INSTITUTO SUPERIOR DE ENGENHARIA DE LISBOA

Departamento de Engenharia de
Electrónica e Telecomunicações e de Computadores

Equilíbrio Dinâmico via Aprendizagem por Reforço

Paulo Fernando Pinho Faustino
(Bacharel)

Dissertação de natureza científica realizada para obtenção do grau de
Mestre em Engenharia Informática e de Computadores

Júri
Presidente
Professor Coordenador Fernando Sousa, ISEL – DEETC
Vogais
Arguente: Professor Adjunto Mestre Paulo Araújo, ISEL – DEETC
Orientador: Professor Adjunto Doutor Luís Morgado, ISEL – DEETC

Setembro 2011

Resumo

A Aprendizagem por Reforço é uma área da Aprendizagem Automática que se preocupa com a forma como um *agente* deve tomar *acções* num *ambiente* de modo a maximizar a noção de *recompensa* acumulada. Esta forma de aprendizagem é inspirada na forma como os humanos aprendem e tem levado à criação de diversos algoritmos de aprendizagem por reforço. Estes algoritmos focam a forma de melhorar o comportamento do agente, assumindo uma independência em relação ao meio que os rodeia.

O presente trabalho estuda a aplicação de métodos de aprendizagem por reforço na resolução do problema do pêndulo invertido. Neste contexto é estudado a importância da variabilidade do ambiente (factores externos ao agente) na execução de agentes de aprendizagem por reforço utilizando um modelo que tenta obter equilíbrio (estabilidade) através de dinamismo – o sistema *Cart-Pole* ou pêndulo invertido. Procurou-se melhorar o comportamento dos agentes autónomos alterando a informação passada a estes, mantendo constantes os parâmetros internos dos agentes (ritmo ou taxa de aprendizagem, factores de desconto, ritmo ou taxa de decaimento, etc.), em vez da vertente clássica de afinar os parâmetros internos dos agentes. Estudaram-se as influências nas alterações no conjunto de *estados* e no conjunto de *acções* na capacidade de um agente de resolver o problema do pêndulo invertido.

Estudou-se o comportamento típico dos agentes de aprendizagem por reforço aplicado ao modelo clássico BOXES, sendo proposto uma nova forma de caracterizar o ambiente utilizando a noção de convergência para um valor de referência. Demonstrou-se o ganho em desempenho deste novo método aplicado a um agente Q-Learning.

Palavras-Chave: Equilíbrio Dinâmico, Aprendizagem por Reforço, Agentes Autónomos, Pêndulo Invertido

Abstract

Reinforcement Learning is an area of Machine Learning that deals with how an *agent* should take *actions* in an *environment* such as to maximize the notion of accumulated *reward*. This type of learning is inspired by the way humans learn and has led to the creation of various algorithms for reinforcement learning. These algorithms focus on the way in which an agent's behaviour can be improved, assuming independence as to their surroundings.

The current work studies the application of reinforcement learning methods to solving the inverted pendulum problem. The importance of the variability of the environment (factors that are external to the agent) on the execution of reinforcement learning agents is studied by using a model that seeks to obtain equilibrium (stability) through dynamism – a Cart-Pole system or inverted pendulum. We sought to improve the behaviour of the autonomous agents by changing the information passed to them, while maintaining the agent's internal parameters constant (learning rate, discount factors, decay rate, etc.), instead of the classical approach of tuning the agent's internal parameters. The influence of changes on the *state* set and the *action* set on an agent's capability to solve the Cart-pole problem was studied.

We have studied typical behaviour of reinforcement learning agents applied to the classic BOXES model and a new form of characterizing the environment was proposed using the notion of convergence towards a reference value. We demonstrate the gain in performance of this new method applied to a Q-Learning agent.

Keywords: Dynamic Equilibrium, Reinforcement Learning, Autonomous Agents, Inverted Pendulum

Índice

Índice de Figuras

1 Equilíbrio Dinâmico via Aprendizagem por Reforço

A Aprendizagem por Reforço é uma área da Aprendizagem Automática (um ramo da Inteligência Artificial) que se preocupa com a forma como um agente deve tomar acções num ambiente de modo a maximizar a noção de recompensa acumulada. Diferentes algoritmos de aprendizagem por reforço têm sido propostos, sendo o foco de muitos dos estudos realizados nesta área a optimização desses algoritmos, ou a criação de algoritmos novos ou melhorados, baseados no conhecimento obtido no estudo dos existentes. Não pondo em causa a vertente do estudo dos aspectos internos dos algoritmos envolvidos, questionamo-nos sobre a possível melhoria do seu comportamento apenas alterando factores externos ao algoritmo de aprendizagem, nomeadamente no que se refere à representação do ambiente.

