Análise da influência da taxa de aprendizado e do fator de desconto sobre o desempenho dos algoritmos Q-learning e SARSA: aplicação do aprendizado por reforço na navegação autônoma

André Luiz Carvalho Ottoni [1]
Erivelton Geraldo Nepomuceno [2]
Marcos Santos de Oliveira [3]
Lara Toledo Cordeiro [4]
Rubisson Duarte Lamperti [5]

Resumo: Nos algoritmos de aprendizado por reforço, a taxa de aprendizado (α) e o fator de desconto (γ) podem ser definidos entre qualquer valor no intervalo entre 0 e 1. Assim, adotando os conceitos de regressão logística, é proposta uma metodologia estatística para a análise da influência da variação de α e γ nos algoritmos Q-learning e SARSA. Como estudo de caso, o aprendizado por reforço foi aplicado em experimentos de navegação autônoma. A análise de resultados mostrou que simples variações em α e γ podem interferir diretamente no desempenho do aprendizado por reforço.

***Palavras-chave*:** Aprendizado por reforço. Navegação autônoma. Regressão logística.

Abstract: *In the reinforcement learning algorithms, the step-size parameter (α) and the discount rate (γ) can be set in the range any value between 0 and 1. Therefore, adopting the concepts of logistic regression, we propose a statistical methodology for the analysis of the variation of the two parameters in the Q-learning and SARSA performance. As a case study, the reinforcement learning was applied in a autonomous navigation experiments. The analysis results showed that simple variations in α and γ can interfere directly in reinforcement learning performance.*

***Keywords*:** *Autonomous navigation. Logistic regression. Reinforcement learning.*

1 Introdução

A técnica de aprendizado por reforço (AR) é amplamente aplicada na robótica para resolução de diferentes problemas e situações [1]. O objetivo do AR é fazer com que um agente possa aprender a tomar decisões a partir de experiências de sucesso e fracasso no ambiente.

Nos algoritmos de aprendizado por reforço Q-learning [2] e SARSA [1] , a taxa de aprendizado (α) e o fator de desconto (γ) podem ser definidos entre qualquer valor no intervalo entre 0 e 1 [1]. Dessa forma, a seleção desses parâmetros torna-se um fator importante, pois o desempenho do AR pode ficar comprometido por uma definição inadequada para o experimento [3] [4].

[1]Programa de Pós-Graduação em Engenharia Elétrica, associação ampla da Universidade Federal de São João del-Rei (UFSJ) e Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG).
{andreottoni@ymail.com}
[2]Departamento de Engenharia Elétrica, Universidade Federal de São João del-Rei (UFSJ).
{nepomuceno@ufsj.edu.br}
[3]Departamento de Matemática e Estatística, Universidade Federal de São João del-Rei (UFSJ).
{mso@ufsj.edu.br}
[4]Departamento de Engenharia Mecatrônica e de Telecomunicações, Universidade Federal de São João del-Rei (UFSJ).
{lara1993gv@hotmail.com}
[5]Campus Medianeira, Universidade Tecnológica Federal do Paraná (UTFPR).
{duartelamperti@yahoo.com.br}
http://dx.doi.org/10.5335/rbca.v8i2.5249

Alguns estudos sobre a definição dos parâmetros do aprendizado por reforço já foram realizados. Em [3], os autores mostraram que a convergência do Q-learning é sensível aos valores de α e γ. Já o trabalho [5], introduz o conceito de meta-parâmetros para o AR. Dessa forma, em [5] é proposto um algoritmo para o ajuste de parâmetros do AR de forma dinâmica. Os autores de [6], por sua vez, apresentam um estudo empírico sobre o efeito da taxa de aprendizado na convergência de algoritmos de AR. Outros trabalhos implementam algoritmos para taxas de aprendizados adaptativas em ambientes dinâmicos [7] [4] [8]. Já o trabalho [9], define os parâmetros α e γ por tentativa e erro para um ambiente de navegação simulada. Em uma publicação recente dos autores deste trabalho, é avaliado o desempenho do algoritmo Q-learning na solução do Problema do Caixeiro Viajante, verificando os resultados da variação da polítca ε-*greedy* e da taxa de aprendizado [10]. Ainda na literatura, o método mais simples e muito utilizado é a definição dos parâmetros α e γ constantes em um única combinação inicial, como nos trabalhos [11] [12] [13] [14] [15].

Dessa forma, baseando-se na importância da definição dos parâmetros de taxa de aprendizado e fator de desconto para o desempenho do aprendizado reforço, neste trabalho é proposta uma metodologia estatística para a análise da influência da variação de α e γ nos algoritmos Q-learning e SARSA, adotando os conceitos de regressão logística [16]. O aprendizado por reforço foi aplicado em dois modelos de navegação autônoma para o aprendizado de desvios de obstáculos e busca por um objetivo.

A navegação autônoma é uma das situações em robótica móvel em que é dado alto enfoque nos estudos de aprendizado por reforço [17] [11] [18] [14] [12]. Nesse tipo de problema, geralmente um agente deve aprender a se movimentar por um ambiente desconhecido, evitando colisões com obstáculos. Além disso, o agente também pode buscar durante a exploração um objetivo específico, como o fim da trajetória.

Este artigo está organizado em seções. Na seção 2, são definidos conceitos teóricos do aprendizado por reforço e regressão logística. Na seção 3, o ambiente implementado é descrito. Já na seção 4, são apresentadas os modelos das estratégias de aprendizagem para as navegações reativa e híbrida, por meio das definições das ações, dos estados, modelagem de recompensa. Uma descrição dos experimentos realizados é apresentada na seção 5. A análise dos resultados obtidos é apresentada na seção 6. Finalmente, na seção 7, são apresentadas as conclusões.

2 Fundamentação teórica

2.1 Aprendizado por reforço

O aprendizado por reforço pode ser entendido como uma forma de os agentes aprenderem o que fazer, particularmente quando não existe nenhum professor dizendo ao agente que ação ele deve executar em cada circunstância. Para isso, o agente precisa saber que algo de bom aconteceu quando ele ganha, e que algo de ruim aconteceu quando perde [19].

A técnica de aprendizagem por reforço é fundamentada nos Processos de Decisão de Markov (MDP - Markov Decision Process). O MDP é a uma forma de modelar processos estocásticos. Um MDP é definido pela quádrupla (S, A, T, R), em que, S é um conjunto de estados do ambiente, A é um conjunto de ações possíveis, T é a função de transição de estado e R é a função de recompensa.

2.1.1 Algoritmo Q-learning

O algoritmo Q-learning, proposto por Watkins [2], é um dos algoritmos de aprendizado por reforço mais adotados. Além disso, possui prova de convergência bem estabelecida [2].

