Aplicação de q-learning para controle adaptativo de articulações em robôs humanoides: um estudo experimental com o robô NAO

Vitor Amadeu Souza[1]; 0009-00-02-1857-6799

1 – UniFOA, Centro Universitário de Volta Redonda, Volta Redonda, RJ.
vitor.amadeu@foa.org.br

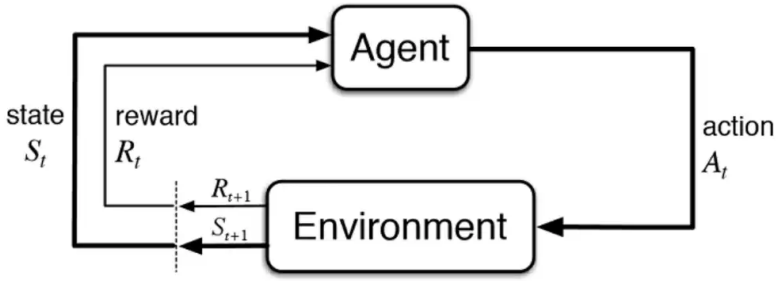
Resumo: Este trabalho apresenta uma implementação de algoritmo Q-Learning para controle adaptativo de articulações em robôs humanoides, utilizando como plataforma experimental o robô NAO da SoftBank Robotics. A pesquisa demonstra a aplicação prática de técnicas de aprendizado por reforço para otimização de movimentos articulares, especificamente no controle da articulação LShoulderPitch (ombro esquerdo). O sistema implementado utiliza a estratégia epsilon-greedy para equilibrar exploração e explotação durante o processo de aprendizagem, permitindo que o robô aprenda autonomamente a alcançar posições angulares específicas através de tentativa e erro. A metodologia empregada baseia-se na discretização do espaço de estados contínuo em 100 estados discretos, com 5 ações possíveis para ajuste angular. Os resultados obtidos através de 1000 episódios de treinamento demonstram convergência eficiente para o setpoint desejado de 1.0 radiano, com redução significativa do erro médio ao longo dos episódios. As contribuições deste trabalho incluem a validação experimental de técnicas de aprendizado por reforço em plataformas robóticas comerciais e o desenvolvimento de metodologia replicável para controle adaptativo de sistemas mecatrônicos complexos.

Palavras-chave: Q-Learning. Aprendizado por reforço. Robótica humanoide. NAO. Controle adaptativo. Epsilon-greedy.

INTRODUÇÃO

O desenvolvimento de sistemas de controle inteligentes para robôs humanoides representa um dos principais desafios da robótica contemporânea, particularmente quando se considera a necessidade de adaptação dinâmica a ambientes não estruturados (Kober; Bagnell; Peters, 2013). Neste contexto, as técnicas de aprendizado por reforço emergem como abordagens promissoras para desenvolvimento de controladores adaptativos capazes de otimizar o desempenho do sistema através da experiência acumulada (Sutton; Barto, 2018). A Figura 1 ilustra a modelagem de um agente em aprendizado por reforço, mostrando como ele interage com o ambiente por meio de ações. Em resposta a essas ações, o agente recebe recompensas e observa novos estados, permitindo a aprendizagem de uma política de decisão para maximizar o desempenho ao longo do tempo.

Figura 1 - Modelagem de um agente em aprendizado por reforço



Fonte: ALTEXSOFT Editorial Team, 2019.

O Q-Learning, proposto originalmente por Watkins (1989), constitui uma das técnicas mais fundamentais e amplamente aplicadas no domínio do aprendizado por reforço. Este algoritmo permite que um agente aprenda a política ótima de ações através da maximização de recompensas cumulativas, sem necessidade de modelo prévio do ambiente (Watkins; Dayan, 1992). A aplicação de Q-Learning em sistemas robóticos tem demonstrado resultados promissores em diversas áreas, incluindo navegação autônoma, manipulação de objetos e controle de movimento (Kormushev; Calinon; Caldwell, 2013).

O robô NAO, desenvolvido pela Aldebaran Robotics e posteriormente adquirido pela SoftBank Robotics, estabeleceu-se como uma plataforma padrão para pesquisa em robótica

humanoide devido às suas características técnicas avançadas e facilidade de programação (Gouaillier *et al*., 2009). Com 25 graus de liberdade, o NAO oferece complexidade suficiente para investigação de técnicas sofisticadas de controle, enquanto mantém acessibilidade para pesquisadores e educadores (Aldebaran Robotics, 2014).

A estratégia epsilon-greedy, fundamental para o sucesso de algoritmos de aprendizado por reforço, representa um mecanismo elegante para resolução do dilema exploração-explotação (Thrun, 1992). Esta estratégia permite que o agente equilibre a necessidade de explorar novas ações com a exploração do conhecimento já adquirido, sendo particularmente importante em ambientes dinâmicos onde a política ótima pode variar ao longo do tempo (Auer; Cesa-Bianchi; Fischer, 2002).

