Análise do Aprendizado por Reforço via Modelos de Regressão Logística: Um Estudo de Caso no Futebol de Robôs

André Luiz C. Ottoni, Marcos S. de Oliveira, Erivelton G. Nepomuceno, Rubisson D. Lamperti.

***Abstract*—O aprendizado por reforço (AR) é um formalismo da Inteligência Artificial que permite a um agente aprender a partir da sua interação com o ambiente no qual ele está inserido. No aprendizado por reforço, quanto mais complexo o ambiente, isto é, quanto maior o número de ações ou a quantidade de agentes, maior é a capacidade computacional necessária para resolver o problema. Dessa forma, verificar o processo de convergência de um sistema de aprendizado por reforço é importante a fins de evitar esforços computacionais desnecessários. Nesse aspecto, o presente trabalho teve sua principal meta de investigação a aplicação da regressão logística para análise do aprendizado por reforço. A análise procurou avaliar a convergência do algoritmo de AR (Q-learning) aplicado em um time de futebol de robôs simulado. Com os resultados dos modelos de regressão logísticos construídos foi possível verificar o ponto de estabilização do aprendizado por reforço.**

Keywords*—*Aprendizado por Reforço, Regressão Logística, Futebol de Robôs

I. Introdução

O aprendizado por reforço (AR) é um formalismo da Inteligência Artificial (IA) que permite a um agente aprender a partir da sua interação com o ambiente no qual ele está inserido [1]. A ideia central da técnica de aprendizado por reforço é que as percepções são utilizadas não apenas para agir, mas ao mesmo tempo, para melhorar a habilidade do agente para agir no futuro. A aprendizagem ocorre à medida que o agente observa suas interações com o ambiente e com seus próprios processos de tomada de decisão [2].

No AR, quanto mais complexo o ambiente, isto é, quanto maior o número de ações ou a quantidade de agentes, maior é a capacidade computacional necessária para resolver o problema [3]. Dessa forma, verificar o processo de convergência de um sistema de aprendizado por reforço é importante a fim de evitar esforços computacionais desnecessários.

Em 1989, Watkins propôs em sua tese de doutorado o Q-learning [4], algoritmo adotado neste trabalho. Desde a elaboração do Q-learning, pesquisas e publicações vem propondo diferentes aplicações e análises para o AR. Robótica móvel [5], otimização na produção de petróleo [6], tráfego aéreo [7] e controle ótimo de descarregadores de navios [8] são alguns exemplos de aplicações do AR encontrados na literatura. Outras pesquisas atuam na linha de tentar diminuir o tempo gasto para convergência dos algoritmos de AR. Esse é o caso do trabalho de Bianchi [3], que propôs heurísticas para a aceleração do aprendizado. As pesquisas relacionadas ao aprendizado reforço em ambientes multiagente também têm seu destaque [9], [10]. Algumas publicações já apresentaram resultados positivos para aplicações do AR na plataforma de simulação da Robocup [11], [12], [13], [14], [15]).

Nos últimos anos, os autores deste trabalho vem estudando algumas metodologias estatísticas para avaliar o comportamento de sistemas multiagentes [16]. O principal estudo de caso adotado é o futebol de robôs simulado da RoboCup[1]. A categoria de simulação 2D da RoboCup simula partidas de futebol de robôs autônomos. Um simulador fornece aos agentes todos os dados que seriam obtidos na realidade por meio dos seus sensores e calcula o resultado das ações de cada agente. Cada jogador é visto como um agente individual, e o time como um sistema multiagente totalizando 11 (onze) jogadores por equipe.

Em trabalhos anteriores [14], [17], estes autores abordaram as metodologias de testes de hipótese, análise de variância e comparações múltiplas, para avaliar o comportamento do sistema de AR. Já neste artigo, o enfoque é dado a análise de convergência do aprendizado por reforço via modelos de regressão logística [18]. Para isso, foi adotado o modelo do sistema de aprendizado por reforço (ações, estados, recompensas) proposto e descrito pelos autores no trabalho [17].