A função $Q(s, a)$ é calculada ao se escolher a ação a no estado s e receber o retorno $r(s_t, a_t)$:

$$Q_{t+1} = Q_t(s_t, a_t) + \alpha[r(s_t, a_t) + max_a Q(s_{t+1}, a_t) - Q_t(s_t, a_t)] \quad (1)$$

em que α é a taxa de aprendizagem, γ é o fator de desconto. O Q-learning é retratado no Algoritmo 1 [2].

2.1.2 Algoritmo SARSA

O algoritmo SARSA [1] é uma modificação do Q-learning. O SARSA não adota a maximização das ações do Q-learning, assim a matriz de aprendizado é atualizada como na (2):

$$Q_{t+1} = Q_t(s_t, a_t) + \alpha[r_t + \gamma Q_t(s_{t+1}, a_{t+1}) - Q_t(s_t, a_t)] \tag{2}$$

O SARSA é retratado no Algoritmo 2.

```
1.  Para cada (s,a) inicialize Q(s,a)=0;
2.  Observe  o estado s;
3.  Repita até o critério de parada ser satisfeito
4.        Selecione a ação a usando a política  e-gulosa;
5.        Execute a ação a;
6.        Receba a recompensa imediata  R(s,a);
7.        Observe o novo estado s';
8.        Atualize o item Q(s,a) de acordo com a equação (1);
9.        s = s';
10. Fim Repita
```

Algoritmo 1: Algoritmo Q-learning.

```
1.  Para cada (s,a) inicialize Q(s,a)=0;
2.  Observe  o estado s;
3.  Selecione a ação a usando a política  e-gulosa;
4.  Repita até o critério de parada ser satisfeito
5.        Execute a ação a;
6.        Receba a recompensa imediata  R(s,a);
7.        Observe o novo estado s';
8.        Selecione a nova ação a' usando a política  e-gulosa;
9.        Atualize o item Q(s,a) de acordo com a equação (2);
10.       s = s';
11.       a = a';
12. Fim Repita
```

Algoritmo 2: Algoritmo SARSA.

2.1.3 Taxa de aprendizado

Nos algoritmos Q-learning e SARSA, a taxa de aprendizado pode ser definida em qualquer valor no intervalo entre 0 e 1, $0 \leq \alpha \leq 1$ [1]. Vale ressaltar que, para $\alpha = 0$ não existe aprendizado, já que a atualização da (1) se simplifica em $Q_{t+1} = Q_t(s_t, a_t)$.

Uma condição de convergência do Q-learning [20], estabelece que cada par estado-ação (s, a) deve ser visitado infinitas vezes, com $0 \leq \alpha < 1$, e atendendo a (3) e (4):

$$\sum_{n=1}^{\infty} \alpha_n(s, a) = \infty, \tag{3}$$

$$\sum_{n=1}^{\infty}[\alpha_n(s,a)]^2 < \infty. \tag{4}$$

Uma maneira de satisfazer essas duas condições, é decaindo o valor da taxa de aprendizado de acordo com o número de acessos a cada par (s, a). Para isso, [20] estabeleceu (5):

$$\alpha_n(s,a) = \frac{1}{1 + visitas_n(s,a)}, \tag{5}$$

em que, n é o número de visitas ao par estado-ação (s, a).

No entanto, o método mais comum é a definição da taxa de aprendizado com um valor constante, ou seja, $\alpha_t = \alpha \in (0,1], \forall t \geq 0$ [4]. Nesse caso, (4) não é satisfeita. Dessa forma, o sistema pode nunca completar a convergência, pois continua variando em reposta às recompensas recentes [1].

2.1.4 Fator de desconto

O fator de desconto (γ) permite ao agente selecionar as ações na tentativa de maximizar a soma de recompensas no futuro. A função G_t, em (6), representa a sequência de retornos descontados no tempo:

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + ... = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}, \tag{6}$$

em que, $\gamma \in [0,1]$ [1].

2.1.5 Política ε-*greedy*

No aprendizado por reforço, a política gulosa (*greedy*) seleciona a ação mais bem estimada até o momento. Esse método pode facilmente encontrar uma solução local e ficar preso nesse ponto indefinidamente. Uma alternativa para esse problema é selecionar ações de forma aleatória com pequena probabilidade ϵ. Esse método é denomidado $\epsilon - greedy$ (ou, quase-guloso) [1]. Neste trabalho, o parâmetro ε foi fixado em 0,2.

Na próxima seção, é apresentada a regressão logística, técnica estatística adotada neste trabalho para analisar a influência dos parâmetros taxa de aprendizado (α) e fator de desconto (γ) sobre os resultados dos algoritmos Q-learning e SARSA.

2.2 Regressão logística

Em modelos de regressão, o valor médio da variável resposta é a quantidade chave, denotada por $E(Y/x)$. Essa é a média condicional ($E(Y/x)$), em que Y representa a variável resposta e x denota o valor da variável independente. Para modelos de regressão linear, $E(Y/x)$ é expressa como uma equação linear em x (ou uma transformação de x ou Y), da seguinte forma [16]:

$$E(Y/x) = \beta_0 + \beta_1 x. \tag{7}$$

Com isso, $E(Y/x)$ pode assumir qualquer valor quando x varia entre $-\infty$ e $+\infty$. Com dados dicotômicos para a variável resposta (0 e 1), a média condicional deve ser maior ou igual a zero e menor ou igual a 1, isto é, $0 \leq E(Y/x) \leq 1$. Para esse caso, um modelo utilizado é o da distribuição logística, cuja forma específica é [16]:

$$E(Y/x) = \pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}. \tag{8}$$

O conceito de razão de chances, ou do inglês, *Odds Ratio (OR)*, é um importante índice oferecido pela regressão logística. O OR é adotado para comparar as chances de sucessos entre dois ou mais indivíduos. A razão de chances, para uma variável dicotômica (ou policotômica) independente, é definida pela equação [16]:

$$OR = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}. \quad (9)$$

3 Simulador desenvolvido

De forma a representar um ambiente para a navegação autônoma e aprendizado de desvio de obstáculos e busca por um objetivo, criou-se um simulador no software MATLAB®. O ambiente tem dimensões 15 por 15 nos eixos X e Y, totalizando 225 células. A Figura 1 (a) apresenta interface gerada pelo simulador. Os obstáculos são indicados pela cor vermelha, o caminho livre pela cor verde, o início pela cor verde claro e o objetivo por um asterisco (*).