Com o intuito de estudar a importância dos factores externos ao algoritmo de aprendizagem na execução de agentes de aprendizagem por reforço, escolhemos estudar a obtenção de equilíbrio dinâmico através de aprendizagem por reforço. Para tal estudamos a forma de, através do uso de agentes de aprendizagem por reforço, resolver o problema do pêndulo invertido (Cart-Pole) que consiste na manutenção em equilíbrio de uma vara colocada em cima de um carro que se pode deslocar lateralmente. Ao mesmo tempo que procurámos melhorar o desempenho do agente implementado como protótipo de suporte ao estudo, procurámos obter respostas a duas questões principais:

a) De que forma a definição do conjunto de estados altera a capacidade do agente de atingir o seu objectivo?

b) De que forma a definição do conjunto de acções altera a capacidade do agente de atingir o seu objectivo?

A presente dissertação estuda a variação do comportamento de um agente que é inserido num ambiente não-linear (Cart-Pole), para diferentes representações de estados e acções, num contexto de três algoritmos de aprendizagem por reforço (*Adaptive Heuristic Critic* - AHC, Q-Learning e Dyna-Q), representando classes principais de métodos de aprendizagem por reforço, nomeadamente, métodos sem modelo baseados em funções valor,

métodos sem modelo baseados em funções acção/valor e métodos baseados em modelos.

Nesta dissertação apresentamos um modelo de estado e de acção alternativo à abordagem clássica do modelo BOXES, criado por Michie e Chambers (1968), utilizado por muitos investigadores como plataforma de teste de algoritmos de aprendizagem por reforço, e demonstramos como o modelo proposto consegue melhorar o desempenho de agentes Q-Learning e Dyna-Q.

1.1 Aprendizagem por Reforço em Agentes Autónomos

No âmbito da inteligência artificial, um agente autónomo é uma entidade computacional que apresenta três propriedades base: autonomia, reactividade e pro-actividade. Em alguns casos uma quarta propriedade complementa as três anteriores, sociabilidade, a qual é definida como a capacidade de um agente de coordenar as suas actividades com outros agentes, e possivelmente com humanos, de modo a alcançar o seu objectivo, e se for caso disso, ajudar os outros agentes a alcançar os deles (Jennings & Woolridge, 1998).

Por autonomia, entende-se a capacidade de um agente actuar sem intervenção directa de humanos ou outros agentes, tendo controlo sobre o seu estado interno e sobre as acções que executa. Por reactividade, entende-se a capacidade de um agente detectar alterações no meio circundante e reagir atempadamente a essas alterações. Por pro-actividade, entende-se a capacidade de um agente reagir não apenas a estímulos do meio circundante, mas também de uma forma orientada de modo a alcançar os seus objectivos, tomando a iniciativa sempre que apropriado. (Jennings & Woolridge, 1998)

1.1.1 Aprendizagem por Reforço

Aprendizagem por reforço é uma área da aprendizagem automática que trata o modo como um agente executa acções num ambiente de forma a maximizar a noção de recompensa acumulada. A noção de recompensa é fundamental para um agente de aprendizagem por reforço pois é esta a característica que o distingue de outros tipos de agentes inteligentes.

A ideia subjacente a este princípio é de que um agente percepcione a sua situação (estado), escolha uma acção para executar e após a execução da acção escolhida, seja informado da recompensa correspondente a ter tomado essa

acção. No entanto, a recepção de uma recompensa significativa (resultante de se cumprir um objectivo, por exemplo) pode ser diferida no tempo, tornando a observação de qual ou quais as acções que originaram essa recompensa uma tarefa difícil ou até impossível. Devido a esse facto, o agente terá de aprender a prever as acções que levem à maior recompensa acumulada. Dá se o nome de *política* ao mapeamento de estados em acções que resulta deste processo de aprendizagem.

1.1.2 Aprender a Prever

Sutton (1988) formalizou a noção de *aprender a prever*, isto é, utilizando experiências passadas num sistema parcialmente desconhecido para prever comportamento futuro. A este método de previsão foi dado o nome de Aprendizagem por Diferença Temporal (*Temporal Difference Learning* - TD).

