APRENDIZADO POR REFORÇO NA SOLUÇÃO DO PROBLEMA DO CAIXEIRO VIAJANTE ASSIMÉTRICO: UMA COMPARAÇÃO ENTRE OS ALGORITMOS Q-LEARNING E SARSA

André Luiz Carvalho Ottoni, andreottoni@ymail.com

Grupo de Controle e Modelagem, Programa de Pós-Graduação em Engenharia Elétrica (Associação Ampla UFSJ & CEFET-MG), Universidade Federal de São João del-Rei (UFSJ)

Erivelton Geraldo Nepomuceno, nepomuceno@ufsj.edu.br

Grupo de Controle e Modelagem, Departamento de Engenharia Elétrica, Universidade Federal de São João del-Rei (UFSJ)

Marcos Santos de Oliveira, mso@ufsj.edu.br

Departamento de Matemática e Estatística, Universidade Federal de São João del-Rei (UFSJ)

Resumo. *O Aprendizado de Máquina é uma subárea da Inteligência Artificial e pode ser classificado em três tipos: Aprendizado Supervisionado, Aprendizado Não-Supervisionado e Aprendizado por Reforço (AR). O Aprendizado por Reforço é uma técnica baseado no aprendizado pelo sucesso e fracasso, e fundamentada nos Processos de Decisão de Markov (PDM). Em uma estrutura comum de AR, o aprendizado acontece a partir da interação direta do sistema inteligente (agente) com o ambiente. Baseando-se nisso, este trabalho visa analisar o desempenho do Aprendizado por Reforço na solução do Problema do Caixeiro Viajante Assimétrico (PCVA). Esse problema se trata de um exemplo clássico da otimização combinatorial. Para isso, são adotados os algoritmos de AR mais tradicionais na literatura: Q-learning e SARSA. Pretende-se analisar a influência da definição de parâmetros no Aprendizado por Reforço nesse tipo de problema de otimização combinatória. Assim, o objetivo é desafiar os algoritmos de AR na solução do PCVA, verificando como seus desempenhos são afetados pelas combinações dos parâmetros de taxa de aprendizado e fator de desconto. Além disso, é proposta uma metodologia baseada em experimentos e análises estatísticas para a seleção desses parâmetros para o PCVA. Foram realizadas simulações com quatro instâncias na biblioteca TspLib: Br17, Ftv33, Ftv44 e Ftv64. A análise dos resultados aponta o desempenho dos algoritmos para cada estudo de caso adotado.*

Palavras-chave: *Inteligência Artificial, Aprendizado por Reforço, Problema do Caixeiro Viajante*

1. INTRODUÇÃO

O Aprendizado de Máquina é um importante campo da Inteligência Artificial (IA) (Russell e Norving, 2013). Os sistemas inteligentes dotados com algoritmos de aprendizado conseguem melhorar seu desempenho em uma tarefa por meio da experiência (Mitchell, 1997). De acordo com (Russell e Norving, 2013), o campo do Aprendizado de Máquina pode ser classificado em três casos: Aprendizado Supervisionado, Aprendizado Não-Supervisionado e Aprendizado por Reforço.

O Aprendizado por Reforço (AR) é uma técnica baseada no aprendizado pelo sucesso e fracasso, e fundamentada nos Processos de Decisão de Markov (PDM) (Sutton e Barto, 1998). Em uma estrutura comum de AR, o aprendizado acontece a partir da interação direta de um agente com o ambiente. Assim, no AR o agente usa sensores para identificar o estado (s) atual do ambiente, em seguida executa a melhor a ação (a), e então recebe um retorno para o par estado-ação (s, a). Geralmente, recompensas positivas indicam sucesso na tomada de decisão. Já as recompensas negativas são as penalidades. Dessa forma, o agente armazena essas informações de sucesso e fracasso para auxiliar nas tomadas de decisões futuras.