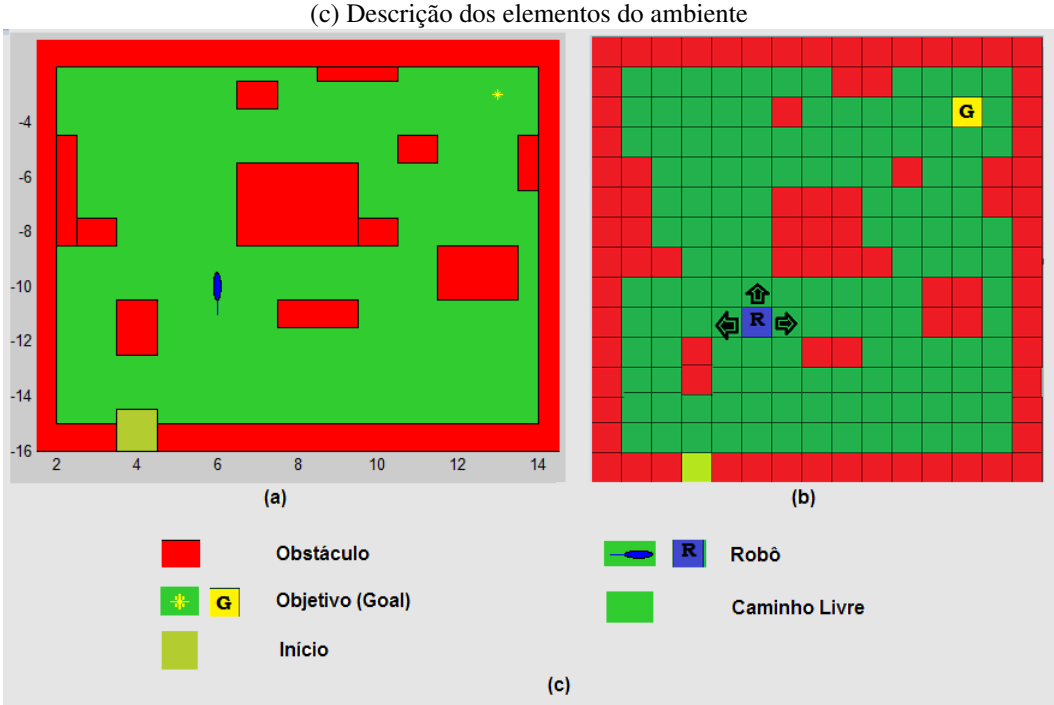
Ao iniciar a simulação, o agente começa do ponto inicial. Quando bate nos obstáculos ou chega ao objetivo, encerra-se o episódio e o agente retorna ao mesmo ponto inicial.

A adoção da simulação de ambientes discretizados para a realização de experimentos de aprendizado por reforço é comum na literatura. Esse domínio é denominado Grid World (Mundo de Grades). Em [1], os autores adotam o Grid World em vários momentos no livro. Já os trabalhos [21] [22] [18], são alguns exemplos de pesquisas que aplicaram o AR no Mundo de Grades na última década. Nos últimos anos, vários trabalhos adotaram esse domínio em análises do aprendizado por reforço, alguns deles são: [13] [23] [24] [25] [26].

Dessa forma, o Grid World representa um problema de constante aplicação em pesquisas de aprendizado por reforço. Provavelmente devido à simplificação desse domínio, que permite aos pesquisadores simularem e comprovarem teorias de mais maneira rápida.

A Figura 1 (b) representa o ambiente simulado no formato de grids. Já a Figura 1 (c), mostra a descrição dos elementos do ambiente.

Figura 1: (a) Interface gerada pelo simulador. (b) Representação do ambiente simulado no formato de grids. (c) Descrição dos elementos do ambiente



Fonte: elaborado pelos autores com base no simulador desenvolvido.

4 Aplicação do aprendizado por reforço no domínio da navegação autônoma

Nesta seção são apresentadas duas aplicações do aprendizado reforço em problemas da navegação autônoma. O primeiro estudo de caso é a navegação reativa. Já a segunda aplicação é a navegação híbrida, em que são interligados conceitos do controle reativo e deliberativo.

A metodologia adotada para o desenvolvimento das estratégias de aprendizagem é dividida em quatro etapas:

1. Definição do conjunto finito de ações que o agente pode realizar;
2. Definição do conjunto finito de estados do ambiente, no qual, o agente está inserido;
3. Definição dos valores dos reforços, para cada par Estado (S) X Ação (A);
4. Aplicação dos algoritmos de aprendizado por reforço, Q-learning e SARSA, no simulador desenvolvido.

4.1 Navegação reativa

A navegação autônoma no formato reativo resume-se basicamente na tomada de decisão a partir de observações imediatas no ambiente. Ou seja, o objetivo no controle reativo é possibilitar ao agente o reconhecimento de objetos e a reação inteligente durante a exploração de um ambiente. Assim, um exemplo da aplicação da navegação reativa é para o desvio de obstáculos em um ambiente desconhecido. Dessa forma, geralmente nesse tipo de navegação, os robôs são dotados por sensores locais, como de odômetria, sonar, laser e visão computacional embarcada.

Em seguida, será apresentada a modelagem do sistema de aprendizado por reforço para o agente reativo. Neste trabalho, o agente simulado para navegação reativa representa um robô com sensores que permitem a percepção de objetos a apenas um passo de distância.

4.1.1 Definição das ações

Nesta etapa são definidas as ações que o agente pode realizar no ambiente do aprendizado. Foram definidas três possíveis ações:

1. Mover para o norte: o agente desloca uma célula para cima no eixo y.
2. Mover para o oeste: o agente desloca uma célula para a esquerda no eixo x.
3. Mover para o leste: o agente desloca uma célula para a direita no eixo x.

Vale ressaltar que a ação mover para o sul não foi considerada por questões de simplificação do modelo e diminuição do custo computacional.

4.1.2 Definição dos estados reativos

As características que compõem os estados do ambiente são caminho livre, obstáculo ou objetivo em algum dos sentidos possíveis de movimentação (norte, leste ou oeste).

- Caminho livre: o agente possui livre acesso e recebe um retorno pouco negativo ao acessá-lo.
- Obstáculo: o agente possui acesso bloqueado e recebe um retorno muito negativo ao acessá-lo.
- Objetivo: representa o final da trajetória. Ao chegar, o agente recebe um retorno muito positivo.

Dessa forma, cada uma dessas três características compõe a definição dos dez estados do ambiente de aprendizado.

1. Caminho livre somente no norte.
2. Caminho livre somente no oeste.
3. Caminho livre somente no leste.
4. Caminho livre somente no norte e oeste.
5. Caminho livre somente no norte e leste.
6. Caminho livre somente no oeste e leste.
7. Caminho livre no norte, oeste e leste.
8. Objetivo no norte.
9. Objetivo no oeste.
10. Objetivo no leste.

4.1.3 Definição das recompensas reativas

Após definir a ações e os estados do ambiente, a próxima etapa foi ponderar os reforços imediatos para cada par Estado (S) $\times$ Ação (A). A ideia do reforço é oferecer ao agente um retorno, positivo (recompensa) ou negativo (penalidade), pela tomada de decisão (ação) em determinada situação (estado) do processo de aprendizado.

