

ANÁLISE DE SENSIBILIDADE DOS PARÂMETROS DO APRENDIZADO POR REFORÇO NA SOLUÇÃO DO PROBLEMA DO CAIXEIRO VIAJANTE: MODELAGEM VIA SUPERFÍCIE DE RESPOSTA

ANDRÉ LUIZ C. OTTONI*, ERIVELTON G. NEPOMUCENO*, MARCOS S. DE OLIVEIRA*

*Programa de Pós-Graduação em Engenharia Elétrica (UFSJ/CEFET-MG)
Universidade Federal de São João del-Rei
São João del-Rei, MG, Brasil

Emails: andreottoni@gmail.com, nepomuceno@ufsj.edu.br, mso@ufsj.edu.br

Abstract— The objective of this work is to perform a sensitivity analysis of the parameters for Reinforcement Learning in the Travelling Salesman Problem Solution. For this it is assumed mathematical modeling via Response Surface Methodology. The adjusted models allow to identify how the performance of Reinforcement Learning is influenced by the levels of learning rate and discount factor.

Keywords— Reinforcement Learning, Traveling Salesman Problem, Response Surface Methodology.

Resumo— O objetivo deste trabalho é realizar uma análise de sensibilidade dos parâmetros do Aprendizado por Reforço na Solução do Problema do Caixeiro Viajante. Para isso, será adotada a modelagem matemática via Metodologia de Superfície de Resposta. Os modelos ajustados permitem identificar como o desempenho do Aprendizado por Reforço é influenciado pelos níveis dos parâmetros taxa de aprendizado e fator de desconto.

Palavras-chave— Aprendizado por Reforço, Problema do Caixeiro Viajante, Metodologia de Superfície de Resposta.

1 Introdução

O Aprendizado por Reforço (AR) é uma técnica baseada no aprendizado pelo sucesso e fracasso, e fundamentada nos Processos de Decisão de Markov (Sutton e Barto, 1998). Em uma estrutura comum do AR, o agente usa sensores para identificar o estado atual do ambiente, para em seguida tomar decisões. Assim, para cada ação executada, o agente recebe um reforço. Essas informações são armazenadas e utilizadas nas escolhas das próximas ações.

O AR possui aplicações diversas, como na robótica, sistemas multiagentes, controle ótimo e otimização (Sutton e Barto, 1998). Nesse sentido, o AR também vem sendo aplicado na resolução de um problema clássico de otimização combinatória, o Problema do Caixeiro Viajante (PCV) (Gambardella e Dorigo, 1995; Bianchi et al., 2009; Liu e Zeng, 2009; Santos et al., 2009; Lima Júnior et al., 2010; Santos et al., 2014; Alipour e Razavi, 2015; Ottoni et al., 2015).

Em Ottoni et al. (2015), são realizados experimentos para analisar a aplicação do AR na solução do PCV, de acordo com a definição dos parâmetros taxa de aprendizado (α) e política ϵ -greedy. Os resultados de Ottoni et al. (2015), apontam que a seleção dos valores de α e ϵ -greedy podem comprometer o desempenho do AR na resolução do PCV.

De fato, estudos já demonstraram que o desempenho do AR pode ser influenciado pela definição de parâmetros, como taxa de aprendizado, fator de desconto e ϵ -greedy (Sutton e Barto, 1998; Schweighofer e Doya, 2003; Even-Dar

e Mansour, 2003; Gatti, 2015).

Assim, a adoção de uma metodologia estatística para analisar os efeitos dos parâmetros dos AR na solução do PCV torna-se um fator relevante (Ottoni et al., 2015). Uma alternativa é a aplicação da Metodologia de Superfície de Resposta (RSM) (Myers et al., 2009). A RSM é uma técnica estatística aplicada nos estudos de processos de otimização (Myers et al., 2009). Trabalhos recentes abordaram a Metodologia de Superfície de Resposta em conjunto com técnicas inteligentes, como Redes Neurais (Gonçalves Júnior et al., 2014) e Algoritmos Genéticos (Mendes et al., 2014). Já em Gatti (2015), a RSM é adotada na análise da influência de parâmetros do AR na convergência do algoritmo $TD(\lambda)$ em dois problemas: *Mountain Car Problem* e *Truck Backer-upper Problem*.