Os métodos de aprendizado por reforço mais adotados são os baseados em diferença temporal, como os algoritmos

Q-learning e SARSA (Watkins e Dayan, 1992; Sutton e Barto, 1998). Nesses métodos de AR, a taxa de aprendizado (α), o fator de desconto (γ), e a política de ações $\epsilon - greedy$, podem ser definidos entre qualquer valor no intervalo entre 0 e 1 (Sutton e Barto, 1998). Dessa forma, a seleção desses parâmetros se torna um fator importante, pois o desempenho do AR pode ficar comprometido por uma definição inadequada para o experimento (Even-Dar e Mansour, 2003).

Alguns estudos sobre a definição dos parâmetros do aprendizado por reforço já foram realizados. Em (Even-Dar e Mansour, 2003), os autores mostraram que a convergência do Q-learning é sensível aos valores de α e γ. Já o trabalho (Schweighofer e Doya, 2003), introduz o conceito de meta-parâmetros para o AR. Dessa forma, em (Schweighofer e Doya, 2003) é proposto um algoritmo para o ajuste de parâmetros do AR de forma dinâmica. Os autores de (Gosavi, 2008), por sua vez, apresentam um estudo empírico sobre o efeito da taxa de aprendizado na convergência de algoritmos de AR.

O Aprendizado por Reforço possui aplicações diversas na literatura como, robótica, sistemas multiagentes, controle ótimo e otimização. Nesse aspecto, o AR também vem sendo aplicado em um campo específico da otimização, os denominados problemas de otimização combinatória. Um dos exemplos mais clássicos de otimização combinatória é o Problema do Caixeiro Viajante (PCV) (Applegate *et al.*, 2007). Seu objetivo é definir a menor rota entre n cidades. Assim, o caixeiro deve passar por todas as cidades uma única vez, exceto aquela na qual se inicia e termina a jornada. Dessa forma, o AR pode ser aplicado ao PCV na tentativa do caixeiro aprender a sequência de cidades que deve acessar para minimizar a distância percorrida na rota (Gambardella e Dorigo, 1995; Santos *et al.*, 2009; Lima Júnior, 2009).

Nesse aspecto, alguns estudos já abordaram a aplicação do Aprendizado por Reforço no Problema do Caixeiro Viajante. Os autores de (Gambardella e Dorigo, 1995), realizam uma conexão entre a técnica de otimização de Colônia de Formigas, em inglês *Ant System (AS)*, e o Aprendizado por Reforço. Dessa forma, é introduzida a classe de algoritmos Ant-Q. Além disso, o Ant-Q é aplicado na solução do Problema do Caixeiro Simétrico e Assimétrico. Outra abordagem recorrente na solução de problemas de otimização combinatória é o desenvolvimento de soluções híbridas entre Algoritmos Genéticos (AGs) e Aprendizado por Reforço (Miagkikh e Punch, 1999; Liu e Zeng, 2009; Santos *et al.*, 2009; Lima Júnior, 2009). Seguindo a mesma linha de conciliar técnicas para a resolução do PCV, o trabalho (**?**) aplica a aceleração por heurísticas no AR.

Vale ressaltar que, em uma publicação recente dos autores deste trabalho, foram realizados alguns estudos sobre a análise do desempenho do Aprendizado por Reforço na solução do Problema do Caixeiro Viajante (Ottoni *et al.*, 2015). Ottoni *et al.* (2015) verificaram para o algoritmo Q-learning os efeitos de combinações da taxa de aprendizado (constante e decaindo) e política $\epsilon - greedy$ na resolução de três instâncias do PCV Simétrico.

Já este trabalho visa analisar o desempenho do Aprendizado por Reforço na solução do Problema do Caixeiro Viajante Assimétrico (PCVA). Além disso, este artigo adota o Q-learning (Watkins e Dayan, 1992) e também outro algoritmo de AR tradicional na literatura, o SARSA (Sutton e Barto, 1998). Assim, o objetivo é desafiar os algoritmos de AR na solução do PCVA, verificando como seus desempenhos são afetados pelas combinações dos parâmetros de taxa de aprendizado e fator de desconto. Pretende-se também formular uma metodologia baseada em experimentos e análises estatísticas para a seleção desses parâmetros para o Problema do Caixeiro Viajante Assimétrico.