Dessa forma, é interessante dar reforços positivos quando o agente chega ao objetivo. Além disso, é importante oferecer penalidades quando a ação executada não for satisfatória.

Assim, os reforços foram definidos da seguinte maneira:

- -1: movimentar por um caminho livre;
- -100: bater em um obstáculo.
- -100: escolher uma ação que não leve ao objetivo, quando esse se encontra a um passo.
- 1000: chegar ao objetivo.

A Tabela 1 apresenta os reforços para cada tomada de decisão em um determinado estado. Por exemplo, se o agente executa a ação 1 (mover para o norte) no estado 2 (caminho livre somente no oeste), então o robô ganha um reforço de -100, já que, ele seguiu em um sentido onde havia obstáculo. No entanto, caso o agente execute a ação 2 (mover para o oeste) no estado 9 (objetivo no oeste), então a recompensa é 1000, pois o robô completou a trajetória.

Tabela 1: Valores de reforços para cada para $S \times A$

Estado/Ação	Norte	Oeste	Leste
1	-1	-100	-100
2	-100	-1	-100
3	-100	-100	-1
4	-1	-1	-100
5	-1	-100	-1
6	-100	-1	-1
7	-1	-1	-1
8	1000	-100	-100
9	-100	1000	-100
10	-100	-100	1000

Fonte: elaborado pelos autores com base na metodologia proposta.

4.2 Navegação híbrida

Arquiteturas de navegação híbrida envolvem componentes dos controles reativos e deliberativos. Conforme apresentado anteriormente, a reatividade está associada à execução de ações predefinidas em resposta a uma informação sensorial obtida localmente [27]. Já o controle deliberativo utiliza informações globais do ambiente para o planejamento da navegação. Geralmente, o robô é dotado com sensores que permitem o acesso a um mapa do ambiente, com as respectivas localizações dos objetos, antes mesmo do início da trajetória.

Em seguida, será apresentada a modelagem do sistema de aprendizado por reforço para a navegação híbrida. Neste trabalho, o agente híbrido tem informações sensoriais locais, como o agente reativo, e também informações do mapa do ambiente (conhecimento deliberativo).

4.2.1 Definição das ações

As ações definidas para a navegação híbrida são equivalentes às ações definidas para o modelo reativo: mover para o norte, mover para o oeste e mover para o leste.

4.2.2 Definição dos estados

Na navegação híbrida o agente possui sua localização no mapa do ambiente. Dessa forma, nesse caso o estado passa ser a posição do agente do mapa. Assim, esse modelo possui 225 estados, ou seja, o número equivalente ao total de células do mapa.

4.2.3 Definição da função de recompensa híbrida

A definição da função de recompensa imediata foi crucial para que esse modelo fosse denominado híbrido. Como pode ser observado em (10), $R(s, a)$ é a soma de duas componentes, a recompensa reativa, definida na seção 4.1.3, e a recompensa deliberativa.

$$R(s, a) = R_{Reativa} + R_{Deliberativa}. \tag{10}$$

A recompensa deliberativa, por sua vez, utiliza as informações sensoriais do mapa do ambiente para o seu cálculo, a partir das posições do agente e do objetivo. Assim, a recompensa deliberativa é definida como a distância entre a posição do agente até o objetivo, multiplicada por -1. Dessa forma, quanto menor a distância para o objetivo, maior será a recompensa deliberativa.

Assim, o valor de $R(s, a)$ aumenta a medida que o agente se aproxima o objetivo, desviando de obstáculos, e alcançando seu valor máximo, $R(s, a)_{Max}$, quando o agente cumpre com sucesso a navegação:

$$R(s, a)_{Max} = R_{ReativaMax} + R_{DeliberativaMax} = 1000 + 0 = 1000. \tag{11}$$

No trabalho [28], os autores aplicam uma técnica semelhante para modelagem de recompensa com reforços intermediários, aplicada ao avanço de agentes em um campo de futebol em busca da área próxima ao gol (objetivo). Conforme [11], reforços intermediários são importantes para acelerar o aprendizado, no entanto, esses reforços devem ter valores inferiores aquele recebido quando o robô atinge o alvo.

5 Experimentos realizados

A aplicação do aprendizado por reforço no simulador desenvolvido visou à navegação autônoma em um ambiente com obstáculos na busca pelo objetivo. Foram realizados quatro experimentos:

- 1º experimento: navegação reativa com o Q-learning;
- 2º experimento: navegação reativa com o SARSA;

- 3º experimento: navegação híbrida com o Q-learning;
- 4º experimento: navegação híbrida com o SARSA.

Os experimentos tiveram como objetivo principal investigar a influência dos parâmetros taxa de aprendizado (α) e fator de desconto (γ) sobre o desempenho dos algoritmos Q-learning e SARSA. Dessa maneira, busca-se entender como esses parâmetros podem interferir nos resultados de sucesso e fracasso do aprendizado por reforço no ambiente em estudo. Para isso, o AR foi simulado com oito distintos valores para α e γ: 0,01, 0,15, 0,30, 0,45, 0,60, 0,75, 0,9 e 0,99. Assim, cada experimento foi testado com 64 combinações desses parâmetros. Além disso, as combinações foram testadas em dez épocas. Ou seja, no total, foram feitas 640 simulações para cada experimento (640 = 64 combinações $\times$ 10 épocas).

Cada simulação foi padronizada com 1000 episódios. Sendo que, para fins de análise com modelos de regressão logística, cada episódio tem como resposta 0 ou 1:

- 0: o agente finalizou o episódio batendo um obstáculo (fracasso);
- 1: o agente finalizou o episódio chegando ao objetivo (sucesso).

Vale ressaltar que, os valores para α e γ foram selecionados na perspectiva de analisar tanto magnitudes baixas e altas desses parâmetros, na compreensão do espaço de definição possível entre 0 e 1. No entanto, a modelagem com quaisquer outros valores para esses parâmetros também é perfeitamente cabível.

6 Análise dos resultados

Neste estudo, foram ajustados quatro modelos de regressão logística binária usando o pacote estatístico MINITAB 14 (versão acadêmica), um para cada experimento realizado. Sendo que, o desfecho de cada modelo logístico indica a probabilidade de sucesso (chegar ao objetivo) de acordo com os parâmetros taxa de aprendizado (α) e fator de desconto (γ) simulados. Dessa forma, cada modelo proposto é descrito como:

- Variável dependente (VD): resultado do episódio (1: sucesso no episódio e 0: facasso no episódio);
- Variável independente 1 (VI_1): taxa de aprendizado (0,01, 0,15, 0,30, 0,45, 0,60, 0,75, 0,9 e 0,99);
- Variável independente 2 (VI_2): fator de desconto (0,01, 0,15, 0,30, 0,45, 0,60, 0,75, 0,9 e 0,99).