O método consiste em efectuar a previsão não da diferença entre a recompensa actualmente recebida e aquela que irá ser obtida no final da sequência de acções (que até poderá nunca ocorrer), mas sim efectuar a previsão da diferença entre a recompensa recebida e aquela que se pensava vir a receber. Esta técnica está na base de alguns dos agentes por nós utilizados neste trabalho.

1.1.3 Aprender sem Modelos do Ambiente

Um modelo do ambiente é uma representação interna que um agente pode utilizar para prever como é que o ambiente irá reagir às suas acções (Sutton & Barto, 1998).

Os algoritmos sem modelo do ambiente são os mais indicados para os agentes que possuem poucos recursos (computacionais e/ou de memória) ou em situações em que a computação é mais dispendiosa em termos de recursos do que as experiências no mundo real.

Um algoritmo sem modelo do ambiente que se utilizou neste trabalho é o *Adaptive Heuristic Critic* (AHC) que utiliza uma função valor (que dá indicação de quão bom é estar num dado estado) para calcular a política óptima.

Um outro algoritmo sem modelo do ambiente utilizado é o algoritmo Q-Learning (Watkins, 1989) que utiliza uma função acção/valor (que dá indicação de quão bom é usar uma dada acção num dado estado) para também calcular a política óptima.

1.1.4 Aprender com Modelos do Ambiente

Os algoritmos com modelos do ambiente, tais como as arquiteturas Dyna propostas por Sutton (1990), utilizam um modelo do ambiente, interno ao agente, para a par das execuções de passos "reais" efectuar planeamento com base em passos de execução simulados internamente, baseados na informação contida no modelo.

Este tipo de algoritmos é adequado a agentes que não têm escassêz de recursos computacionais ou quando a computação é menos dispendiosa do que as experiências do mundo real. Idealmente estes algoritmos convergem para a solução óptima mais depressa do que os equivalentes sem modelo.

1.2 O Sistema em Estudo – Pêndulo Invertido

O pêndulo invertido consiste num sistema em que existe uma vara (que pode ou não ter um peso adicionado ao ponto oposto à base) que faz de pêndulo. Este pêndulo está "preso" à base de um carro que se pode mover livremente numa dimensão. O pêndulo, embora "preso" na base, pode oscilar livremente no mesmo plano de movimento do carro.

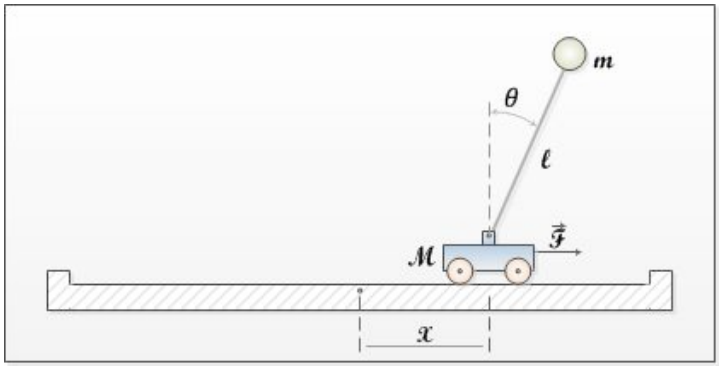

Figura 1 – Representação de um Pêndulo Invertido

No estudo clássico deste sistema (modelo BOXES de Michie e Chambers, 1968), as percepções dadas ao agente são compostas por quatro variáveis X, $\dot{X}$, Θ, $\dot{\Theta}$. Estas variáveis representam respectivamente a posição do carro relativamente ao centro da pista, a velocidade linear do carro, o ângulo do pêndulo respeitante ao eixo vertical do sistema e a velocidade angular do pêndulo. O modelo BOXES obtém o seu nome pela forma como o espaço é particionado em regiões (em caixas ou "boxes" no original em Inglês). Cada

região corresponde a uma combinação das quatros variáveis num dos intervalos possíveis para cada variável:

- Para a posição do carro X:
  - –2.4 m <= X < –0.8 m
  - –0.8 m <= X < 0.8 m
  - 0.8 m <= X <= 2.4 m

- Para a velocidade do carro $\dot{X}$:
  - –infinito m/s < $\dot{X}$ < –0.5 m/s
  - –0,5 m/s <= $\dot{X}$ < 0.5 m/s
  - 0.5 m/s <= $\dot{X}$ < +infinito m/s