2. ALGORITMOS DE APRENDIZADO POR REFORÇO

O Q-learning proposto por (Watkins e Dayan, 1992) é um dos algoritmos de Aprendizado por Reforço mais conhecidos e adotados. O método se baseia na atualização da matriz de aprendizado Q, a partir da Eq. 1.

$$Q_{t+1}(s,a) = Q_t(s,a) + \alpha[r(s,a) + \gamma max_{a'} Q(s') - Q_t(s,a)], \tag{1}$$

em que:

- $Q_t(s,a)$ é valor no instante t na matriz de aprendizado Q para o par estado (s) $\times$ ação (a);

- $Q_{t+1}(s,a)$ é a atualização da matriz de aprendizado no instante $t+1$ pela execução da ação a no estado s;
- $r(s,a)$ é o recompensa imediata para a execução da ação a no estado s;
- $max_{a'}Q(s')$ é a utilidade de s', ou seja, o valor máximo na matriz de aprendizado na linha do novo estado s'.
- α é a taxa de aprendizado;
- γ é o fator de desconto;
- $s = s_t$, $a = a_t$, $s' = s_{t+1}$ e $a' = a_{t+1}$.

O Algoritmo SARSA (Sutton e Barto, 1998) 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 Eq. 2:

$$Q_{t+1}(s,a) = Q_t(s,a) + \alpha[r_t + \gamma Q_t(s',a') - Q_t(s',a')]. \tag{2}$$

O Algoritmo 1 retrata o SARSA.

Algoritmo 1: SARSA.

Definir os parâmetros: α, γ e ϵ
Para cada par ***s,a*** inicialize a matriz ***Q(s,a)***=0
Observe o estado ***s***
Selecione a ação ***a*** usando a política $\epsilon - greedy$
repita
  Execute a ação ***a***
  Receba a recompensa imediata ***r(s,a)***
  Observe o novo estado ***s'***
  Selecione a nova ação ***a'*** usando a política ε-gulosa
  $Q_{t+1}(s,a) = Q_t(s,a) + \alpha[r_t + \gamma Q_t(s',a') - Q_t(s',a')]$
  s = s'
  a = a'
até *o critério de parada ser satisfeito*;

3. PROBLEMA DO CAIXEIRO VIAJANTE

O Problema do Caixeiro Viajante, no inglês Traveling Salesman Problem, consiste em determinar a menor rota entre um conjunto de cidades, $C = (c_1, c_2, c_3, ..., c_n)$ (Lima Júnior, 2009). A cada par de cidades é dada uma distância (ou custo) associado, d_{ij}. Como restrição, cada localidade deve ser visitada uma única vez. Além disso, o caixeiro deve iniciar e finalizar o percurso na mesma cidade.

O PCV é dito simétrico se satisfaz $d_{ij} = d_{ji}$. No entanto, se $d_{ij} \neq d_{ji}$, então o problema é assimétrico (Lima Júnior, 2009). Ou seja, a distância entre duas cidades é diferente de acordo com o sentido da rota.

4. MODELAGEM DO PROBLEMA

Para a solução do Problema do Caixeiro Viajante via métodos de Aprendizado por Reforço são necessárias algumas definições iniciais, como ações, estados e reforços do modelo. O caixeiro viajante passa então a ser considerado como um agente de aprendizado que deve buscar otimizar as tomadas de decisões (ações) na seleção da ordem das localidades (estados) que deve visitar. Assim, a metodologia adotada para o desenvolvimento da estratégia de aprendizagem é a mesma adotada em (Ottoni *et al.*, 2015):

1. Definição do conjunto finito de estados do ambiente: Nesse caso, os estados são todas as localidades em que o caixeiro viajante (agente) deve acessar.

2. Definição do conjunto finito de ações que o agente pode realizar: Cada ação foi definida como sendo intenção de ir para outra localidade (estado) do problema. Vale ressaltar que, para evitar a repetição de localidades na rota, as ações que levem aos estados já visitados não devem estar disponíveis (Lima Júnior, 2009).

3. Definição dos valores dos reforços, para cada par estado (s) $\times$ ação (a): Os reforços foram definidos como as distâncias entre as localidades multiplicada por -1. Assim, quanto maior a distância, mais negativo é o reforço. Dessa forma, espera-se que o agente procure encontrar a distância mais curta entre duas localidades para diminuir a penalidade. Essa abordagem é a mesma adotada por Bianchi (2004).