Este trabalho está organizado em oito seções. Na seção 2 são apresentados conceitos do aprendizado por reforço. Já a seção 3 mostra conceitos da teoria de regressão logística. A seção 4 apresenta o futebol de robôs simulado. Já a seção 5, mostra a modelagem do aprendizado por reforço para o estudo de caso. As seções 6 e 7 abordam os experimentos realizados e a análise de resultados, respectivamente. Finalmente, na seção 8 são apresentadas as conclusões.

André Luiz C. Ottoni é graduando em Engenharia Elétrica na Universidade Federal de São João del-Rei. Email: andreottoni@ymail.com.

Marcos S. de Oliveira é professor Doutor no Departamento de Matemática e Estatística da Universidade Federal de São João del-Rei. Email: mso@ufsj.edu.br.

Erivelton G. Nepomuceno é professor Doutor no Departamento de Engenharia Elétrica na Universidade Federal de São João del-Rei. Email: nepomuceno@ufsj.edu.br.

Rubisson D. Lamperti é professor Mestre no Campus Medianeira na Universidade Tecnológica Federal do Paraná. Email: duartelamperti@yahoo.com.br.

[1] Robocup Federation: http://www.robocup.org.

II. Aprendizado por Reforço

A. Processos de Decisão de Markov

Um Processo de Decisão de Markov (MDP - Markov Decision Process) é uma forma de modelar processos, na qual as transições entre estados são probabilísticas.

Uma especificação das probabilidades de resultados para cada ação em cada estado possível é chamada de modelo de transição, denotado por $T(s, a, s')$. $T(s, a, s')$ é utilizado para denotar a probabilidade de alcançar o estado s' se a ação a for executada no estado s [2].

Um MDP é definido pela quádrupla (S, A, T, R) onde [3]:

- S: é um conjunto finito de estados do ambiente;
- A: é um conjunto finito de ações que o agente pode realizar;
- $T : S \times A \rightarrow \Pi(S)$: é a função de transição de estado, em que $\Pi(S)$ é uma função de probabilidades sobre o conjunto de estados S. $T(s_t, a_t, s_{t+1})$ define a probabilidade de realizar a transição do estado s_t para o estado s_{t+1} quando se executa a ação a_t.
- $R : S \times A \rightarrow R$: é a função de recompensa, que especifica a tarefa do agente, definindo a recompensa recebida (ou o custo esperado), ao longo do tempo.

Resolver um MDP consiste em computar a política π: $S \times A$ que maximiza (ou minimiza) alguma função, geralmente a recompensa recebida (ou o custo esperado), ao longo do tempo [3].

A técnica de Aprendizagem por Reforço é fundamentada nos Processos de Decisão de Markov.

B. Algoritmo Q-learning

O Q-learning (Algoritmo 1), proposto por Watkins [1], é um método de aprendizado por reforço usado para propósitos de controle. Esse algoritmo foi minuciosamente estudado e tem prova de convergência bem estabelecida [3]. A ideia básica do Q-learning é que o algoritmo aprende uma função de avaliação ótima sobre todo o espaço de pares estado-ação $S \times A$. Quando a função ótima Q for aprendida, o agente saberá qual ação resultará na maior recompensa em uma situação particular futura [19].

A função $Q(s, a)$ de recompensa futura esperada ao se escolher a ação a no estado s, é aprendida por meio de tentativas e erros segundo a equação (1):

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

em que α é a taxa de aprendizagem, r_t é a recompensa, resultante de tomar a ação a no estado s, γ é o fator de desconto e o termo $V_t(s_{t+1}) = max_a Q(s_{t+1}, a_t)$ é a utilidade do estado s resultante da ação a, obtida utilizando a função Q que foi aprendida até o presente [19].

Em um sistema de AR, uma política representa um conjunto de ações e estados que levam ao objetivo final. Dessa forma, o processo de aprendizado por reforço, pode ser expresso em termos de convergência até uma política que conduz o problema de forma ótima [7].