- Para o ângulo do pêndulo Θ:
  - $-90^{o} <= \Theta < -6^{o}$
  - $-6^{o} <= \Theta < -1^{o}$
  - $-1^{o} <= \Theta < 0^{o}$
  - $0^{o} <= \Theta < 1^{o}$
  - $1^{o} <= \Theta < 6^{o}$
  - $6^{o} <= \Theta <= +90^{o}$

- Para a velocidade angular do pêndulo $\dot{\Theta}$:
  - –infinito deg/s < $\dot{\Theta}$ < –50 deg/s
  - –50 deg/s <= $\dot{\Theta}$ < 50 deg/s
  - 50 deg/s <= $\dot{\Theta}$ < +infinito deg/s

1.3 O Problema

O problema de "equilibrar" consiste em encontrar uma política comportamental adequada que impeça o carro de atingir o fim da pista, mantendo-o dentro dos limites (2,4 metros para cada lado do centro), ao mesmo tempo que tenta impedir o pêndulo de cair (inclinar-se mais do que 12^{o} para cada lado do eixo vertical do sistema).

O que é que o problema de equilibrar um pêndulo invertido em cima de um carro em movimento tem de especial?

A resposta pode ser encontrada se consideramos que em aprendizagem por reforço um agente tenta resolver o problema por simples tentativa e erro. Ou seja, o agente deve aprender a equilibrar o pêndulo, de forma autónoma, sem ter qualquer conhecimento prévio do problema ou forma de obter informação para além da observação das variáveis de ambiente e de um sinal

de reforço produzido de cada vez que o agente tenta efectuar alguma acção para equilibrar o pêndulo. O conjunto de acções permitidas ao agente é o de "empurrar o carro para a esquerda" e o de "empurrar o carro para a direita", aplicando uma força de valor constante de cada vez que exerce uma acção.

1.4 Estabelecimento de uma Referência

No decurso das nossas investigações, ao recolher informação sobre o desempenho de três tipos de agentes de aprendizagem por reforço: Adaptive Heuristic Critic (AHC), Q-Learning e Dyna-Q, começamos por utilizar as condições de ambiente especificadas por Sutton e Barto (1998). A informação recolhida serviu como referência para as experiências subsequentes.

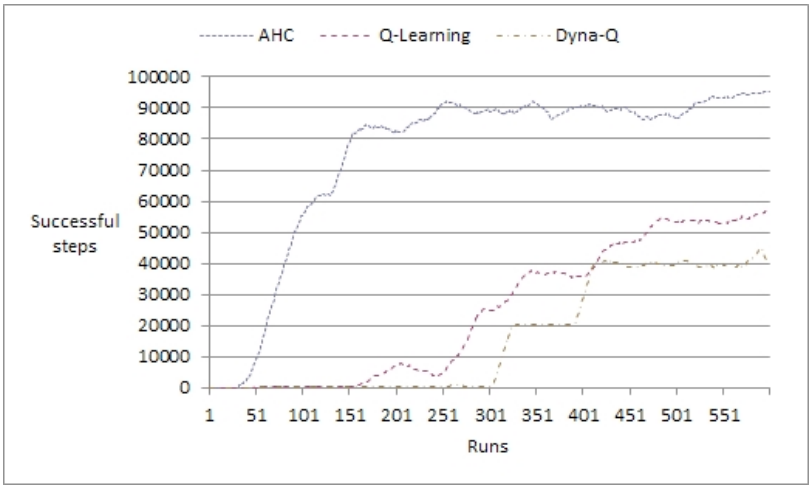

Figura 2 – Obtenção dos valores de referência

Com as condições iniciais de referência, o agente AHC conseguiu resolver o problema do pêndulo invertido mas os agentes Q-Learning e Dyna-Q não o conseguiram (não atingiram valores médios próximos dos 100,000 passos estipulados).

1.5 Estudo de Modelos Alternativos de Estado e de Acção

Com o intuito de estudar o efeito de alterações na representação de estado e de acção no comportamento dos agentes, o conjunto de estados foi reduzido e a semântica das acções foi alterada.

