UTILIZANDO TRANSFERÊNCIA DE CONHECIMENTO PARA ACELERAR O APRENDIZADO POR REFORÇO

Luiz A. Celiberto Jr*, Murilo F. Martins†, Reinaldo A. C. Bianchi‡, Jackson P. Matsuura†

*Departamento de Sistemas e Controle - Instituto Tecnológico da Aeronáutica
Pça Mal-do-Ar Eduardo Gomes, 50 São José dos Campos -SP-BR

†Departamento de Engenharia Elétrica e Eletrônica - Imperial College - Londres,UK

‡Departamento de Engenharia Elétrica - Centro Universitário da FEI
Av. Humberto de Alencar Castelo Branco, 3972 São Bernardo do Campo-SP-BR

Emails: celibertojr@gmail.com, mfmartin@imperial.ac.uk, rbianchi@fei.edu.br, jackson@ita.br

Abstract— Reinforcement Learning is a very well known technique for the solution of problems when the agent needs to act with success in an unknown environment through trial and error. However, this technique is not efficient enough to be used in applications with real world demands due to the time that the agent needs learn. This paper investigates the use of a Transfer Learning (TL) between agents to speed up the well known Reinforcement Learning algorithm in the Littman domain. The experiences were made comparing the use with Reinforcement Learning algorithm, Q-learning, with heuristic accelerated version generated by Transfer Learning of the same algorithm. The results show that the use of a Transfer Learning can lead to a significant improvement in the performance of the agent.

Keywords— Reinforcement Learning, Transfer Learning.

Resumo— O Aprendizado por Reforço é uma técnica muita conhecida para a solução de problemas quando o agente precisa atuar com sucesso em um local desconhecido por meio de tentativa e erro. Porém, esta técnica não é eficiente o bastante para ser usada em aplicações com exigências do mundo real, devido ao tempo que o agente leva para aprender. Este artigo investiga a utilização do uso de transferência de conhecimento entre agentes para acelerar o algoritmo de aprendizado por reforço no domínio desenvolvido por Littman.
Foram realizadas experiências comparando o uso de um algoritmo de aprendizado por reforço bem conhecido, o Q-learning, com uma versão acelerada por heurísticas do mesmo algoritmo geradas pela transferência de conhecimento. Os resultados mostram que o uso de transferência de conhecimento pode conduzir a uma significante melhora na atuação do agente.

Palavras-chave— Aprendizado por Reforço, Transferência de Conhecimento.

1 Introdução

Aprendizado por Reforço (AR) é uma técnica muito atraente quando se deseja solucionar uma variedade de problemas de controle e planejamento quando não existem modelos disponíveis *a priori*, pois o agente irá aprender a cumprir uma função de maneira correta em um ambiente desconhecido por meio de tentativa e erro.

No aprendizado por reforço, o agente aprende por meio da interação direta entre o agente e o ambiente através de recompensas. Estas recompensas são dadas por meio de reforços positivos e negativos, que são usadas para sinalizar ao agente se ele esta tomando as ações corretas ou não. O objetivo do agente sempre será acumular o máximo reforço positivo. Porém, aprender não é suficiente: o agente precisa aprender rapidamente, pois o ambiente em que ele esta localizado pode sofrer constantes mudanças. Infelizmente, o aprendizado por reforço é muito lento, necessitando de enorme quantidade de ações para que um agente aprenda corretamente: a convergência dos algoritmos de AR só pode ser atingida após uma extensiva exploração do espaço de estados-ações.

Bianchi (Bianchi, 2004) mostrou que a velocidade de convergência de um algoritmo de AR pode ser acelerada ao se utilizar funções heurísticas para guiar a exploração do espaço de estados-ações. Ele também propôs diversos algoritmos acelerados por heurísticas, entre eles, o Q-Learning Acelerado por Heurísticas, baseado no conhecido algoritmo Q-Learning (Watkins, 1989).

O uso de técnicas de transferência de conhecimento tem recebido grande atenção recentemente no aprendizado de agentes(Konidaris and Barto, 2007; Taylor and Stone, 2005; Asgharbeygi et al., 2006; Ferns et al., 2006), fornecendo assim um meio mais rápido para o aprendizado dos mesmos (Asgharbeygi et al., 2006).