Para cada s,a inicialize ***Q(s,a)***=0 ;
Observe ***s*** ;
while *o critério de parada não seja satisfatório* **do**
  Selecione a ação ***a*** usando a política de ações ε-gulosa ;
  Execute a ação ***a*** ;
  Receba a recompensa imediata ***r(s,a)*** ;
  Observe o novo estado ***s'*** ;
  Atualize o item ***Q(s,a)*** de acordo com a equação (1) ;
  s $\leftarrow$ ***s'*** ;
end

Algoritmo 1: Forma procedimental do algoritmo Q-learning.

III. Regressão Logística

Métodos de regressão tornaram-se ferramentas importantes para qualquer análise de dados em que se tem a preocupação de descrever a relação entre uma variável resposta (também chamada de variável dependente) e uma ou mais variáveis explicativas (também conhecidas com variáveis independentes). O exemplo mais comum de modelagem é o uso do modelo de regressão linear onde a variável resposta é assumida ser contínua. Por outro lado, é comum termos uma variável dependente discreta, assumindo dois ou mais valores possíveis. Nesses casos, o método padrão de análise é a regressão logística [18].

O que distingue um modelo de regressão logístico de um modelo de regressão linear é que a variável resposta no modelo logístico é binária ou dicotômica. Essa diferença entre a regressão logística e a linear afeta tanto a escolha do modelo paramétrico quanto nas suposições do modelo.

Em qualquer problema de regressão, a quantidade chave é o valor médio da variável resposta dado os valores das variáveis independentes. Essa quantidade é chamada de média condicional e é expressa por $E(Y/x)$, em que Y denota a variável resposta e x denota o valor da variável independente. Em regressão linear assumi-se que essa média pode ser expressa como uma equação linear em x (ou uma transformação de x ou Y), da seguinte forma:

$$E(Y/x) = \beta_0 + \beta_1 x. \quad (2)$$

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:

$$E(Y/x) = \pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}. \quad (3)$$

Uma transformação de $\pi(x)$ que é fundamental em regressão logística é a transformação logito. Essa transformação é definida, em termos de $\pi(x)$, como:

$$g(x) = ln\left(\frac{\pi(x)}{1 - \pi(x)}\right) = \beta_0 + \beta_1 x. \quad (4)$$

A importância dessa transformação é que $g(x)$ possui muitas das propriedades desejadas do modelo de regressão linear. O logito, $g(x)$, é linear em seus parâmetros, pode ser contínuo

tendo uma amplitude de $-\infty$ e $+\infty$ dependendo da amplitude de valores de x. Maiores detalhes do modelo de regressão logística podem ser obtidas em [18].

IV. Futebol de Robôs Simulado

A categoria de simulação 2D da RoboCup simula partidas de futebol de robôs autônomos. O simulador (Figura 1) apresenta características de um ambiente dinâmico, ruidoso, cooperativo e coordenado [20]. Nesta liga não existem robôs/agentes físicos, todo o ambiente e os agentes são virtuais. O simulador fornece aos agentes todos os dados que são obtidos na realidade por meio de sensores e calcula o resultado das ações de cada agente. A simulação 2D permite definir as estratégias de jogo para um time de robôs formado por doze agentes robôs.

Cada partida realizada tem duração de aproximadamente dez minutos, valor este correspondente a seis mil ciclos de simulação, sendo separado em dois tempos de aproximadamente cinco minutos [21].

O UaiSoccer2D é um time de futebol de robôs simulado da UFSJ, vice-campeão na Competição Brasileira de Robótica em 2013. O UaiSoccer2D que é analisado neste trabalho, é resultado de implementações e modificações na estrutura do Agent2D (Helios base) [22].