4. Aplicação dos algoritmos de aprendizado por reforço Q-learning e SARSA no simulador desenvolvido: Foi desenvolvido um simulador no software *Matlab* para realizar os experimentos.

5. METODOLOGIA DE DEFINIÇÃO DOS PARÂMETROS

Nesta seção é proposta uma metodologia para análise e definição de parâmetros no AR: taxa de aprendizado e fator de desconto. Em seguida, a sequência de passos da metodologia proposta é descrita:

1. Definição dos critérios de desempenho.

   - D_C: Distância calculada por combinação de parâmetros em cada episódio.
   - M_C: Média da distância calculada por combinação de parâmetros ao longo do conjunto de épocas.
   - Critérios de desempate: Média da distância calculada por parâmetros ao longo do conjunto de épocas.
     - M_α: Média para a taxa de aprendizado.
     - M_γ: Média para o fator de desconto.

2. Análise e definição da taxa de aprendizado (α) e fator de desconto (γ).

   (a) Definição do conjunto de valores para γ, espaçados entre 0 e 1. Ver Tab. 2.

   (b) Definição do conjunto de valores para α, espaçados entre 0 e 1. Ver Tab. 2.

   (c) Definição de um valor para o parâmetro ϵ da política de seleção de ações $\epsilon - greedy$. Ver Tab. 2.

   (d) Experimentos com todas as combinações de α e γ.

   (e) Análise dos critérios de desempenho para cada combinação de α e γ .

   (f) Definição dos parâmetros: taxa de aprendizado (α) e fator de desconto (γ).

6. FORMULAÇÃO DOS EXPERIMENTOS

Para cada algoritmo, foram realizados testes com o PCV adotando quatro problemas assimétricos da biblioteca TSPLIB: Br17, Ftv33, Ftv44 e Ftv64. A Tab. 1 especifica o número de localidades e a solução ótima conhecida para cada um dos problemas estudados.

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 possível.

Cada combinação de parâmetros foi simulada em cinco épocas (repetições) com 1000 (mil) episódios. Sendo que, cada episódio teve como resposta a distância total percorrida pelo agente na rota da instância.

A Tab. 2 resume as definições dos experimentos da 2ª Etapa, para cada algoritmo e instância adotados.

Tabela 1. Problemas da TSPLIB estudados.

Problema	**Cidades**	**Solução Ótima Conhecida**
Br17	17	39
Ftv33	34	1286
Ftv44	45	1613
Ftv64	65	1839

Tabela 2. Resumo dos Experimentos na 2ª Etapa da Metodologia.

	Quantidade	**Valores**
α	8	0,01; 0,15; 0,30; 0,45; 0,60; 0;75; 0,90; 0;99
γ	8	0,01; 0,15; 0,30; 0,45; 0,60; 0;75; 0,90; 0;99
ϵ	1	0,01
Combinações	$8 \times 8 \times 1 = 64$	-
Épocas por Combinação	5	-
Episódios por Época	1000 (mil)	-
Episódios por Combinação	5×1000 5000	-
Total de Épocas	$64 \times 5 = 320$	-
Total de Episódios	$320 \times 1000 =$ 320000	-

7. ANÁLISE DOS RESULTADOS

Os resultados dos experimentos apontam uma alta sensibilidade à definição dos parâmetros (α e γ) no desempenho do AR na solução do PCVA. Para exemplificar essa afirmação, as Figs. 1 e 2 apresentam a Média da Distância (M_C) calculada por combinação de parâmetros para as instâncias Br17 e Ftv64, respectivamente. Nessas Figuras é possível observar, por exemplo, uma tendência na diminuição do M_C com valores mais baixos para o fator de desconto. Além disso, para α = 0,01, em todos os casos gerou resultados altos para a Média da Distância calculada.

A Tab. 3 apresenta os melhores resultados dos algoritmos de AR para cada uma das quatros instâncias do Problema do Caixeiro Viajante Assimétrico estudadas: Br17, Ftv33, Ftv44 e Ftv64.