Dessa forma, o objetivo deste trabalho é expandir os estudos de análise de sensibilidade dos parâmetros do AR na solução do PCV. Para isso, será adotada a modelagem matemática via Metodologia de Superfície de Resposta.

Este trabalho está organizado em seções. A seção 2 apresenta conceitos teóricos iniciais do Problema do Caixeiro Viajante e Metodologia de Superfície. Em seguida, a seção 3 descreve fundamentos do sistema de Aprendizado por Reforço. Os experimentos realizados e a estrutura da modelagem matemática proposta são apresentados nas seções 4 e 5, respectivamente. Já a seção 6, descreve a análise dos resultados. Finalmente, na seção 7 são apresentadas as conclusões.

2 Fundamentação Teórica

2.1 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, $C = (c_1, c_2, c_3, \dots, c_n)$ (Applegate et al., 2007; Lima Júnior et al., 2010). A cada par de cidades é dada uma distância (ou custo) associado, c_{ij} . Como restrição, cada localidade deve ser visitada uma única vez. Além disso, o caixeiro deve iniciar e finalizar o percurso na mesma cidade.

Geralmente, o PCV é formulado sobre um grafo $G = (N, A)$, em que, N é conjunto de nós (vértices), e A é o conjunto de arcos (i, j) do problema (Goldbarg e Luna, 2005).

Neste trabalho, o PCV será abordado sobre dois paradigmas: Simétrico e Assimétrico. No caso Simétrico, o custo associado ao deslocamento de uma cidade i para uma localidade j é equivalente ao custo de ir de j para i . Já no problema Assimétrico, o sentido de realização da rota pode alterar o valor da distância total percorrida.

Os experimentos foram realizados adotando problemas da TSPLIB¹ (Reinelt, 1991). A biblioteca TSPLIB, *A Traveling Salesman Problem Library*, é um repositório de dados aberto que reúne opções de estudos de caso do PCV. O repositório TSPLIB apresenta problemas (instâncias) tanto para o PCV Simétrico, quanto para casos Assimétricos. Além disso, a base de dados fornece o valor da solução ótima conhecida para cada instância da biblioteca.

2.2 Metodologia de Superfície de Resposta

A Metodologia de Superfície de Resposta, em inglês *Response Surface Methodology (RSM)*, reúne um conjunto de técnicas estatísticas para a otimização de processos (Myers et al., 2009). A medida de desempenho é denominada resposta. Já as variáveis de entrada são ditas variáveis independentes (VIs) (Myers et al., 2009).

Os modelos de superfície de resposta apresentam a mesma estrutura dos modelos de regressão linear múltipla (Myers et al., 2009). Assim, as Equações 1 e 2 apresentam a estrutura dos modelos RSM de 1ª e 2ª ordem, respectivamente, com duas VIs (x_1 e x_2):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon, \quad (1)$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \epsilon. \quad (2)$$

O efeito do erro na resposta é representado por ϵ . Para a estimação dos coeficientes do modelo (β), normalmente é adotado o método de mínimos quadrados, assumindo distribuição normal com média zero e variância constante (Myers et al., 2009).

¹<http://comopt.ifi.uni-heidelberg.de/software/TSPLIB95/>

De acordo com Myers et al. (2009), os modelos de 2ª ordem são mais adotados por serem mais flexíveis e mais bem ajustáveis em problemas de superfície reais.

3 Sistema de Aprendizagem por Reforço

No Aprendizagem por Reforço, agente e ambiente se interagem em uma sequência de passos de tempo discretos ($t = 0, 1, 2, 3, \dots$). A cada instante de tempo t , o agente recebe uma representação do ambiente, por meio do estado, $s_t \in S$, e seleciona uma ação $a_t \in A$. Um instante de tempo depois, $t + 1$, o agente recebe um reforço, $r_{t+1} \in R$, e observa o novo estado s_{t+1} . Sendo que, S é o conjunto de todos os estados, A é o conjunto das ações e R é a função de recompensa (Sutton e Barto, 1998).

Os parâmetros taxa de aprendizado (α) e fator de desconto (γ) estão presentes em boa parte dos métodos de AR. Esses parâmetros geralmente são definidos no intervalo entre 0 e 1.

A taxa de aprendizado é responsável por controlar efeitos das novas atualizações na matriz de aprendizado. Assim, para $\alpha = 0$ não existe aprendizado, já que a atualização do algoritmo se simplifica para $Q_{t+1}(s, a) = Q_t(s, a)$ (Ottoni et al., 2015).