Este trabalho tem como objetivo investigar o uso de heurísticas por meio da transferência de conhecimento adquirido por um agente durante seu treinamento para outro agente que não foi previamente treinado. Para esta experiência, inicialmente, um jogador utilizando o Q-Learning fará uma partida contra outro jogador utilizando Minimax-Q (Littman, 1994). Quando concluído

todos os treinamentos a tabela de aprendizado do Q-Learning é armazenada que será utilizada como conhecimento adquirido e será a heurística para um novo jogo. Neste novo jogo o algoritmo Minimax-Q fará novas partidas com o algoritmo Q-Learning Acelerado por Heurísticas (HAQL) e depois seus treinamentos serão comparados.

Este artigo é organizado da seguinte maneira: na próxima seção são apresentadas, de forma sucinta, características do algoritmo de aprendizado por reforço e a seção 3 introduz a aceleração do aprendizado por reforço. A seção 4 descreve a proposta da experiência e os resultados e a última seção apresenta a conclusão deste trabalho.

2 Aprendizado por Reforço

No Aprendizado por Reforço, um agente sem conhecimentos prévios aprende por meio de interações com o ambiente, recebendo recompensas por suas ações e assim descobrindo a política ótima para a resolução de um determinado problema. A suposição principal do Aprendizado por Reforço é a existência de um agente que pode aprender a escolher suas ações que resultarão em um melhor resultado futuro na realização de uma tarefa (Pegoraro, 2001).

O aprendizado por reforço é uma técnica de aprendizado não supervisionado devido a não existência de uma representação de pares de entrada e de saída. Para cada movimentação do agente não é fornecida nenhum tipo de informação externa que ajude seu deslocamento, tirando aquela que ele mesmo percebe da sua interação com o ambiente (Kaelbling et al., 1996).

O Aprendizado por Reforço funciona da seguinte maneira: em um ambiente, a cada intervalo de tempo o agente executa uma ação a_t. Esta ação é determinada pela política já aprendida e faz o agente ir para o estado s_{t+1} e tendo em vista a recompensa $r_{s,a}$ que irá ganhar. A recompensa pode ser dada por valores positivos ou negativos, indicando se o agente esta seguindo corretamente para o objetivo ou não. A Fig. 1 apresenta um esboço do funcionamento do aprendizado por reforço.

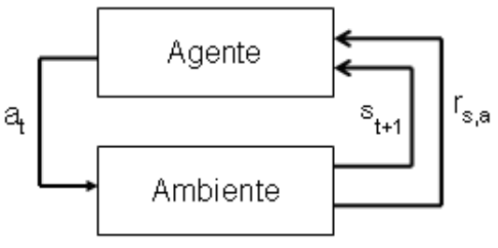

Figura 1: Aprendizado por Reforço.

Entre os diversos algoritmos de Aprendizado por Reforços existentes, o mais conhecido é o Q-Learning (Watkins, 1989; Watkins and Dayan, 1992) considerado por vários autores um algoritmo de fácil implementação (Kaelbling et al., 1996) e por este motivo, utilizado em uma grande quantidade de domínios (Mitchell, 1997). No Q-Learning, para cada ação realizada pelo agente é computado sua recompensa e o valor esperado ao seguir a melhor política com um desconto. Esta política é aprendida por meio da interação com o ambiente e, assim, aprendidos quais as melhores ações para chegar a um objetivo. A informação da política é armazenada em uma matriz Q(s,a), que guarda os valores estimados para cada par de estado,ação (Uther and Veloso, 1997).

O algoritmo Q-Learning foi proposto como uma maneira de aprender iterativamente a política ótima π quando o modelo do sistema não é conhecido (Watkins, 1989). O algoritmo Q-Learning propõe que o agente aprenda uma função Q de recompensa esperada com desconto, conhecida como função valor-ação. Ele aproxima iterativamente $\hat{Q}$ a estimativa de Q^*(s, a) no instante t utilizando a seguinte regra de aprendizado:

$$\hat{Q}(s,a) \leftarrow \hat{Q}(s,a) + \alpha \left[r + \gamma \max_{a'} \hat{Q}(s',a') - \hat{Q}(s,a) \right] \quad (1)$$

onde: s é o estado atual; a é a ação realizada em s; r é o reforço recebido após realizar a em s ; s' é o novo estado; γ é o fator de desconto ($0 \leq \gamma < 1$); α é a taxa de aprendizagem ($0 < \alpha < 1$).