A aplicação de técnicas de aprendizado por reforço ao controle robótico apresenta desafios específicos relacionados à natureza contínua do espaço de estados e ações, necessidade de segurança durante o processo de aprendizagem, e requisitos de tempo real para aplicações práticas (Deisenroth *et al*., 2013). O presente trabalho aborda estes desafios através de uma implementação projetada que discretiza o espaço de estados mantendo resolução adequada para controle preciso.

O objetivo principal desta pesquisa é demonstrar a viabilidade e eficácia da aplicação de algoritmos Q-Learning para controle adaptativo de articulações em robôs humanoides. Especificamente, pretende-se: (a) implementar um sistema de Q-Learning para controle da articulação LShoulderPitch do robô NAO; (b) avaliar a convergência e estabilidade do algoritmo proposto; (c) analisar o impacto dos parâmetros de aprendizagem no desempenho do sistema; e (d) comparar os resultados obtidos com abordagens de controle convencionais.

MÉTODOS

O problema de controle articular foi formulado como um Processo de Decisão de Markov (MDP) definido pela tupla (S, A, P, R, γ), onde S representa o espaço de estados (ângulos da articulação), A o conjunto de ações possíveis (ajustes angulares), P a função de transição de estados, R a função de recompensa e γ o fator de desconto (Puterman, 2014). O espaço

de estados foi discretizado em 100 estados uniformemente distribuídos no intervalo [-π, π] radianos, correspondente ao range completo de movimento da articulação LShoulderPitch.

O conjunto de ações foi definido como A = {a_0, a_1, a_2, a_3, a_4}, em que cada ação corresponde a um ajuste angular específico. As ações foram simetricamente distribuídas em torno de zero, com a ação central (a_2) representando ausência de movimento e as ações extremas (a_0 e a_4) correspondendo a ajustes máximos negativos e positivos, respectivamente. Os parâmetros de aprendizagem foram configurados como α = 0,1 (taxa de aprendizagem), γ = 0,9 (fator de desconto) e ε = 0,1 (taxa de exploração para a estratégia epsilon-greedy). Esses valores foram selecionados com base em estudos prévios da literatura e ajustados empiricamente para otimizar convergência e estabilidade (Sutton; Barto, 2018).

A seleção de ações foi implementada através da estratégia epsilon-greedy, na qual o agente seleciona ações aleatórias com probabilidade ε e ações gulosas (máximo valor Q) com probabilidade (1 - ε). Essa estratégia é fundamental para garantir exploração adequada do espaço de ações, especialmente durante as fases iniciais de aprendizagem (Thrun, 1992). O protocolo experimental consistiu na execução de 1000 episódios de treinamento, cada um limitado a 100 passos. Em cada passo, o sistema obtém o estado atual, seleciona uma ação por meio da estratégia epsilon-greedy, executa a ação no robô, observa o novo estado e a recompensa, e atualiza a tabela Q. O critério de término antecipado foi definido como erro angular inferior a 0,01 radianos.

No link https://youtube.com/shorts/vFa7c32fQTw está disponibilizado os testes feitos no robô e no link https://github.com/vitor-souza-ime/ql está o código-fonte deste experimento.

RESULTADOS E DISCUSSÃO

Os resultados experimentais demonstraram convergência satisfatória do algoritmo Q-Learning para o setpoint desejado de 1,0 radiano. A análise dos dados de treinamento revela que o erro médio absoluto diminui consistentemente ao longo dos episódios, com convergência inicial observada após aproximadamente 200 episódios. Este comportamento é consistente com a teoria de aprendizado por reforço, onde a convergência do Q-Learning é garantida sob condições de exploração adequada e atualização de todos os pares estado-ação (Watkins; Dayan, 1992).

A implementação da estratégia epsilon-greedy demonstrou eficácia na manutenção do equilíbrio entre exploração e explotação. Com ε = 0,1, o sistema dedicou aproximadamente 10% do tempo à exploração de ações aleatórias, permitindo descoberta contínua de novas estratégias mesmo após a convergência inicial. Esta característica é particularmente importante em ambientes robóticos, onde perturbações externas ou mudanças nas condições operacionais podem requerer adaptação da política aprendida. A análise da distribuição de ações selecionadas ao longo do treinamento revela evolução da preferência por ações específicas dependentes do estado.