Tabela 3. Parâmetros com melhores critérios de desempenho para cada problema e algoritmo.

		Q-learning			**SARSA**		
Problema	**Critério**	**Valor**	α	γ	**Valor**	α	γ
Br17	*Menor* D_C	39	*	*	39	**	**
Br17	*Menor* M_C	92,707	0,99	0,01	92,892	0,90	0,01
Ftv33	*Menor* D_C	1454	0,75	0,30	1382	0,6; 0,99	0,15
Ftv33	*Menor* M_C	1792,66	0,99	0,01	1776,40	0,99	0,01
Ftv44	*Menor* D_C	1906	0,90	0,01	1795	0,99	0,01
Ftv44	*Menor* M_C	2360,3	0,75	0,01	2327,8	0,90	0,01
Ftv64	*Menor* D_C	2197	0,75	0,01	2140	0,99	0,15
Ftv64	*Menor* M_C	3061,7	0,99	0,15	3093,2	0,99	0,01

*25 combinações dos parâmetros α e γ alcançaram o D_C = 39 para o Q-learning.
**26 combinações dos parâmetros α e γ alcançaram o D_C = 39 para o SARSA.

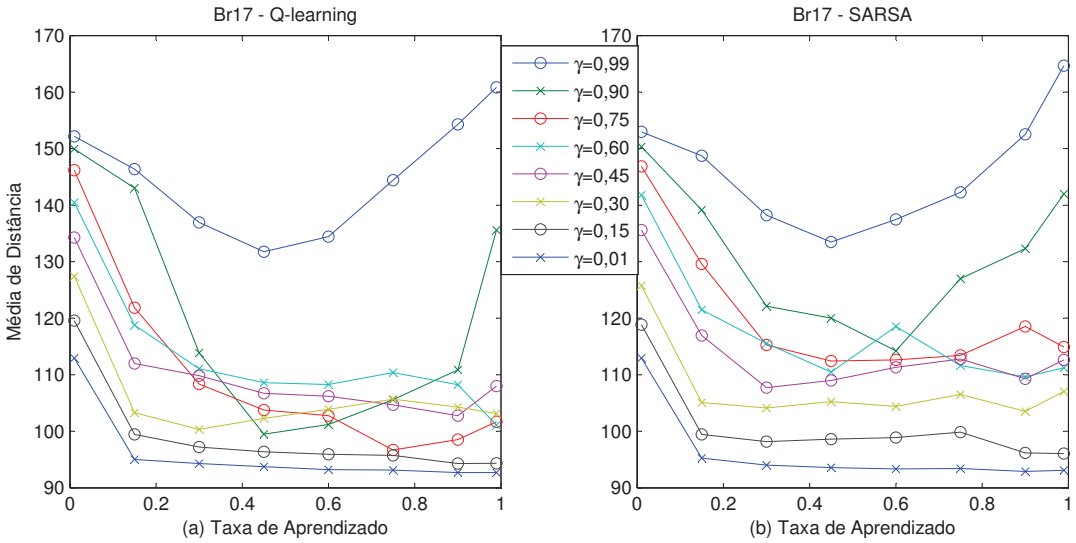

Figura 1. Média de distância para a instância Br17 versus combinações de parâmetros (α e γ) na 2ª Etapa da Metodologia. (a) Q-learning. (b) SARSA.