As variáveis independentes do modelo são do tipo politocotômico, tendo como grupo de referência $\alpha = 0,01$ (VI_1) ou $\gamma = 0,01$ (VI_2). Dessa forma, o modelo oferece a probabilidade de sucesso no episódio para cada parâmetro sobre a chance de sucesso do grupo de referência. Como as variáveis independentes tem oito categorias, para a devida especificação, são adotadas variáveis auxiliares, definidas como Dummy ($D_1 ... D_{n-1}$). A especificação das variáveis independentes em formato policotômico é apresentada na Tabela 2:

Tabela 2: Especificação das variáveis independentes, adotando 0,01 como grupo de referência

Valor do parâmetro	**Código**	D_1	D_2	D_3	D_4	D_5	D_6	D_7
0,01	0	0	0	0	0	0	0	0
0,15	1	1	0	0	0	0	0	0
0,30	2	0	1	0	0	0	0	0
0,45	3	0	0	1	0	0	0	0
0,60	4	0	0	0	1	0	0	0
0,75	5	0	0	0	0	1	0	0
0,90	6	0	0	0	0	0	1	0
0,99	7	0	0	0	0	0	0	1

Fonte: elaborado pelos autores com base nos experimentos propostos.

6.1 Resultados para a navegação reativa

Os resultados para os modelos ajustados, referentes aos dois primeiros experimentos, são resumidos na Tabela 3.

Tabela 3: Modelos de regressão logística ajustados para o 1º e 2º experimentos

	1º experimento		**2º experimento**	
VI$_1(\alpha)$	β_1	**Razão de chances**	β_1	**Razão de chances**
0,15	0,514936	1,67	-0,544167	0,58
0,30	0,137196	1,15	-0,613778	0,54
0,45	-0,227354	0,80	-0,523326	0,59
0,60	-0,626922	0,53	-0,448643	0,64
0,75	-0,987304	0,37	-0,396049	0,67
0,90	-1,17689	0,31	-0,445057	0,64
0,99	-1,64923	0,19	-0,503566	0,60
VI$_2(\gamma)$				
0,15	0,421941	1,52	0,190834	1,21
0,30	0,670768	1,96	0,298296	1,35
0,45	0,834835	2,30	0,342653	1,41
0,60	0,986633	2,68	0,395253	1,48
0,75	1,11395	3,05	0,369268	1,45
0,90	1,12773	3,09	0,369268	1,45
0,99	0,897571	2,45	0,379084	1,46
β_0	-3,29275		β_0	-3,03294
p-valor	0,000		**p-valor**	0,000

Fonte: elaborado pelos autores com base nos resultados dos experimentos realizados.

Adotando efeitos de significância a um nível para p<0,05, como em ambos os casos o p-valor geral é 0,000, é indicado que se deve aceitar os modelos logísticos ajustados.

No 1º experimento (Q-learning), analisando a razão de chances para a taxa de aprendizado, nota-se que $\alpha = 0{,}15$ alcançou o maior valor (1,67). Dessa forma, a razão de chances indica que $VI_1 = 0{,}15$ tem 1,67 vezes mais chances de alcançar sucesso em um episódio do que o grupo de referência ($\alpha = 0{,}01$). Já analisando a razão de chances para o fator de desconto, nota-se que $\gamma = 0{,}9$ alcançou o maior valor (3,09). Dessa forma, a razão de chances indica que $VI_2 = 0{,}90$ tem mais de três vezes mais chances de alcançar sucesso em um episódio do que o grupo de referência ($\gamma = 0{,}01$). Vale ressaltar que, $\gamma = 0{,}75$ também alcançou índice de razão de chances superior a três vezes ao grupo de referência.

Já o 2º experimento (SARSA), revela um cenário distinto ao encontrado no 1º experimento. Isso porque, analisando a razão de chances para a taxa de aprendizado, o melhor desempenho foi para $\alpha = 0{,}01$. Com relação a razão de chances para o fator de desconto, nota-se que $\gamma = 0{,}6$ alcançou o maior valor (1,48).

Vale ressaltar também que, o desempenho do algoritmo Q-learning foi mais sensível a variações dos parâmetros α e γ, em relação ao SARSA.

O gráficos da Figura 2 deixam mais claro a influência dos parâmetros de taxa de aprendizado e fator de desconto na chance de sucesso em um episódio para as duas implementações de navegação reativa.

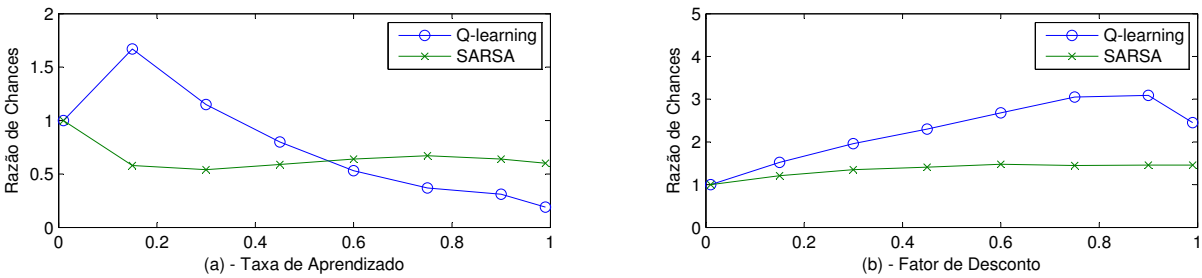
6.2 Resultados para a navegação híbrida

Os resultados para os modelos ajustados, referentes aos 3º e 4º experimentos, são resumidos na Tabela 4.

Adotando efeitos de significância a um nível para p<0,05, como em ambos os casos o p-valor geral é 0,000, é indicado que se deve aceitar os modelos logísticos ajustados.

No 3º experimento (Q-learning), analisando a razão de chances para a taxa de aprendizado, nota-se que $\alpha = 0{,}45$ alcançou o maior valor (4,28). Dessa forma, a razão de chances indica que $VI_1 = 0{,}45$ tem mais de quatro

Figura 2: Influência dos parâmetros de taxa de aprendizado e fator de desconto no desempenho dos algoritmos de aprendizado por reforço na navegação reativa. (a) Taxa de aprendizado (α) *versus* a sua razão de chances. (b) Fator de desconto (γ) versus a sua razão de chances



Fonte: elaborado pelos autores com base nos resultados dos experimentos realizados.

Tabela 4: Modelos de regressão logística ajustados para o 3º e 4º experimentos.