Criou-se um novo conjunto de percepções baseadas na convergência para um valor de referência das variáveis de percepção criando assim um novo

conjunto de estados de agente. Ao eliminar a dependência sobre as variáveis continuas no ambiente (posição X, velocidade $\dot{\mathrm{X}}$, ângulo Θ e velocidade angular $\dot{\Theta}$), e apenas considerando a sua convergência para os respectivos "zeros" (em valores absolutos), conseguiu-se diminuir o conjunto de estados do agente.

$$\delta \mathrm{X} = \begin{cases} 1 : \left|\mathrm{X}_{t-1} < \mathrm{X}_{t}\right| \\ 0 : \left|\mathrm{X}_{t-1} \geq \mathrm{X}_{t}\right| \end{cases} \tag{1.1}$$

$$\delta \dot{\mathrm{X}} = \begin{cases} 1 : \left|\dot{\mathrm{X}}_{t-1} < \dot{\mathrm{X}}_{t}\right| \\ 0 : \left|\dot{\mathrm{X}}_{t-1} \geq \dot{\mathrm{X}}_{t}\right| \end{cases} \tag{1.2}$$

$$\delta \Theta = \begin{cases} 1 : \left|\Theta_{t-1} < \Theta_{t}\right| \\ 0 : \left|\Theta_{t-1} \geq \Theta_{t}\right| \end{cases} \tag{1.3}$$

$$\delta \dot{\Theta} = \begin{cases} 1 : \left|\dot{\Theta}_{t-1} < \dot{\Theta}_{t}\right| \\ 0 : \left|\dot{\Theta}_{t-1} \geq \dot{\Theta}_{t}\right| \end{cases} \tag{1.4}$$

Igualmente, para que esta alteração pudesse resultar, alterou-se a semântica do conjunto de acções do agente de "empurrar para a esquerda" e "empurrar para a direita" para "empurrar para o centro" e "empurrar para fora do centro" e utilizaram-se as percepções originais para obter informação de lateralidade no ambiente (para saber se em determinado momento o "empurrar para o centro" significa efectivamente empurrar para a esquerda ou para a direita).

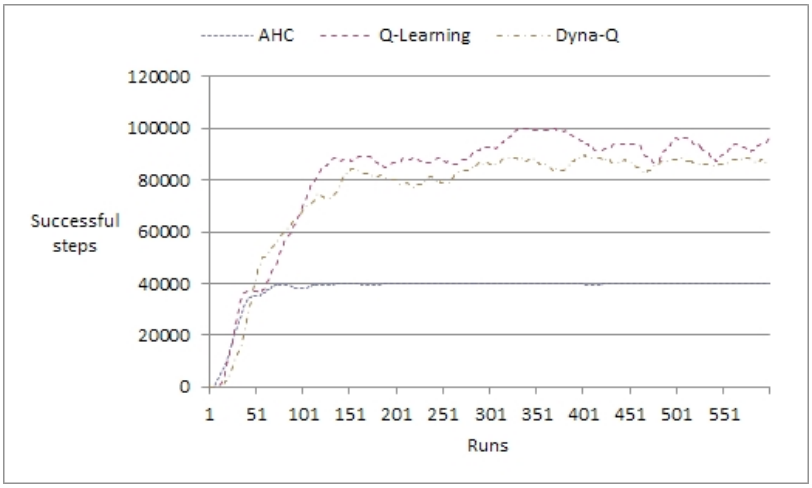

Figura 3 – Uso da posição X como variável de medição de convergência

Com estas novas condições tanto o agente de Q-Learning como o agente de Dyna-Q foram capazes de resolver o problema do pêndulo invertido com uma média de 90,000 passos com sucesso por episódio. No entanto, com estas novas condições o agente de AHC, que anteriormente era capaz de resolver o problema, deixou de o conseguir.

Os agentes utilizados nas experiências com os novos conjuntos de estados e acções sofreram apenas alterações de modo a reflectir as novas representações desses conjuntos. Os parâmetros internos de todos os agentes foram propositadamente mantidos nos seus valores prévios (de referência). Isto foi feito para se avaliar a influência das representações de estado e de acção nos comportamentos dos agentes, devido apenas a alterações realizadas nessas representações (os conjuntos de estados e acções). Qualquer alteração efectuada aos parâmetros nesta altura poderia influenciar o comportamento dos agentes corrompendo a avaliação do impacto das alterações efectuadas.

No que respeita aos mecanismos de aprendizagem por reforço, nenhumas alterações foram efectuadas a quaisquer dos agentes de modo a estudar-se o efeito das alterações nas representações internas independentemente dos mecanismos de aprendizagem.