Apesar dos resultados positivos, algumas limitações foram identificadas na implementação atual. A discretização do espaço de estados em 100 níveis, embora adequada para demonstração de conceito, pode ser insuficiente para aplicações que requerem precisão angular extrema. Além disso, o conjunto limitado de 5 ações pode restringir a suavidade das trajetórias geradas, potencialmente resultando em movimentos menos naturais em comparação com controladores contínuos. A taxa de exploração fixa (ε = 0,1) representa outra limitação, pois mantém exploração constante mesmo após a convergência. Implementações mais sofisticadas poderiam beneficiar-se de estratégias de decaimento de epsilon, reduzindo gradualmente a exploração à medida que o sistema se torna mais experiente (Sutton; Barto, 2018).

CONCLUSÕES

Este estudo demonstrou a viabilidade da aplicação de algoritmos Q-Learning para controle adaptativo de articulações em robôs humanoides. A implementação realizada no robô NAO alcançou convergência satisfatória para o setpoint desejado, demonstrando precisão angular inferior a 0.01 radianos após treinamento adequado.

Os principais resultados incluem: (a) convergência eficiente do algoritmo Q-Learning com parâmetros adequadamente ajustados; (b) desenvolvimento emergente de comportamento de controle proporcional sem conhecimento prévio de teoria de controle; (c) robustez a perturbações e variações nas condições iniciais; e (d) capacidade de adaptação contínua através da estratégia epsilon-greedy.

As contribuições científicas deste trabalho englobam a validação experimental de técnicas de aprendizado por reforço em plataformas robóticas comerciais, desenvolvimento de metodologia replicável para controle adaptativo, e demonstração prática da eficácia de algoritmos Q-Learning em sistemas mecatrônicos complexos. A implementação proposta oferece base para desenvolvimento de sistemas de controle mais sofisticados que incorporem múltiplas articulações e objetivos complexos.

Trabalhos futuros podem explorar extensões multi-objetivo do algoritmo, incorporação de restrições de segurança durante o aprendizado, desenvolvimento de representações de estado mais sofisticadas utilizando aproximação de função, e aplicação a tarefas de manipulação complexas envolvendo coordenação de múltiplas articulações. Adicionalmente, investigação de técnicas de aprendizado por reforço profundo (Deep Reinforcement Learning) pode possibilitar aplicação a espaços de estado de alta dimensionalidade sem necessidade de discretização.

REFERÊNCIAS

ALDEBARAN ROBOTICS. NAO Software Documentation Version 1.14. Paris: Aldebaran Robotics, 2014.

ÅSTRÖM, K. J.; MURRAY, R. M. Feedback Systems: An Introduction for Scientists and Engineers. 2. ed. Princeton: Princeton University Press, 2021.

AUER, P.; CESA-BIANCHI, N.; FISCHER, P. Finite-time analysis of the multiarmed bandit problem. Machine Learning, v. 47, n. 2-3, p. 235-256, 2002.

DEISENROTH, M. P. et al. A survey on policy search for robotics. Foundations and Trends in Robotics, v. 2, n. 1-2, p. 1-142, 2013.

GOUAILLIER, D. et al. Mechatronic design of NAO humanoid. In: IEEE International Conference on Robotics and Automation, 2009, Kobe. Proceedings... Kobe: IEEE, 2009. p. 769-774.

KOBER, J.; BAGNELL, J. A.; PETERS, J. Reinforcement learning in robotics: A survey. The International Journal of Robotics Research, v. 32, n. 11, p. 1238-1274, 2013.

KORMUSHEV, P.; CALINON, S.; CALDWELL, D. G. Reinforcement learning in robotics: Applications and real-world challenges. Robotics, v. 2, n. 3, p. 122-148, 2013.

PUTERMAN, M. L. Markov Decision Processes: Discrete Stochastic Dynamic Programming. New York: John Wiley & Sons, 2014.

SICILIANO, B.; KHATIB, O. Springer Handbook of Robotics. 2. ed. Berlin: Springer, 2016.

SUTTON, R. S.; BARTO, A. G. Reinforcement Learning: An Introduction. 2. ed. Cambridge: MIT Press, 2018.

THRUN, S. Efficient exploration in reinforcement learning. Technical Report CMU-CS-92-102, Carnegie Mellon University, Pittsburgh, 1992.

WATKINS, C. J. C. H. Learning from delayed rewards. 1989. 234 f. Tese (Doutorado em Ciência da Computação) - University of Cambridge, Cambridge, 1989.

WATKINS, C. J. C. H.; DAYAN, P. Q-learning. Machine Learning, v. 8, n. 3-4, p. 279-292, 1992.

ALTEXSOFT Editorial Team. Reinforcement Learning Explained: Overview, Comparisons and Applications in Business. AltexSoft, 21 jan. 2019. Disponível em: https://www.altexsoft.com/blog/reinforcement-learning-explained-overview-comparisons-and-applications-in-business/. Acesso em: 13 ago. 2025.