	3º experimento		**4º experimento**	
VI$_1(\alpha)$	β_1	**Razão de chances**	β_1	**Razão de chances**
0,15	1,34919	3,85	1,34909	3,85
0,30	1,44109	4,23	1,47760	4,38
0,45	1,45488	4,28	1,46241	4,32
0,60	1,43989	4,22	1,36296	3,91
0,75	1,36674	3,92	1,19261	3,30
0,90	1,28337	3,61	0,945152	2,57
0,99	1,19512	3,61	0,688887	1,99
VI$_2(\gamma)$				
0,15	0,0807075	1,08	-0,236079	0,79
0,30	0,105384	1,11	-0,370496	0,69
0,45	0,0788984	1,08	-0,411041	0,66
0,60	0,108864	1,12	-0,410430	0,66
0,75	0,156223	1,17	-0,410430	0,67
0,90	0,149951	1,16	-0,416670	0,66
0,99	0,0523950	1,05	-0,490181	0,61
β_0	-1,92050		β_0	-1,52080
p-valor	0,000		**p-valor**	0,000

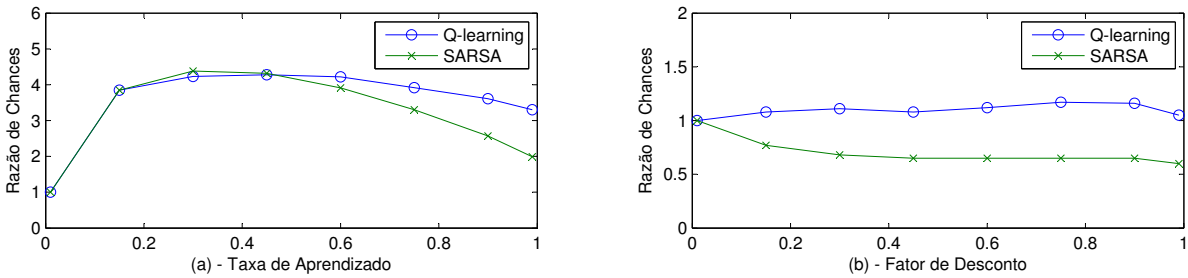
Fonte: elaborado pelos autores com base nos resultados dos experimentos realizados.

vezes mais chances de alcançar sucesso em um episódio do que o grupo de referência ($\alpha = 0{,}01$). No entanto, vale ressaltar, que para todos os valores de $\alpha > 0,01$, o modelo ajustado registrou desempenho superior ao grupo de referência. Já analisando a razão de chances para o fator de desconto, nota-se que $\gamma = 0{,}75$ alcançou o maior valor (1,17).

Já no 4º experimento (SARSA), nota-se que $\alpha = 0{,}30$ alcançou o maior valor índice de razão de chances (4,38). Além disso, como no modelo ajustado para o 3º experimento, para todos os valores de $\alpha > 0,01$ analisados, o modelo ajustado registrou desempenho superior ao grupo de referência. Já analisando a razão de chances para o fator de desconto, nota-se que que o grupo de referência, $\gamma = 0{,}01$, obteve melhor desempenho.

O gráficos da Figura 3 deixam mais claro a influência dos parâmetros de taxa de aprendizado e fator de desconto na chance de sucesso em um episódio para as duas implementações de navegação híbrida.

Figura 3: Influência dos parâmetros de taxa de aprendizado e fator de desconto no desempenho dos algoritmos de aprendizado por reforço na navegação híbrida. (a) Taxa de aprendizado (α) *versus* a sua razão de chances. (b) Fator de desconto (γ) versus a sua razão de chances.



Fonte: elaborado pelos autores com base nos resultados dos experimentos realizados.

6.3 Discussões gerais

As análises revelam que para a mesma tarefa e ambiente, o desempenho de aprendizado do robô simulado sofreu alterações devido a três principais condições:

1. Algoritmo de aprendizado por reforço: Q-learning ou SARSA;
2. Tipo de navegação: reativa ou híbrida;
3. Valor dos parâmetros de taxa de aprendizado (α) e fator de desconto (γ).

A Figura 4 ilustra a diferença de desempenho no processo de aprendizado por reforço. Esse gráfico apresenta a média de passos necessários para o agente alcançar o objetivo nos quatro experimentos com a adoção de dois pares dos parâmetros: (i) $\alpha = 0{,}15$ e $\gamma = 0{,}9$; (ii) $\alpha = 0{,}99$ e $\gamma = 0{,}9$. A média de passos (iterações) foi calculada com dados de dez épocas de simulação para cada combinação. Dessa forma, na Figura 4 percebe-se que os mesmos pares de α e γ podem apresentar comportamentos totalmente distintos para cada tipo de navegação e algoritmo.

Como se observa na Figura 4 (a), o desempenho do agente adotando a combinação $\alpha = 0,15$ e $\gamma = 0,9$ é superior ao par $\alpha = 0,99$ e $\gamma = 0,9$. Ou seja, a alteração do valor da taxa de aprendizado aumentou significativamente a média de passos necessários para que o agente encontrasse o objetivo. No entanto, na Figura 4 (c), 3º experimento, ao adotar $\alpha = 0,15$ e $\gamma = 0,9$ ou $\alpha = 0,99$ e $\gamma = 0,9$, o desempenho de ambos os pares de parâmetros é semelhante. Dessa forma, é necessário o ajuste desses parâmetros de acordo com o método de navegação e algoritmo adotado.

Vale ressaltar que a Figura 4 é apenas uma amostra visual dos efeitos dos pares de parâmetros sobre o desempenho do AR. Assim, quaisquer outros pares de α e γ poderiam ser utilizados para representar esse exemplo.

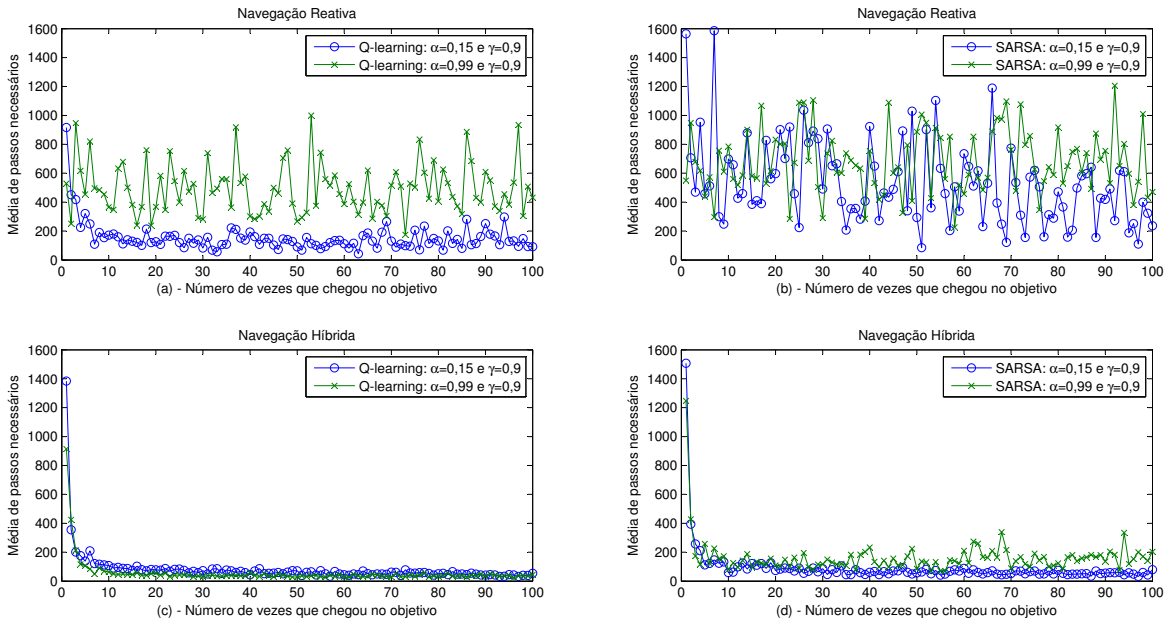
A Tabela 5 apresenta um resumo dos parâmetros com maior razão de chances para cada experimento realizado.