Para a escolha das ações é utilizada uma estratégia conhecida como exploração aleatória ε-Greedy. O agente irá executar a ação que contenha o maior valor Q e probabilidade 1-ε e escolhe uma ação aleatória com probabilidade ε. A regra de transição de estados será dada pela seguinte equação:

$$\pi = \begin{cases} a_{random} & \text{se } q \leq \varepsilon, \\ \arg\max_a Q(s,a) & \text{caso contrario,} \end{cases} \quad (2)$$

Sendo q uma variável randômica com probabilidade [0,1] e ε ($0 \leq$ p ≤ 1) um parâmetro que define a exploração/explotação.

Apesar de ter sido aplicado em uma vasta gama de problemas de maneira bem sucedida, uma desvantagem do Aprendizado por Reforço é o tempo que o agente leva para aprender. Em muitos casos são necessárias uma enorme quantidade de interações (ou episódios) para o agente poder aprender sobre o ambiente onde ele esta ou sobre a tarefa que deve realizar.

Devido a esta demora, é comum o uso de alguma forma de aceleração do aprendizado. A seção a seguir apresenta uma técnica que pode ser usadas para diminuir o tempo de aprendizado dos algoritmos de AR.

3 Aceleração do Aprendizado por Reforço

Um dos métodos de acelerar o algoritmo Q-Learning é aplicar uma heurística que acelere o aprendizado. Heurísticas são mêtodos ou principio de decisões, que indicam as melhores alternativas de ações que levem a solução de um problema mais facilmente (Pearl, 1984).

Utilizar uma heurística em um aprendizado por reforço é fornecer uma ajuda ao agente para que ele possa conseguir chegar a um objetivo mais facilmente. Bianchi (Bianchi, 2004) define a heurística como uma técnica que no caso médio melhora a eficiência do algoritmo. Em seu trabalho, a heurística é usada para que o agente seja influenciado a escolher as ações durante o aprendizado.

Assim é obtida uma heurística definida por H(s,a) no AR que irá indicar a importãncia de executar em um estado st uma ação at. A política estará fortemente vinculada a sua função heurística. Sendo assim podemos dizer que teremos uma função heurística que é definida por uma "Política Heurística" (Bianchi, 2004).

Conhecendo as informações existentes no domínio ou em estágios iniciais, pode ser definida uma heurística que poderá ser usada para acelerar o aprendizado, porém devido as características do AR o uso de uma heurística inadequada pode causar um atraso no sistema, mas não impede o algoritmo de convergir para uma política ótima (Bianchi, 2004).

Bianchi (Bianchi, 2004) também propôs diversos algoritmos acelerados por heurísticas, entre eles, o Q-Learning Acelerado por Heurísticas (HAQL), baseado no conhecido algoritmo Q-Learning. O HAQL utilizada uma modificação da regra E-Greedy para a regra de transição de estados, e incorpora a função heurística como uma somatória simples ao valor da função valor-ação. A regra de transição de estados é mostrada abaixo:

$$\pi = \begin{cases} \arg\max_a [\hat{Q}(s,a) + \xi H(s,a)] & \text{se } q \leq \varepsilon, \\ a_{random} & \text{caso contrario} \end{cases} \quad (3)$$

aonde:

- $\mathcal{H} : \mathcal{S} \times \mathcal{A} \rightarrow \Re$: é uma função heurística que influência a escolha da ação.
- ξ: é uma variável real usada para determinar a influência da heurística.
- q: é uma variável randômica com probabilidade [0,1] e ε ($0 \leq p \leq 1$) é um parâmetro que define a exploração/explotação: quanto maior o valor de ε menor será a probabilidade de uma escolha randômica.
- a_{random} : ação randômica em um estado s

Vale notar que as únicas modifiações em relação ao algoritmo Q-Learning se referem ao uso da função heurística para a escolha da ação a ser executada e a existência de um passo de atualização da função H(s,a). A convergência deste algoritmo é demonstrada por Bianchi (Bianchi, 2004).

