
0,03	-0,78	0,52	0,8	-0,88	-0,78	-0,59	-0,14	0,1
-0,528	0,3	0,11	-0,88	-0,53	-0,03	0,52	0,07	0,23
-0,84	0,13	0,67	0,76	0,523	0,07	-0,52	-0,69	0,012
0,1	-0,33	-0,304	0,28	0,56	0,26	0,64	-0,34	0,1
-0,36	0,21	-0,13	-0,2	8,424	-0,46	0,24	-0,391	-0,368

4. Conclusão

Este artigo apresentou o desenvolvimento de aplicações com aprendizagem por reforço. Dois exemplos foram apresentados, o primeiro baseado no modelo clássico, utilizando o algoritmo *Q-Learning*, e o segundo baseado em rede neural artificial. Ambos os exemplos foram casos de sucesso, onde o agente aprendeu a executar sua tarefa em tempo real.

Referências

Sutton, R. S.; Barto, A. G. (1991) "Reinforcement learning: An introduction". Massachusetts: MIT Press.

Haykin, S. (2001) Redes Neurais: princípios e práticas. Porto Alegre: Bookman.

Heertz, J., Krogh, A. and Palmer, R.G. (1991) Introduction to the Theory of Neural Computing. Redwood City: Addison-Wesley Publishing Co.

Alves, F. A. S., Neumann, A. M. M., Gouvêa Jr., M. M. (2014) "Bengala Inteligente Neural Baseada em Aprendizagem por Reforço para Deficientes Visuais", Encontro Nacional de Inteligência Artificial e Computacional, São Carlos.