Em relação aos resultados produzidos pelo agente de AHC, Kaelbling *et al* (1996) fornece uma justificação sugerindo que o algoritmo do AHC tende a ser demasiado dependente nos valores dos parâmetros internos para qualquer dado problema. Verificamos este facto efectuando uma alteração a um dos cinco parâmetros internos do algoritmo, o parâmetro β (taxa de aprendizagem para os pesos do crítico), alterando o valor de 0,5 para 0,7. Esta simples alteração fez com que a média de valores passasse de 40,000 para 60,000.

1.6 Conclusão

Com os resultados apresentados nesta dissertação, acreditamos ter conseguido atingir os objectivos iniciais de resolver o problema do pêndulo invertido através de agentes de aprendizagem por reforço. Implementámos uma nova forma de interpretar as variáveis de percepção permitindo a um agente simples (como é o caso de Q-Learning) resolver com sucesso este problema. Para além disso, obtivemos as respostas às questões iniciais:

- De que forma a definição do conjunto de estados altera a capacidade do agente de atingir o seu objectivo?

- Para os agentes com mecanismos de aprendizagem Q-Learning e Dyna-Q a redução no conjunto de estados foi fundamental no processo de resolver o problema. O número de estados reduzido levou a um aumento do número de vezes que as acções fossem escolhidas para um determinado estado, fazendo com que a convergência para os valores óptimos se realizasse mais rapidamente, obtendo-se assim a política óptima. Para o agente com mecanismo de aprendizagem AHC, a mudança no conjunto de estados (sem adaptar os parâmetros internos) foi suficiente para o agente deixar de resolver o problema, pois os resultados deste mecanismo são dependentes dos valores dos parâmetros internos para uma determinada configuração de problema.

- De que forma a definição do conjunto de acções altera a capacidade do agente de atingir o seu objectivo?
    - A semântica correcta das acções é importante na resolução do problema do pêndulo invertido. O facto de uma alteração na semântica ter sido necessária após a alteração do conjunto de estados mostra o acoplamento estrito que existe entre o agente e o ambiente nos conjuntos de estados e de acções.

1.6.1 Trabalho Futuro

Um trabalho de investigação de natureza científica nunca está terminado. Há sempre mais qualquer coisa que se pode fazer ou estudar. Nesse sentido deixamos algumas indicações de possível trabalho futuro associado ao presente estudo.

- Utilização do modelo de estados reduzido na resolução de problemas de natureza distinta da do pêndulo invertido;

- Implementação de um agente físico para resolução do pêndulo invertido, utilizando como base os agentes estudados neste trabalho;

- Determinação dos limites da capacidade dos agentes estudados, nomeadamente estudando a noção de *robustez* (Sammut, 1994) aplicando alterações ao ambiente e verificando a reação dos agentes.

2 Bibliografia

Alavala C. R. (2008), *Fuzzy Logic and Neural Networks: Basic concepts and applications*, New Age International Publisher, 2008

Barto A. G., Sutton R. S., Anderson C. W. (1983), *Neuronlike Adaptive Elements That Can Solve Difficult Learning Control Problems*. IEEE Transaction on Systems, Man, and Cybernetics, Vol. SMC-13, No. 5, September/October

Blynel J. (2000), *Reinforcement Learning on Real Robots*. University of Aarhus

Braitenberg V. (1984), *Vehicles: Experiments in Synthetic Psychology*, Bradford Book, MIT Press, 1984

Brooks R. (1989), *A Robot that walks: Emergent behaviour form a carefully evolved network*. Neural Computation, 1(2) Summer 1989 pp. 253-262, MIT Artificial Intelligence Laboratory, Cambridge, MA 02139, USA

Brooks R. (1990), *Elephants don't play chess*, Robotics and Autonomous Systems 6 (1990) 3-15, MIT Artificial Intelligence Laboratory, Cambridge, MA 02139, USA

van Hasselt H., Wiering M. A. (2009), *Using continuous action spaces to solve discrete problems*. In: Neural Networks, 2009. IJCNN 2009. International Joint Conference on. 2009. pp. 1149–1156.