Já o fator de desconto permite ao agente selecionar as ações na tentativa de maximizar a soma de recompensas no futuro. A função G_t , na Equação 3, representa a sequência de retornos descontados no tempo:

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}, \quad (3)$$

em que $\gamma \in [0, 1]$ (Sutton e Barto, 1998).

Em seguida, uma breve descrição do algoritmo de AR adotado neste trabalho.

3.1 Algoritmo SARSA

O SARSA (Sutton e Barto, 1998) é um tradicional algoritmo de Aprendizagem por Reforço. O algoritmo SARSA (Algoritmo 1) recebeu esse nome pois envolve na sua atualização os termos: s_t (estado no instante t), a_t (ação executada no instante t), $r(s_t, a_t)$ (retorno para o par $s_t \times a_t$), s_{t+1} (estado no instante $t + 1$) e a_{t+1} (ação executada no instante $t + 1$).

A Equação 4 descreve a atualização da matriz de aprendizado Q pelo SARSA, com a execução da ação a no estado s :

$$Q_{t+1} = Q_t(s, a) + \alpha[r(s_t, a_t) + \gamma Q_t(s', a') - Q_t(s, a)]. \quad (4)$$

No Algoritmo 1, a política denominada ϵ - greedy (ou, quase-guloso) é responsável pelo controle entre gula e aleatoriedade na seleção das ações (Sutton e Barto, 1998).

Definir os parâmetros: α, γ e ε
Para cada s, a inicialize $Q(s, a) = 0$
Observe o estado s
Selecione a ação a usando a política

ε -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 -greedy
Atualize $Q(s, a)$ com a Eq. (4)
 $s = s'$
 $a = a'$

até o critério de parada ser satisfeito;

Algoritmo 1: SARSA.

3.2 Modelo de Aprendizado por Reforço

A aplicação do Aprendizado por Reforço na solução do Problema do Caixeiro Viajante requer a definição de um modelo com um conjunto de estados (S), ações (A) e recompensas (R). Assim, neste trabalho é adotada a estrutura apresentada em Ottoni et al. (2015):

- **Estados:** Localidades que o caixeiro viajante deve acessar.
- **Ações:** As ações representam a intenção de ir para outro estado (localidade) no problema.
- **Recompensas:** Os reforços são as distâncias entre as localidades multiplicada por -1. Assim, quanto maior a distância entre duas localidades, mais negativa é a recompensa.

Dessa forma, o modelo adotado tem como objetivo permitir ao caixeiro viajante (agente) aprender a se deslocar pelas localidades (estados) otimizando a escolha pelos próximos destinos (ações) na rota (ambiente).

4 Experimentos Realizados

Os experimentos foram realizados no *software MATLAB*[®] e compreenderam testes em cinco instâncias da TSPLIB, conforme Tabela 1:

Tabela 1: Problemas da TSPLIB estudados.

Tipo	Problema	Cidades	Solução Conhecida
Simétrico	berlin52	52	7542
	brazil58	58	25395
	kroA200	200	29368
Assimétrico	ftv33	34	1286
	ftv44	45	1613

Para cada problema do PCV abordado foram realizadas simulações envolvendo um conjunto de

64 combinações dos parâmetros taxa de aprendizado (α) e fator de desconto (γ). Os valores para α e γ são:

- α : 0,01, 0,15, 0,30, 0,45, 0,60, 0,75, 0,90 e 0,99.
- γ : 0,01, 0,15, 0,30, 0,45, 0,60, 0,75, 0,90 e 0,99.

Além disso, cada combinação foi simulada em cinco épocas (repetições) com 1000 episódios. Vale ressaltar que, a resposta de um episódio é a distância total (custo) percorrida pelo caixeiro na rota (Ottoni et al., 2015).

5 Modelagem Matemática

Neste trabalho, foram ajustados modelos matemáticos de Superfície de Resposta de 2ª ordem. Esses modelos têm como objetivo representar a sensibilidade aos parâmetros (α e γ) no desempenho do AR na solução de cada instância do Problema do Caixeiro Viajante estudada:

- 1º Modelo: berlin52.
- 2º Modelo: brazil58.
- 3º Modelo: kroA200.
- 4º Modelo: ftv33.
- 5º Modelo: ftv44.