Tabela 5: Parâmetros com maior razão de chances para cada experimento realizado

Experimento	**Algoritmo**	**Navegação**	α	γ
1º	Q-learning	Reativa	0,15	0,9
2º	SARSA	Reativa	0,01	0,6
3º	Q-learning	Híbrida	0,45	0,75
4º	SARSA	Híbrida	0,30	0,01

Fonte: elaborado pelos autores com base nos resultados dos experimentos realizados.

Figura 4: Média de iterações necessárias para alcançar o objetivo versus (i) $\alpha = 0{,}15$ e $\gamma = 0{,}9$; (ii) $\alpha = 0{,}99$ e $\gamma = 0{,}9$



Fonte: elaborado pelos autores com base nos resultados dos experimentos realizados.

7 Conclusão

Este trabalho teve como objetivo principal estudar os efeitos da seleção dos parâmetros taxa de aprendizado (α) e fator de desconto (γ) no desempenho do aprendizado por reforço. Para isso, os algoritmos Q-learning e SARSA foram aplicados em um ambiente simulado para aprendizado de navegação autônoma.

A metodologia de análise proposta, utiliza do índice de razão de chances da regressão logística para levantar quais os melhores valores para os parâmetros α e γ para o estudo de caso. Dessa forma, os resultados indicam em cada experimento qual par (α, γ) representa ter mais chance de sucesso no processo de aprendizado de por reforço em questão. A análise de resultados mostrou que simples variações em α e γ podem interferir diretamente no desempenho do aprendizado por reforço.

Os experimentos revelaram que, apesar da similaridade estrutural dos códigos dos algoritmos Q-learning e SARSA, a variação das taxas proporcionou em certos casos resultados distintos. Para o modelo reativo, o algoritmo Q-learning mostrou-se mais sensível a definição de α e γ, conforme Figura 2. Já para a navegação híbrida, a principal diferença entre os algoritmos aconteceu no desempenho com relação ao fator de desconto, como visto na Figura 3 - b.

A taxa de aprendizado apresentou comportamento interessante. Para os experimentos do modelo reativo, percebe-se a superioridade no desempenho do aprendizado ao adotar valores de α mais próximos de zero, como $\alpha = 0,01$ (Q-learning e SARSA) e $\alpha = 0,15$ (Q-learning). Já para o modelo híbrido, essa relação é oposta, pois o pior desempenho de ambos os algoritmos foi ao adotar a menor taxa de aprendizado ($\alpha = 0,01$). Dessa forma, como na navegação híbrida o agente tem mais informações sensoriais do ambiente do que na navegação reativa, levanta-se a seguite hipótese:

- Hipótese: para um bom desempenho do aprendizado por reforço, a magnitude da taxa de aprendizado deve ser ajustada de acordo com a quantidade de informações do modelo do ambiente disponíveis para o robô.

Quanto ao ambiente de aplicação adotado, buscou-se um estudo de caso simples e tradicional na literatura,

o Grid World. Assim, pretende-se facilitar que outros pesquisadores repliquem as simulações e proponham novos experimentos e análises.

É importante ressaltar que se espera que os estudos realizados neste trabalho atuem como fator motivacional para que outros autores também verifiquem a influência da variação dos parâmetros do AR nas suas respectivas aplicações. Assim, trabalhando na busca por modelos de aprendizado por reforço com desempenhos mais satisfatórios.

Em trabalhos futuros, pretende-se analisar a influência da variação dos parâmetros, α, γ e da política ε-gulosa em outros estudos de casos. Além disso, verificar a hipótese levantada neste trabalho para ambientes de navegação reais.

Agradecimentos

Agradecemos à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação de Amparo a Pesquisa do Estado de Minas Gerais (FAPEMIG), Universidade Federal de São João del-Rei (UFSJ), Universidade Tecnológica Federal do Paraná (UTFPR) e Programa de Pós-Graduação em Engenharia Elétrica (PPGEL - UFSJ/CEFET-MG) pelo apoio.

Referências

[1] SUTTON, R.; BARTO, A. *Reinforcement learning: an introduction*. [S.l.]: Cambridge, MA: MIT Press, 1998.

[2] WATKINS, C. J.; DAYAN, P. Technical note q-learning. *Machine Learning, Kluwer Academic Publishers, Boston*, v. 8, n. 3, p. 279292, 1992. Disponível em: <http://dx.doi.org/10.1023/A:1022676722315>.

[3] EVEN-DAR, E.; MANSOUR, Y. Learning rates for q-learning. *Journal of Machine Learning Research*, v. 5, p. 1–25, 2003. Disponível em: <http://www.jmlr.org/papers/volume5/evendar03a/evendar03a.pdf>.

[4] DABNEY, W. *Adaptive step-sizes for reinforcement learning*. Tese (Doctoral Dissertations in Computer Science) — University of Massachusetts Amherst, Amherst, MA, EUA, 2014. Disponível em: <http://scholarworks.umass.edu/dissertations_2/173>.

[5] SCHWEIGHOFER, N.; DOYA, K. Meta-learning in reinforcement learning. *Neural Networks*, v. 16, n. 1, p. 5–9, 2003. Disponível em: <http://dx.doi.org/10.1016/S0893-6080(02)00228-9>.

[6] GOSAVI, A. On step sizes, stochastic shortest paths, and survival probabilities in reinforcement learning. In: 2008 WINTER SIMULATION CONFERENCE. *Proceedings of the 2008 Winter Simulation Conference*. Austin, TX: IEEE, 2008. p. 525 – 531. Disponível em: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4736109&isnumber=4736042>.

[7] NODA, I. Recursive adaptation of stepsize parameter for non-stationary environments. In: ____. *Adaptive and Learning Agents: Second Workshop, ALA 2009, Held as Part of the AAMAS 2009 Conference in Budapest, Hungary, May 12, 2009. Revised Selected Papers*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010. p. 74–90. ISBN 978-3-642-11814-2. Disponível em: <http://dx.doi.org/10.1007/978-3-642-11814-2_5>.