4 Experimento e Resultados Obtidos

Este trabalho tem como objetivo investigar o uso de heurísticas por meio da transferência de conhecimento adquirido por um agente durante seu treinamento para outro agente que não foi previamente treinado.

Nas experiências iniciais foi utilizado o domínio desenvolvido por Littman (Littman, 1994), modelado a partir do futebol de robôs que propõem um jogo de Markov com soma zero entre os dois agentes. No domínio utilizado, dois jogadores chamados de A e B competem em um grade de 10x10 células. Neste jogo, cada célula pode ser ocupada por um dos jogadores, que podem executar ações de deslocamento (para cima, para baixo, esquerda, direita) ou ficar imóvel.

A posse da bola é randomicamente concedida a um agente no começo de cada partida, e esta irá sempre acompanhar os jogadores. Quando é realizada uma ação que leva a bola á célula aonde esta um adversário, o jogador perde a bola. Quando o jogador com a bola, entra na área do gol do adversário, o time marca um gol e uma nova partida (ou *episodio*) é começada. Quando o jogador faz uma ação que leva a bola para fora da grade, o jogador fica imóvel. A Fig. 2 mostra um exemplo do ambiente proposto por Littman.

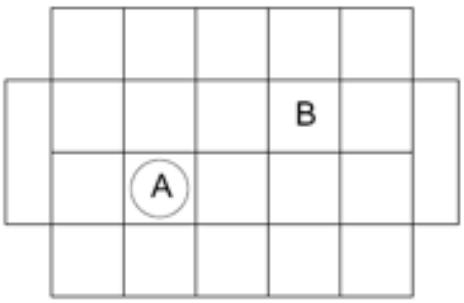

Figura 2: Ambiente proposto por Littman

Para esta experiência, inicialmente, um jogador utilizando o Q-Learning (jogador A) faz uma partida contra outro jogador (jogador B) utilizando Minimax-Q (Littman, 1994). Ambos os algoritmos possuem o valor α de 1 com taxa de decaimento de 0,9999954 para cada ação realizada, taxa de exploraçcão/explotação de 0,2 e fator de desconto γ de 0,9 (Bianchi, 2004; Celiberto et al., 2007). O reforço utilizado foi de +100 quando o agente chega ao gol e -100 por gol mar-

cado pelo adversário. Foram realizados um total de 30 treinamentos, sendo que cada treinamento consiste de 500 episódios. Cada episódios representa 10 tempos de jogos. Cada tempo de jogo termina quando um jogador faz ou recebe um gol ou quando é realizado até 50 interações com a bola.

Quando concluído todos os treinamentos, a tabela de aprendizado do Q-Learning é armazenada. Esta tabela será então utilizada como conhecimento adquirido e será a heurística para um novo jogo. Neste novo jogo o algoritmo Minimax-Q (jogador B) fará novas partidas com jogador A, este formado pelo algoritmo Q-Learning Acelerado por Heurísticas (HAQL).

Para o HAQL, foi utilizado um valor ε de 1 com decaimento de 0.9999 para cada ação realizada. Este valor ε e seu decaimento (retirado por meio de testes empíricos) representa o grau de confiança que o jogador deve ter na heurística enquanto nao tiver ainda conhecimento pleno sobre o domínio. Sendo assim, inicialmente com o valor de ε alto o aprendizado depende fortemente do conhecimento anterior, porém o conhecimento é mais importante no começo do jogo, ajudando o agente a entender o domínio e suas regras.

Na Fig. 3 é possível comparar os resultado do aprendizado somente com Q-Learning (QL), Q-Learning Acelerado por Heurísticas (HAQL) e de somente o uso de heurísticas (H). Os dados demostraram que utilizar o conhecimento anterior de uma partida em formato de heurísticas, resultou em uma aceleração do aprendizado.