Assim, a estrutura dos modelos propostos é composta por três variáveis: y , α e γ . A variável resposta (y) representa a média da distância percorrida pelo caixeiro na rota. Para cada uma das cinco repetições de cada instância, foi calculada uma média para cada combinação de α e γ . Além disso, as variáveis independentes são $VI_1 = \alpha$ e $VI_2 = \gamma$. Dessa forma, os modelos possuem o formato da Equação 5:

$$y = \beta_0 + \beta_1\alpha + \beta_2\gamma + \beta_3\alpha^2 + \beta_4\gamma^2 + \beta_5\alpha\gamma. \quad (5)$$

Para o ajuste dos modelos foi adotado o pacote RSM do *software* estatístico R (Lenth, 2009; R Core Team, 2013).

6 Análise dos Resultados

Os resultados para o ajuste dos modelos de superfície de resposta são descritos em seguida. Além disso, são apresentados alguns gráficos de contornos e superfícies de resposta. Também são apresentados os pontos estacionários no fim desta seção.

6.1 Modelos Ajustados

Os modelos ajustados devem satisfazer algumas medidas de adequação. Uma dessas análises visa verificar se os resíduos dos modelos estão distribuídos normalmente (Hines et al., 2006). Define-se os resíduos como $e_i = y_i - \hat{y}_i$, $i = 1, 2, \dots, n$, em que y_i é uma observação e $\hat{y}_i$ é o correspondente valor estimado a partir do modelo de regressão (Hines et al., 2006). Utilizou-se o teste de Kolmogorov-Smirnov (KS) (Razali e Wah, 2011) para a verificação dessa suposição. Os correspondentes p-valores do teste KS para os modelos foram: $p = 0,2491$, $p = 0,3649$, $p = 0,1943$, $p = 0,3692$ e $p = 0,1803$, confirmando a hipótese de normalidade dos resíduos.

Outros componentes de análise da adequação de um modelo de superfície de resposta são: coeficiente de determinação múltipla (R^2) e coeficiente de determinação múltipla ajustado (R_a^2) (Myers et al., 2009). Esses coeficientes, definidos entre 0 e 1, indicam o quanto da variabilidade é explicada pelo modelo. Assim, se R^2 e R_a^2 se aproximam de 1, apontam um bom ajuste do modelo à amostra (Hines et al., 2006). A Tabela 2 apresenta os valores ajustados para R^2 e R_a^2 .

Tabela 2: Coeficientes de determinação múltipla.

Modelo	R^2	R_a^2
berlin52	0,9037	0,9022
brazil58	0,9080	0,9065
kroA200	0,8965	0,8948
ftv33	0,8872	0,8854
ftv44	0,8932	0,8915

A Tabela 3 apresenta os coeficientes ajustados para cada modelo em estudo. Os testes de significância de coeficientes individuais para este trabalho revelaram que, para os cinco modelos, todos os coeficientes são altamente significantes ($p < 0,001$).

6.2 Gráficos de Contornos e Superfícies de Resposta

A metodologia de superfície de resposta fornece duas ferramentas gráficas para análise: gráfico de contornos e superfície de resposta (Myers et al., 2009).

Neste trabalho, o gráfico de contornos apresenta em duas dimensões a relação entre a taxa de aprendizado (α) e o fator de desconto (γ). Assim, a partir de linhas de contorno é possível identificar regiões que se aproximam do mínimo ou máximo da resposta. As Figuras 1 e 2 apresentam os gráficos de contornos para o 1º e 4º modelos, referentes respectivamente as instâncias berlin52 e ftv33.

Já para a análise em três dimensões, a ferramenta adotada é o gráfico de superfície de resposta. As Figuras 3 e 4 apresentam essa visualiza-

Tabela 3: Coeficientes para cada modelo.