[8] RYZHOV, I. O.; FRAZIER, P. I.; B., P. W. A new optimal stepsize for approximate dynamic programming. *IEEE Transactions on Automatic Control*, v. 60, p. 743–757, 2015. ISSN 0018-9286. Disponível em: <http://ieeexplore.ieee.org/document/6897935/>.

[9] BAL, S. J.; MAHALIK N, P. A simulation study on reinforcement learning for navigation application. *Artificial Intelligence and Applications*, v. 1, n. 2, p. 43–53, 2014. Disponível em: <http://www.scipublish.com/journals/AIA/papers/819>.

[10] OTTONI, A. L. C. et al. Análise do desempenho do aprendizado por reforço na solução do problema do caixeiro viajante. In: SIMPóSIO BRASILEIRO DE AUTOMAçãO INTELIGENTE, 12. *Anais do XII SBAI - Simpósio Brasileiro de Automação Inteligente*. Natal, RN: SBA, 2015. Disponível em: <http://www.sbai2015.dca.ufrn.br/download/artigo/17>.

[11] SELVATICI, A. H. P.; COSTA, A. H. R. Aprendizado da coordenação de comportamentos primitivos para robôs móveis. *Revista Controle & Automação*, v. 18, n. 2, p. 173–186, 2007. ISSN 0103-1759. Disponível em: <http://dx.doi.org/10.1590/S0103-17592007000200004>.

[12] MONTEIRO, S. T.; RIBEIRO, C. H. C. Desempenho de algoritmos de aprendizagem por reforço sob condições de ambiguidade sensorial em robótica móvel. *Revista Controle & Automação*, v. 15, n. 3, p. 320–338, 2004. ISSN 0103-1759. Disponível em: <http://dx.doi.org/10.1590/S0103-17592004000300008>.

[13] PERICO, D. H.; BIANCHI, R. A. C. Uso de heurísticas obtidas por meio de demonstrações para aceleração do aprendizado por reforço. In: SIMPóSIO BRASILEIRO DE AUTOMAçãO INTELIGENTE, 10. *Anais do X SBAI - Simpósio Brasileiro de Automação Inteligente*. São João del-rei, RN: SBA, 2011. Disponível em: <http://www.sba.org.br/rsv/SBAI/SBAI2011/85784.pdf>.

[14] BENICASA, A. X. Navegação autônoma de robôs baseada em técnicas de mapeamento e aprendizagem de máquina. *Revista Brasileira de Computação Aplicada*, v. 4, n. 1, p. 102–111, 2012. Disponível em: <http://dx.doi.org/10.5335/rbca.2013.1810>.

[15] OTTONI, A. L. C. et al. Desenvolvimento de um sistema de aprendizado por reforço para times de robôs - uma análise de desempenho por meio de testes estatísticos. In: CONGRESSO BRASILEIRO DE AUTOMáTICA, 19. *Anais do XIX CBA - Congresso Brasileiro de Automática*. Campina Grande, PB: SBA, 2012. p. 3557–3564.

[16] HOSMER, D. W.; LEMESHOW, S.; STURDIVANT, R. X. *Applied Logistic Regression*. [S.l.]: 3th. ed. New York: John Wiley & Sons, 2013.

[17] FARIA, G.; ROMERO, R. A. F. Navegação de robôs móveis utilizando aprendizado por reforço e lógica fuzzy. *Revista Controle & Automação*, v. 13, n. 3, p. 219–230, 2002. ISSN 0103-1759. Disponível em: <http://dx.doi.org/10.1590/S0103-17592002000300002>.

[18] OLIVEIRA, A. G.; MELO, J. D.; DORIA, A. D. Análise comparativa de algumas técnicas para o estabelecimento de trajetórias em ambientes com obstáculos usando aprendizagem por reforço. In: CONGRESSO BRASILEIRO DE REDES NEURAIS, 8. *Anais do VIII CBRN - Congresso Brasileiro de Redes Neurais*. Florianópolis, SC: SBRN, 2007. p. 1–6. Disponível em: <http://abricom.org.br/wp-content/uploads/2016/03/50100097.pdf>.

[19] RUSSELL, S. J.; NORVING, P. *Inteligência Artificial*. [S.l.]: Campus, 3st ed., 2013.

[20] MITCHELL, T. M. *Machine Learning*. [S.l.]: McGraw-Hill Science, 1997.

[21] FOSTER, D.; DAYAN, P. Structure in the space of value functions. *Machine Learning*, v. 49, p. 325–346, 2002. ISSN 1573-0565. Disponível em: <http://dx.doi.org/10.1023/A:1017944732463>.

[22] BIANCHI, R. A. C.; RIBEIRO, C. H. C.; COSTA, A. H. R. Accelerating autonomous learning by using heuristic selection of actions. *Journal of Heuristics*, v. 14, n. 2, p. 135–168, 2008. ISSN 1572-9397. Disponível em: <http://dx.doi.org/10.1007/s10732-007-9031-5>.

[23] ZHANG, Q. et al. Reinforcement learning in robot path optimization. *Journal of Software*, v. 7, n. 3, p. 657–662, 2012. Disponível em: <http://dx.doi.org/10.4304/jsw.7.3.657-662>.

[24] CAMPBELL, J. S.; GIVIGI, S. N.; SCHWARTZ, H. M. Multiple-model q-learning for stochastic reinforcement delays. In: *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. [s.n.], 2014. p. 1611–1617. ISSN 1062-922X. Disponível em: <http://dx.doi.org/10.1109/SMC.2014.6974146>.

[25] HAFEZ, M. B.; LOO, C. K. Topological q-learning with internally guided exploration for mobile robot navigation. *Neural Computing and Applications*, v. 26, n. 8, p. 19391954, 2015. Disponível em: <http://dx.doi.org/10.1007/s00521-015-1861-8>.

[26] TAMASSIA, M. et al. Learning options for an mdp from demonstrations. In: CHALUP, S.; BLAIR, A.; RANDALL, M. (Ed.). *Artificial Life and Computational Intelligence*. Springer International Publishing, 2015, (Lecture Notes in Computer Science, v. 8955). p. 226–242. ISBN 978-3-319-14802-1. Disponível em: <http://dx.doi.org/10.1007/978-3-319-14803-8_18>.

[27] ROMERO, R. A. F. et al. *Robótica móvel*. [S.l.]: LTC, 2014.

[28] OTTONI, A. L. C. et al. Análise do aprendizado por reforço via modelos de regressão logística: Um estudo de caso no futebol de robôs. *Revista Junior de Iniciação Científica em Ciências Exatas e Engenharia*, v. 1, n. 10, p. 44–49, 2015. Disponível em: <http://c3.furg.br/components/download_categoria/baixar.php?arquivo=f47330643ae134ca204bf6b2481fec47>.