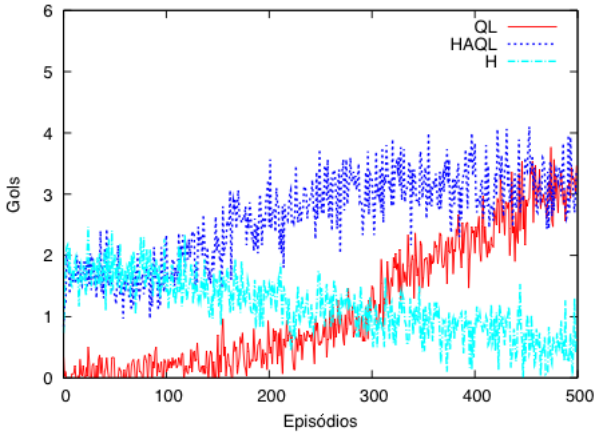

Figura 3: Evolução do saldo de gols para os algoritmos Q-Learning, HAQL e somente heurísticas.

Foi utilizado o teste t de Student (Spiegel, 1984), para verificar se a hipótese de que o uso do algorítmo do HAQL, utilizando como heurísticas conhecimento adquirido, acelera o aprendizado em relação ao algoritmo Q-Learning é válida. O resultado para o teste t de Student computado utilizando os dados da Fig. 3 é apresentado na Fig. 4. Nesta figura é possível ver que os algoritmos são significativamente diferentes, com nível de confiança de 99 %. Isto confirma que a heurísticas usadas tornam o algoritmo HAQL mais rápido que o Q-Learning.

Tabela 1: Total de gols feitos pelos jogadores (média e desvio padrão)

Algoritmo	Jogador	Gols
Q–Leaning	Jogador A	(876,06 ± 145,69)
Minimax-Q	Jogador B	(299,6 ± 48,91)
HAQL	Jogador A	(1807,8 ± 94,03)
Minimax-Q	Jogador B	(496,96 ± 91,42)

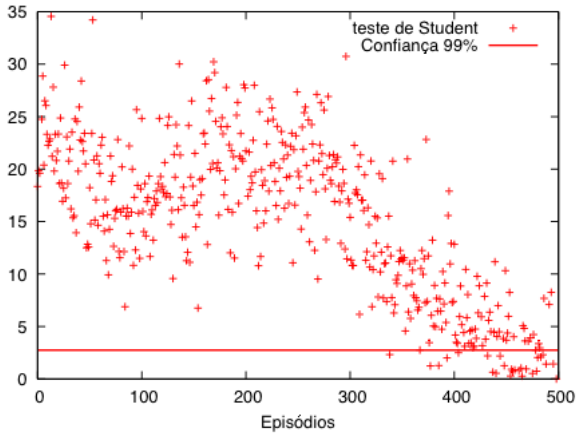

Figura 4: Resultado do teste t de Student para os algoritmos Q-Learning e HAQL

Também foi observado o quanto a heurística ajuda no desenvolvimento da partida, como mostrado na tabela 1, que representa a média dos gols dos jogadores em 500 episódios. É interessante notar que o jogador A após o uso das heurísticas aumenta a sua media de gols em mais de 100%, mas por outro lado o jogador B também aumenta a sua média de gols em mais de 65%, isto deixa amostra claramente que em certas oportunidades a heurística conduziu para uma ação não eficaz do jogador, mostrando que nem todo o conhecimento anterior é válido.

5 Conclusões

Este artigo propôs o uso de transferência de conhecimento entre agentes para acelerar o algoritmo de aprendizado por reforço Q-Learning no domínio desenvolvido por Littman (Littman, 1994).

Os resultados obtidos neste domínio mostram que utilizando o conhecimento adquirido em outros jogos como heurísticas é possível fazer com que um novo jogador aprenda mais rápido que usando apenas o Q-Learning, pois proporciona um menor espaço de busca explorado pelo agente, não precisando assim, o mesmo, ficar passando por todos os estados para aprender o que fazer, obtendo assim um aprendizado com uma performance melhor.