Modelo	Coef.	β	p
berlin52	Intercepto	18656	$< 2 \times 10^{-16}$
	α	-22730	$< 2 \times 10^{-16}$
	γ	-7097	$1,9 \times 10^{-11}$
	α^2	14985	$< 2 \times 10^{-16}$
	γ^2	14487	$< 2 \times 10^{-16}$
	$\alpha\gamma$	6718	$< 2 \times 10^{-16}$
brazil58	Intercepto	68807	$< 2 \times 10^{-16}$
	α	-85423	$< 2 \times 10^{-16}$
	γ	-30211	$2,16 \times 10^{-13}$
	α^2	54619	$< 2 \times 10^{-16}$
	γ^2	58197	$< 2 \times 10^{-16}$
	$\alpha\gamma$	30530	$< 2 \times 10^{-16}$
kroA200	Intercepto	191276	$< 2 \times 10^{-16}$
	α	-236975	$< 2 \times 10^{-16}$
	γ	-75721	$1,07 \times 10^{-11}$
	α^2	147295	$< 2 \times 10^{-16}$
	γ^2	147568	$< 2 \times 10^{-16}$
	$\alpha\gamma$	62817	$1,59 \times 10^{-14}$
ftv33	Intercepto	3110	$< 2 \times 10^{-16}$
	α	-3888	$< 2 \times 10^{-16}$
	γ	-1281	$4,64 \times 10^{-14}$
	α^2	2658	$< 2 \times 10^{-16}$
	γ^2	2278	$< 2 \times 10^{-16}$
	$\alpha\gamma$	877	$9,28 \times 10^{-13}$
ftv44	Intercepto	4298	$< 2 \times 10^{-16}$
	α	-5442	$< 2 \times 10^{-16}$
	γ	-1593	$3,93 \times 10^{-11}$
	α^2	3550	$< 2 \times 10^{-16}$
	γ^2	3079	$< 2 \times 10^{-16}$
	$\alpha\gamma$	1401	$3,29 \times 10^{-15}$

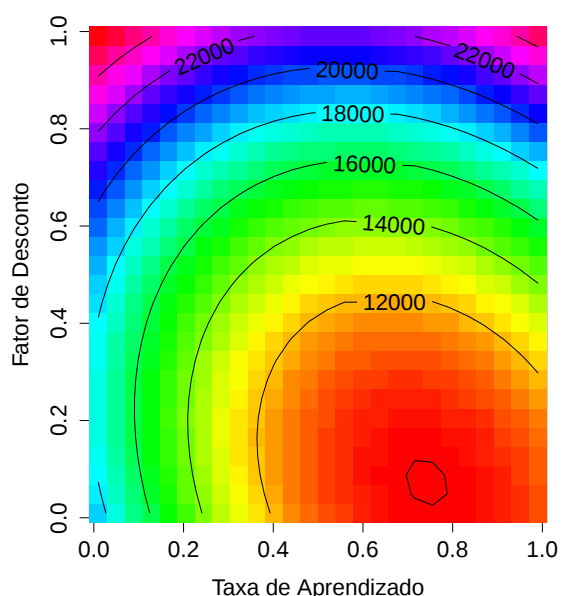


Figura 1: Gráfico de contornos para o 1º modelo (berlin52).

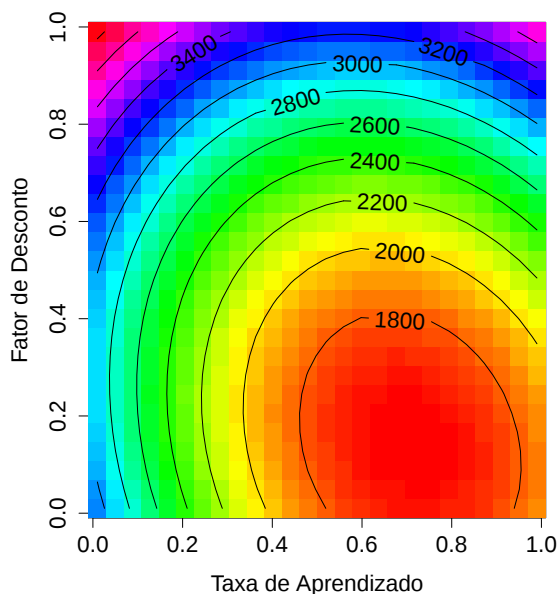


Figura 2: Gráfico de contornos para o 4º modelo (ftv33).

ção em superfície para os modelos 1 e 4. De forma similar ao gráfico de contornos, é possível identificar regiões de α e γ que se aproximam do mínimo da variável resposta (distância percorrida).