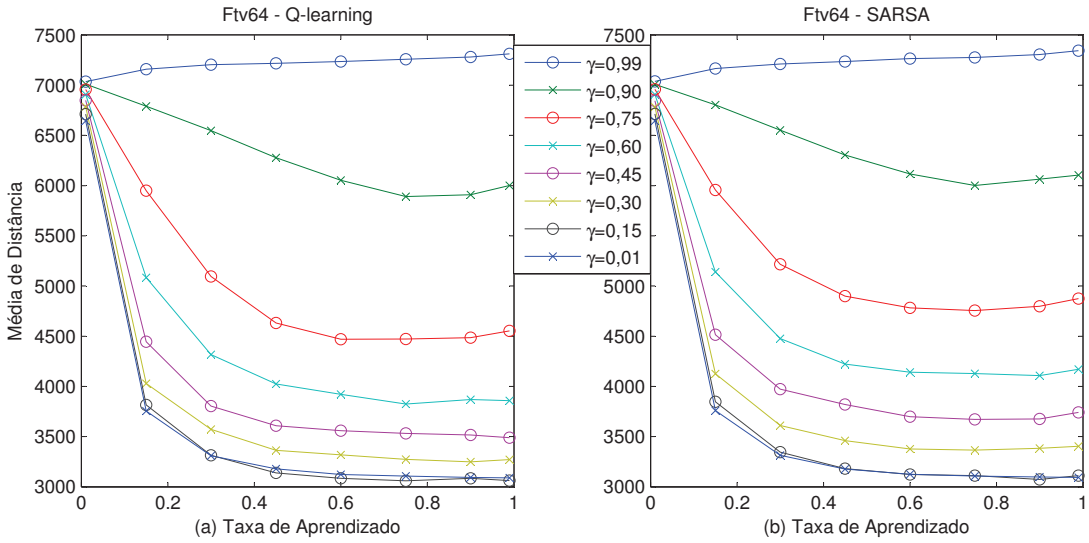

Figura 2. Média de distância para a instância Ftv64 versus combinações de parâmetros (α e γ) na 2ª Etapa da Metodologia. (a) Q-learning. (b) SARSA.

A metodologia proposta na Seção 5, prevê no item 2(f) , a análise dos critérios de desempenho para as combinações de parâmetros. Dessa forma, em seguida são propostas uma sequência de três regras gerais para serem adotadas na análise no passo 2(f) da metodologia:

1. Verificar a combinação de parâmetros que calculou a menor valor para D_C. Definir a 1ª combinação (C_1).
2. Verificar a combinação que gerou a menor média M_C ao longo do processo. Definir a 2ª combinação (C_2).
3. Comparação: Se a primeira combinação de parâmetros é igual à segunda, ($C_1 = C_2$), então definir a combinação $C = C_1 = C_2$. Se não, definir os parâmetros, a partir de C_1 e C_2, observando os critérios de desempate de menores valores para M_α e M_γ.

Assim, após a análise dos critérios de desempenho foram definidos os melhores parâmetros para cada instância e algoritmo, conforme Tab. 4.

Tabela 4. Definição dos parâmetros para cada problema e algoritmo estudado.

Problema	**Algoritmo**	α	γ
Br17	Q-learning	0,99	0,01
Br17	SARSA	0,90	0,01
Ftv33	Q-learning	0,75	0,01
Ftv33	SARSA	0,99	0,01
Ftv44	Q-learning	0,75	0,01
Ftv44	SARSA	0,90	0,01
Ftv64	Q-learning	0,75	0,15
Ftv64	SARSA	0,99	0,01

Para comparar os desempenhos dos algoritmos Q-learning e SARSA, é apresentada a Eq. 3:

$$D_{ij} = \left(1 - \frac{MS_{ij} - SO_i}{SO_i}\right) \times 100\%, \tag{3}$$

em que, D_{ij} é o desempenho do algoritmo i para a instância j, MS_{ij} é a melhor solução do algoritmo i na instância j, SO_j é a solução ótima conhecida da instância j. Assim, a Tab. 5 mostra os desempenhos de Q-learning e SARSA para cada problema analisado.

Tabela 5. Desempenho dos algoritmos.

Problema	**Solução Ótima Conhecida**	**Melhor Solução Q-learning**	**Melhor Solução SARSA**	**Desempenho Q-learning**	**Desempenho SARSA**
Br17	39	39	39	100%	100%
Ftv33	1286	1454	1382	86,93%	92,54%
Ftv44	1613	1906	1795	81,83%	88,72%
Ftv64	1839	2197	2140	80,53%	83,63%

8. CONCLUSÃO

Este trabalho teve como objetivo estudar os efeitos da definição da taxa de aprendizado e fator de desconto sobre o desempenho do AR na solução do PCVA. Para isso, foram adotados os algoritmos Q-learning e SARSA. Assim, a metodologia de definição de parâmetros proposta visa avaliar de α e γ sobre os resultados em cada problema estudado.