Outro fator importante que se deve destacar é apesar dos jogos serem com algoritmos simila-

res e em mesmo domínio, as partidas não são as mesmas, isto pode ser muito bem observado por meio do gráfico H da Fig. 3, que representa somente o uso das heurísticas para determinar as ações do agente (jogador A), é possível observar que com o uso de somente as heurísticas, estas conseguem marcar alguns gols, porém quando o adversário aprende o jogo, eles não servem mais, pois são fixas e o agente praticamente não consegue mais marcar gol. Pode-se então concluir que as heurísticas aceleram aprendizado, mas não são tão fortes que eliminam a necessidade do mesmo.

A partir dos resultados deste trabalho, os trabalhos futuros vão se concentrar na extração da informações necessárias do conhecimento para através delas gerar regras que possam orientar outros agentes em domínios parecidos ou mesmo até em domínios diferentes. Estas regras podem fornecer informações do que o agente pode ou não fazer com base no que já tenha acontecido com outro agente, mesmo em domínios diferentes, como tambem fazendo que agentes que estão aprendendo mais devagar possam receber regras de agentes que estão aprendendo mais rápido, acelerando globalmente o aprendizado.

Dentre das áreas de contribuição deste trabalho, é possível destacar: Aprendizado de Máquinas ; Sistemas Multi-agentes; Controle de Tráfego Veicular Urbano, Políticas de Investimento; etc.

Referências

Asgharbeygi, N., Stracuzzi, D. J. and Langley, P. (2006). Relational temporal difference learning, *ICML*, ICML, pp. 49–56.

Bianchi, R. A. C. (2004). *Uso de Heurísticas para Aceleração do Aprendizado por Reforço*, PhD thesis, Escola Politécnica da Universidade de São Paulo, São Paulo. Tese de Doutorado em Engenharia Elétrica.

Celiberto, Jr., L. A., Ribeiro, C. H., Costa, A. H. and Bianchi, R. A. (2007). Heuristic reinforcement learning applied to robocup simulation agents, pp. 220–227.

Ferns, N., Castro, P. S., Precup, D. and Panangaden, P. (2006). Methods for computing state similarity in markov decision processes, *In Proceedings of UAI*, AUAI Press.

Kaelbling, L. P., Littman, M. L. and Moore, A. W. (1996). Reinforcement learning: A survey, *Journal of Artificial Intelligence Research* **4**: 237–285.

Konidaris, G. and Barto, A. G. (2007). Building portable options: Skill transfer in reinforcement learning, *IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, January 6-12, 2007*, pp. 895–900.

Littman, M. L. (1994). Markov games as a framework for multi-agent reinforcement learning, *Proceedings of the 11th International Conference on Machine Learning (ML-94)*, Morgan Kaufmann, New Brunswick, NJ, pp. 157–163.

Mitchell, T. (1997). *Machine Learning*, McGraw-Hill Education (ISE Editions).

Pearl, J. (1984). *Heuristics: Intelligent Search Strategies for Computer Problem Solving*, Addison-Wesley.

Pegoraro, R. (2001). *Agilizando Aprendizagem por Reforço em Robótica Móvel Através do Uso de Conhecimento Sobre o Domínio*, PhD thesis, Escola Politécnica da Universidade de São Paulo, São Paulo. Tese de Doutorado em Engenharia Elétrica.

Spiegel, M. R. (1984). *Estatística. 2. ed.*, McGraw-Hill.

Taylor, M. E. and Stone, P. (2005). Behavior transfer for value-function-based reinforcement learning, *in* F. Dignum, V. Dignum, S. Koenig, S. Kraus, M. P. Singh and M. Wooldridge (eds), *The Fourth International Joint Conference on Autonomous Agents and Multiagent Systems*, ACM Press, New York, NY, pp. 53–59.

Uther, W. and Veloso, M. (1997). Adversarial reinforcement learning, *Technical report*, In Proceedings of the AAAI Fall Symposium on Model Directed Autonomous Systems.

Watkins, C. J. C. H. (1989). *Learning from Delayed Rewards*, PhD thesis, University of Cambridge.

Watkins, C. J. C. H. and Dayan, P. (1992). Q-learning; machine learning, **8**: 279–292.