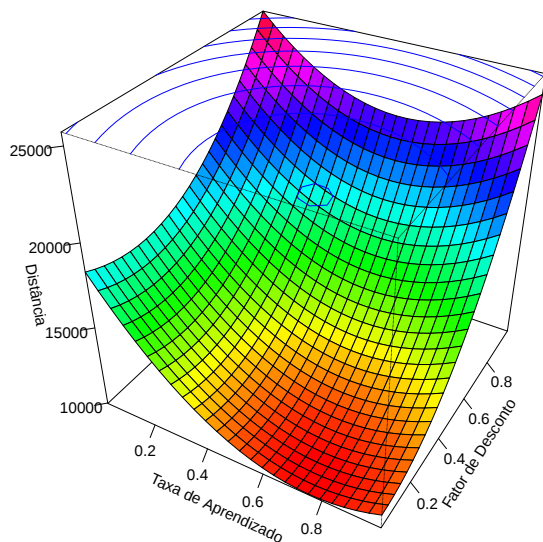


Figura 3: Superfície de resposta para o 1º modelo (berlin52).

6.3 Pontos Estacionários

Na metodologia de superfície de resposta, a identificação dos pontos estacionários (mínimo ou máximo) é interessante para verificar os valores que

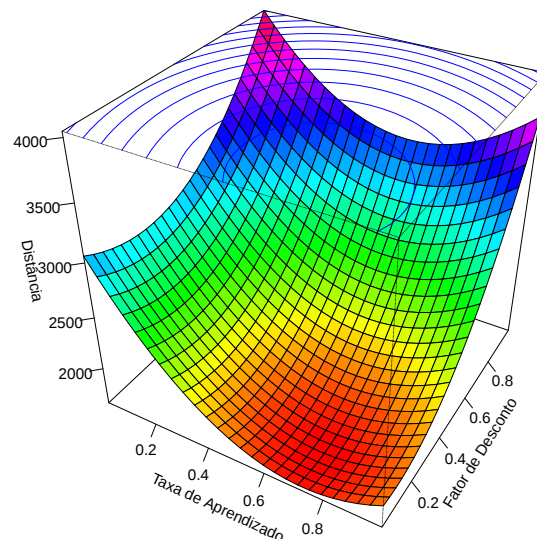


Figura 4: Superfície de resposta para o 4º modelo (ftv33).

otimizam as resposta predita no modelo (Myers et al., 2009). No Problema do Caixeiro Viajante, o objetivo é minimizar a distância percorrida na rota. Sendo assim, os pontos estacionários desejados nas superfícies modeladas são os mínimos das funções.

Neste trabalho, os pontos estacionários foram obtidos a partir da análise canônica no *software* R (Myers et al., 2009; Lenth, 2009), e são apresentados na Tabela 4.

Tabela 4: Pontos estacionários.

Modelo	α	γ
berlin52	0,7420503	0,0729088
brazil58	0,7655741	0,0587433
kroA200	0,7853619	0,0894069
ftv33	0,7074018	0,1449799
ftv44	0,7490532	0,0882221

Em seguida, foi realizada uma nova fase de experimentos para verificar o desempenho do AR com a adoção dos valores dos pontos estacionários para os parâmetros α e γ . Assim, as combinações foram simuladas em cinco repetições com 10000 episódios. A Tabela 5 apresenta os melhores resultados encontrados para cada instância, com os valores de taxa de aprendizado e fator de desconto ajustados (pontos estacionários).

Além disso, foram realizados experimentos adotando os parâmetros utilizados em outros trabalhos: $\alpha = 0,1$ e $\gamma = 0,3$ (Gambardella e Dorigo, 1995; Bianchi et al., 2009), $\alpha = 0,9$ e $\gamma = 1$ (Lima Júnior et al., 2010) e $\alpha = 0,99$ e $\gamma = 0,01$ (Ottoni et al., 2015). Essas combinações também foram simuladas em cinco repetições com 10000

episódios e os melhores resultados são apresentados na Tabela 5.

Os parâmetros ajustados pela RSM alcançaram os melhores resultados em três instâncias: brazil58, ftv33 e ftv44.

Tabela 5: Melhor solução encontrada para cada problema adotando os valores dos pontos estacionários (PE) e parâmetros de outros trabalhos.

Prob.	TSP	D95	L10	O15	PE
berlin52	7542	9169	16441	8046	8048
brazil58	25395	28284	49156	27150	26685
kroA200	29368	38468	121879	33125	35311
ftv33	1286	1533	2245	1458	1381
ftv44	1613	2033	2692	1863	1812

Prob.: Problema. TSP: Solução conhecida da TSPLIB.

D95: solução com parâmetros de Gambardella e Dorigo (1995).

L10: solução com parâmetros de Lima Júnior et al. (2010).

O15: solução com parâmetros de Ottoni et al. (2015).