O algoritmo SARSA alcançou desempenho superior ao Q-learning na maioria das instâncias adotadas. Além disso, vale ressaltar que, o desempenho de ambos algoritmos diminuiu com o aumento da complexidade dos problemas, ou seja, crescimento do número de localidades nas instâncias.

Em trabalhos futuros, pretende-se aprimorar a metodologia para avaliar os parâmetros α, γ e $\epsilon - greedy$. Assim, unindo aspectos deste trabalho e do artigo anterior destes autores (Ottoni *et al.*, 2015). Dessa forma, apresentar a análise da sensibilidade desses três parâmetros para problemas simétricos e assimétricos.

AGRADECIMENTOS

Agradecemos à UFSJ, FAPEMIG, CAPES, CNPq e PPGEL (Associação Ampla UFSJ & CEFET-MG).

NOMENCLATURA

Em seguida, a nomenclatura adotada neste trabalho:

AR	Aprendizado por Reforço	**Letras gregas**	
PCV	Problema do Caixeiro Viajante	α	taxa de aprendizado
PCVA	Problema do Caixeiro Viajante Assimétrico	γ	fator de desconto

REFERÊNCIAS

Applegate, D., Bixby, R.E., Chvátal, V. e Cook, W., 2007. *The Traveling Salesman Problem: A Computational Study*. Princeton University Press Princeton.

Bianchi, R.A.C., 2004. *Uso de Heurística para a aceleração do aprendizado por reforço.* Tese (Doutorado), Escola Politécnica da Universidade de São Paulo.

Even-Dar, E. e Mansour, Y., 2003. "Learning rates for q-learning". *Journal of Machine Learning Research*, Vol. 5, pp. 1–25.

Gambardella, L.M. e Dorigo, M., 1995. "Ant-q: A reinforcement learning approach to the traveling salesman problem". *Proceedings of the 12th International Conference on Machine Learning*.

Gosavi, A., 2008. "On step sizes, stochastic shortest paths, and survival probabilities in reinforcement learning". *Proceedings of the 2008 Winter Simulation Conference*.

Lima Júnior, F.C., 2009. *Algoritmo Q-learning como Estratégia de Exploração e/ou Explotação para as Mataheurísticas GRASP e Algoritmo Genético*. Tese (Doutorado), Programa de Pós-Graduação em Eng. Elétrica e de Computação da UFRN.

Liu, F. e Zeng, G., 2009. "Study of genetic algorithm with reinforcement learning to solve the tsp". *Expert Systems with Applications*, Vol. 36.

Miagkikh, V. e Punch, W.F., I., 1999. "Global search in combinatorial optimization using reinforcement learning algorithms". In: *Evolutionary Computation, 1999. CEC 99. Proceedings of the 1999 Congress on*. Vol. 1, p. 196 Vol. 1. doi:10.1109/CEC.1999.781925.

Mitchell, T.M., 1997. *Machine Learning*. McGraw-Hill Science.

Ottoni, A.L.C., Nepomuceno, E.G., Cordeiro, L.T., Lamperti, R.D. e Oliveira, M.S., 2015. "Análise do desempenho do aprendizado por reforço na solução do problema do caixeiro viajante". *XII SBAI - Simpósio Brasileiro de Automação Inteligente*.

Russell, S.J. e Norving, P., 2013. *Inteligência Artificial*. Campus, 3rd edi.

Santos, J.Q., Lima Junior, F., Magalhaes, R., de Melo, J. e Neto, A., 2009. "A parallel hybrid implementation using genetic algorithm, grasp and reinforcement learning". In: *Neural Networks, 2009. IJCNN 2009. International Joint Conference on*. pp. 2798–2803. ISSN 1098-7576. doi:10.1109/IJCNN.2009.5178938.

Schweighofer, N. e Doya, K., 2003. "Meta-learning in reinforcement learning". *Neural Networks*, Vol. 16, pp. 5–9.

Sutton, R. e Barto, A., 1998. *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press, 1st edi.

Watkins, C.J. e Dayan, P., 1992. "Technical note q-learning". *Machine Learning*.

NOTA DE RESPONSABILIDADE

Os autores são os únicos responsáveis pelo material reproduzido nesse artigo.