O time Helios [22] do Japão foi o campeão da categoria de simulação 2D da RoboCup em 2010 e 2012. Os desenvolvedores resolveram disponibilizar na *Internet* o código base do time. O código base do Helios (Agent2D) foi escolhido por ter funções básicas como interceptação de bola, chute, drible e movimentação. Além disso, já oferece implementado a conexão com o RoboCup Soccer Server, servidor de execução dos jogos 2D da Robocup, criada através de um *socket*, que possibilita enviar e receber mensagens.

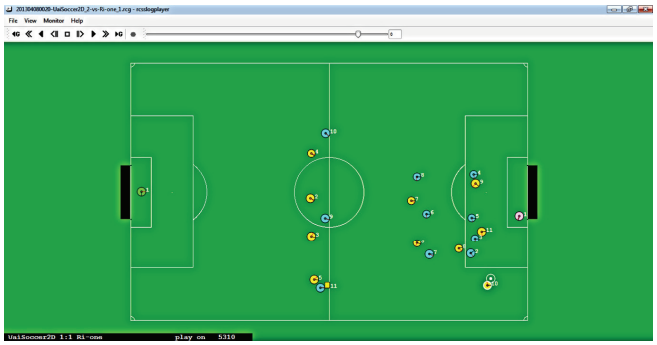

Fig. 1. Imagem do monitor do simulador RcSoccerSim da RoboCup.

V. Modelagem do Sistema de Aprendizado por Reforço

Procurou-se com este trabalho avaliar a convergência do algoritmo Q-learning aplicado em um time de futebol de robôs simulado. Para isso, foi adotado o modelo do sistema de aprendizado por reforço (ações, estados, recompensas) proposto e descrito no trabalho [17]. Vale ressaltar que, o modelo proposto é válido para o aprendizado da tomada de decisão com posse de bola. As ações sem posse de bola foram mantidas na mesma estrutura de programação do código base do Agent2D. Ou seja, em ações sem bola não existe atuação do aprendizado, como posicionamento, movimentação e marcação, e o time mantém um comportamento fixo.

A metodologia adotada para o desenvolvimento da estratégia de aprendizagem é dividida em quatro etapas:

1) Definição do conjunto finito de ações que os agentes podem realizar;
2) Definição do conjunto finito de estados do ambiente, no qual, os agentes estão inseridos;
3) Definição dos valores dos reforços, para cada par Estado (S) X Ação (A);
4) Aplicação do algoritmo de aprendizado por reforço Q-learning ao time de futebol de robôs na plataforma de simulação 2D da Robocup.

A. Definição das Ações

Nesta etapa são definidas as possíveis ações de um agente no campo de futebol de robôs simulado 2D. As ações abaixo são apenas para o agente com posse de bola.

1) Ação: Drible A (Carregar a bola em direção ao gol com drible A);
2) Ação: Drible B (Carregar a bola em direção ao gol com drible B);
3) Ação: Passe A (Tocar a bola para um companheiro com tipo de passe A);
4) Ação: Passe B (Tocar a bola para um companheiro com tipo de passe B);
5) Ação: Lançamento de bola.
6) Ação: Chute (Chutar a bola em direção ao gol).

B. Definição dos Estados

A interação dos agentes com o mundo virtual é interpretado por meio dos estados do ambiente. Nesses estados são definidas as características do ambiente durante uma partida de futebol de robôs. As características levadas em consideração são o posicionamento dos robôs da própria equipe com a posse da bola no plano (X, Y) do campo e a distância dos adversários.

As Figuras 2 e 3 apresentam o sistema de eixos (X, Y) na plataforma de futebol de robôs simulado da Robocup. O centro do campo corresponde ao ponto (X=0, Y=0). O eixo X varia de -52,5 a 52,5 de uma extremidade a outra na horizontal e Y de -34 a 34 na vertical.

Para caracterizar o ambiente de atuação dos agentes, o campo de jogo é dividido em cinco zonas. Cada zona, por sua vez, possui três células, totalizando quinze células no ambiente. As coordenadas X e Y do campo são utilizadas para a definição de cada um desses trechos. Essa estrutura é mostrada na Figura 4.