PE: solução com parâmetros de pontos estacionários.

7 Conclusão

Os modelos de Superfície de Reposta ajustados permitem identificar como o desempenho do Aprendizado por Reforço é influenciado pelos níveis dos parâmetros taxa de aprendizado e fator de desconto. Isso é possível graças ao conjunto de ferramentas disponíveis pela RSM. Os gráficos de contornos e superfície de resposta oferecem um importante aspecto visual quanto a sensibilidade do AR aos valores de α e γ . Já a análise de pontos estacionários permite inferir para cada modelo quais os valores dos parâmetros otimizam a resposta.

Em trabalhos futuros, pretende-se aprimorar os estudos de análise de sensibilidade dos parâmetros de AR na solução do Problema do Caixeiro Viajante, aplicando a modelagem por Superfície de Resposta para outros algoritmos, como o *Q-learning*. Pretende-se também, investigar a sensibilidade dos parâmetros do AR em outros problemas de otimização combinatória e também outros domínios tradicionais de aplicação do AR, como robótica móvel e sistemas multiagentes.

Agradecimentos

Agradecemos à CAPES, CNPq, FAPEMIG e UFSJ.

Referências

Alipour, M. M. e Razavi, S. N. (2015). A new multiagent reinforcement learning algorithm to solve the symmetric traveling salesman problem, *Multiagent and Grid Systems* **11**(2): 107–119.

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., Ribeiro, C. H. C. e Costa, A. H. R. (2009). On the relation between ant colony optimization and heuristically accelerated reinforcement learning,

1st International Workshop on Hybrid Control of Autonomous System pp. 49–55.

Even-Dar, E. e Mansour, Y. (2003). Learning rates for q -learning, *Journal of Machine Learning Research* **5**: 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* pp. 252–260.

Gatti, C. (2015). *Design of Experiments for Reinforcement Learning*, Springer International Publishing.

Goldberg, M. C. e Luna, H. P. L. (2005). *Otimização Combinatória e Programação Linear*, Elsevier/Campus.

Gonçalves Júnior, A. M., Rocha e Silva, V. V., Baccarini, L. M. R. e Reis, M. L. F. (2014). Three-phase induction motors faults recognition and classification using neural networks and response surface models, *Journal of Control, Automation and Electrical Systems* **25**(3): 330–338.

Hines, W. W., Montgomery, D. C., Goldsman, D. M. e Borror, C. M. (2006). *Probabilidade e Estatística na Engenharia*, LTC.

Lenth, R. V. (2009). Response-surface methods in r, using rsm, *Journal of Statistical Software* **32**(7): 1–17.

Lima Júnior, F. C., Neto, A. D. D. e Melo, J. D. (2010). *Hybrid Metaheuristics Using Reinforcement Learning Applied to Salesman Traveling Problem, Traveling Salesman Problem, Theory and Applications*, Prof. Donald Davendra (Ed.), InTech.

Liu, F. e Zeng, G. (2009). Study of genetic algorithm with reinforcement learning to solve the tsp, *Expert Systems with Applications* **36**(3): 6995 – 7001.

Mendes, L. F. S., Baccarini, L. M. R. e Abreu Júnior, L. (2014). Diagnóstico de falhas em motores de indução utilizando superfície de resposta e algoritmos genéticos, *XX CBA - Congresso Brasileiro de Automação* pp. 2946–2953.

Myers, R. H., Montgomery, D. C. e Anderson-Cook, C. M. (2009). *Response surface methodology: process and product optimization using designed experiments*, 3 edn, John Wiley & Sons.

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* pp. 43–48.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Razali, N. M. e Wah, Y. B. (2011). Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests, *Journal of Statistical Modeling and Analytics* **2**(1): 21–33.

Reinelt, G. (1991). Tsp lib - a traveling salesman problem library, *ORSA Journal on Computing* **3**(4): 376–384.

Santos, J. P. Q., Melo, J. D., Duarte Neto, A. D. e Aloise, D. (2014). Reactive search strategies using reinforcement learning, local search algorithms and variable neighborhood search, *Expert Systems with Applications* **41**(10): 4939–4949.

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, *Neural Networks, 2009. IJCNN 2009. International Joint Conference on*, pp. 2798–2803.

Schweighofer, N. e Doya, K. (2003). Meta-learning in reinforcement learning, *Neural Networks* **16**(1): 5–9.

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