Hyman S. E., Malenka R. C., Nestler E. J. (2006), *Neural Mechanismsof Addiction: The Role of Reward-Related Learning and Memory*. The Annual Review of Neuroscience. Doi: 10.1146/annurev.neuro.29.051605.113009, 29: pp. 565-598

Jennings N., Wooldridge M. (1998), *Applications of Intelligent Agents*, In N. Jennings, M. Wooldridge, (Eds.), Agent Technology - Foundations, Applications, and Markets, Springer-Verlag, 1998

Kaelbling L. P., Littman M. L., Moore A. W. (1996), *Reinforcement learning: A survey*. Arxiv preprint cs/9605103

Kobori N., Suzuki K,, Hartono P., Hashimoto S. (2003), *Learning to control a joint driven double inverted pendulum using nested actor/critic algorithm*. In: Neural Information Processing, 2002. ICONIP'02. Proceedings of the 9th International Conference on. 2003. pp. 2610–2614.

Mataric M. J. (2007), *The Robotics Primer*, MIT Press, Cambridge 2007

Michie D., Chambers R. A. (1968), *Boxes: An experiment in adaptive control*. In E. Dale and D. Michie (Eds.), *Machine Intelligence 2*. Edinburgh: Oliver and Boyd

Morgado L. G. (2010), *Agentes Inteligentes: Introdução*, Mestrado em Engenharia Informática e de Computadores, ISEL-DEETC, 2010-2011

Morgado L. G. (2011), *Aprendizagem por Reforço*, Mestrado em Engenharia Informática e de Computadores, ISEL-DEETC, 2010-2011

Morgado L. G. (2008), *Complementos de Inteligência Artificial: Aprendizagem por Reforço*, Mestrado em Engenharia Informática e de Computadores, ISEL-DEETC, 2008-2009

Norvig P., Russell S. (2003), *Artificial Intelligence, A Modern Approach*. 2nd ed. Prentice Hall

Pollack M. E., Ringuette M. (1990), *Introducing the Tileworld: Experimentally Evaluating Agent Architectures*. National Conference on Artificial Intelligence – AAAI, pp. 183-189

Reinforcement learning (2011) – Wikipedia, the free encyclopedia [Internet]. [cited 2011 Jan 20]; Available from: http://en.wikipedia.org/wiki/Reinforcement_learning

Ribeiro C. H. (1999), *A tutorial on reinforcement learning techniques*, Conference Paper, Supervised Learning track tutorials of the 1999 International Joint Conference on Neuronal Networks, 1999

Sammut C. A. (1994), *Recent Progress with BOXES*. In K. Furakawa, Michie, D. & S. Muggleton (Eds.), Machine Intelligence 13. Oxford: The Clarendon Press, OUP, pp. 363-384

Segway (2011) – The leader in personal, green transportation [Internet]. [cited 2011 Jan 20]; Available from: http://www.segway.com/

Singh S. (2011), Large state spaces are hard for RL [Internet]. [cited 2011 May 21]; Available from: http://umichrl.pbworks.com/w/page/7597584/Large-state-spaces-are-hard-for-RL

Stang J. (2005), *The Inverted Pendulum*. Design Project, Cornell University

Sutton R. S. (1988), *Learning to predict by the methods of temporal differences.* Machine Learning 3: pp. 9-44, erratum p.377

Sutton R. S. (1990), *Integrated Architectures for Learning, Planning, and Reacting Based on Approximating Dynamic Programming.* Proceedings of the Seventh International Conference on Machine Learning, pp.216-224, Morgan Kaufmann

Sutton R. S., Barto A. G. (1998), *Reinforcement Learning: An Introduction.* MIT Press/Bradford Books, Cambridge, MA

Sutton R. S. (1999), *Reinforcement Learning,* University of Alberta Webdocs; Available from http://webdocs.cs.ualberta.ca/~sutton/papers/Sutton-99-MITECS.pdf

Tesauro G. (1992), *Practical Issues in Temporal Difference Learning*, Machine Learning 8, pp. 257-277, Kluwer Academic Publisher, 1992

Watkins C. J. (1989), *Learning from delayed rewards.* PhD Thesis, University of Cambridge, England

Wooldridge M.(2002), *Multiagent Systems* , John Wiley & Sons, 2002

Woolridge M. (2009), *An Introduction to Multiagent Systems*, John Wiley and Sons, 2009

Yang L. (2008), *Understanding and analyzing approximate dynamic programming with gradient-based framework and direct heuristic dynamic programming*, Doctoral Thesis, Arizona State University, 2008

O autor:

(Paulo Fernando Pinho Faustino)

O orientador:

(Luís Filipe Graça Morgado)