Outra informação levada em consideração para definição do estado do agente com bola é a distância do adversário mais próximo (*dist*). Nesse caso, para *dist* menor que quatro é dito que o adversário está *próximo*. Caso contrário, o adversário está *distante*. É adotado o valor de 4 unidades considerando essa distância igual a soma do diâmetro de dois robôs, visto que, o raio de um agente é próximo de 1 unidade.

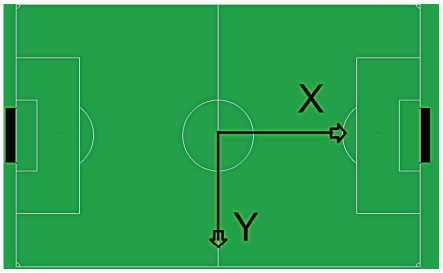

Fig. 2. Sistemas de coordenadas X,Y do campo de futebol simulado 2D para o time que está atacando para a direita.

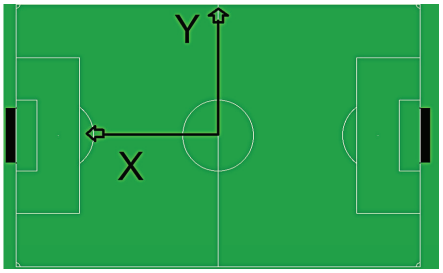

Fig. 3. Sistemas de coordenadas X,Y do campo de futebol simulado 2D para o time que está atacando para a esquerda.

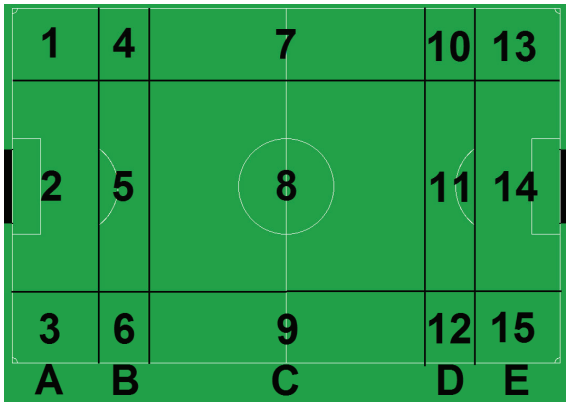

Fig. 4. Esquema de divisão proposto em zonas e células do campo de futebol de robôs simulado. A estrutura é válida para o time atacando da esquerda para a direita.

C. Definição da Matriz de Recompensas Imediatas

O ambiente do futebol de robôs simulado envolve uma grande complexidade, em termos de número de ações, para que o time de robôs alcance a recompensa principal ao marcar um gol. Um método comum, usado originalmente no treinamento de animais, é chamado de modelagem de recompensa, no qual, fornece recompensas adicionais por ”progressos feitos” [2]. Dessa forma, o objetivo de ”marcar um gol” pode ser desmembrado em ”obter posse de bola”, ”driblar em direção à meta” e ”chutar em direção ao gol”. Reforços intermediários são importantes para acelerar o aprendizado, no entanto, esses reforços devem ter valores inferiores àquele recebido quando o robô atinge o alvo [5].

A partir disso, ao definir as recompensas imediatas, o objetivo é valorizar cada passo necessário para que o time de robôs marque um gol. Ou seja, o objetivo é que o time aprenda uma estratégia de jogo visando um comportamento ofensivo com posse de bola. Essa abordagem é distinta de usar reforços somente quando há gols (recompensas) ou perda de bola (penalizações). Para isso, as recompensas são propostas para aumentarem de valor à medida que o time avance as zonas de divisão do campo, em busca da Zona E e Célula 14. Nesse trecho do campo, o agente estará mais próximo de cumprir a meta de marcar um gol. Dessa forma, para cada Zona é definido um valor de penalidade e um valor de reforço. A penalidade corresponde a um número inferior ao reforço na Zona. Isso porque, o valor de reforço é destinado a execução da ação ”correta” na Zona. No caso da célula 14, a ação ”correta” escolhida é o chute.

A Tabela I apresenta as penalidades e reforços definidos para cada Zona do campo. Vale notar que o valor do reforço aumenta à medida que o agente com posse de bola está mais próximo do gol adversário (Zona E e Célula14).

TABLE I. VALORES DE REFORÇOS E PENALIDADES PARA CADA ZONA DO CAMPO.

Zona	Penalidade	Reforço
A	-10	-1
B	-1	0
C	0	1
D	1	10
E	10	20
E (Célula14)	10	40

VI. EXPERIMENTOS REALIZADOS

Na etapa experimental foram simuladas 600 partidas de futebol de robôs na plataforma de RcSoccerSim da RoboCup. A simulação foi feita adotando o time UaiSoccer2D (UFSJ), vice-campeão brasileiro da categoria de simulação 2D da RoboCup . O código do UaiSoccer2D foi adapatado para realizar o experimento com o aprendizado por reforço. Já o adversário escolhido foi o time Ri-one (Japão). Ao final de cada partida foi armazenado o resultado final do jogo (vitória ou não vitória) e número de toques na bola de cada robô (jogador) na partida.

As simulações foram divididas em três experimentos, sendo que cada um refere-se a uma repetição do algoritmo de aprendizado por reforço:

- Experimento 1: 1ª repetição do algoritmo Q-learning (200 simulações);
- Experimento 2: 2ª repetição do algoritmo Q-learning (200 simulações);
- Experimento 3: 3ª repetição do algoritmo Q-learning (200 simulações).

Vale ressaltar que, no início de cada uma das repetições não havia aprendizado acumulado pelos agentes, ou seja, o sistema de AR adquiriu conhecimento ao longo dos jogos do experimento.

Os parâmetros do algoritmo Q-learning, a taxa de aprendizagem (α) e o fator de desconto (γ) foram fixados em 0,125 e 0,9 respectivamente. Esse valor para γ é o mesmo utilizado por [3], [11]. Já a definição de α = 0,125 foi baseada nos bons resultados de [12], [11].

Neste trabalho foram ajustados oitos modelos de regressão logística binária usando o pacote estatístico MINITAB[2] 14 (Versão Acadêmica). Sendo que, o desfecho de cada modelo logístico indica a probabilidade de vitória de acordo com o número de toques na bola efetuados pelos agentes do time. Dessa forma, cada modelo proposto é descrito como:

- Variável Dependente (VD): Resultado do jogo (1: Vitória e 0: Derrota/Empate);
- Variável Independente (VI): Soma de toques na bola de todos os jogadores do time.

A descrição do agrupamento dos dados para cada modelo logístico é mostrada na Tabela II, em que o total de simulações (n) em cada modelo logístico equivale a soma da amostra de jogos analisados em cada repetição. Por exemplo, o modelo 1 foi ajustado a partir dos dados das 25 simulações iniciais de cada uma das três repetições, totalizando n = 75. Já a coluna Etapa, faz referência ao tempo de treinamento do algoritmo de AR.

TABLE II. AGRUPAMENTO DOS DADOS PARA CADA MODELO LOGÍSTICO.

Modelo	Etapa	Simulações por Experimento	Total de Simulações
1	Inicial	1 até 25	75
2	Inicial	1 até 50	150
3	Inicial	1 até 75	225
4	Intermediária	1 até 100	300
5	Intermediária	1 até 125	375
6	Intermediária	1 até 150	450
7	Final	1 até 175	525
8	Final	1 até 200	600

VII. RESULTADOS

Os modelos de regressão logística ajustados para os dados experimentais são apresentados na Tabela III. Neste estudo, os valores do coeficiente (β_1) e p-valor (em inglês, *p-value*) em cada modelo são analisados e comparados a fim de comprovar a convergência do AR ao longo do processo de simulação.

Foram aplicadas as seguintes hipóteses sobre os modelos de regressão logística:

- H_0: não é válido afirmar que a soma de toques na bola dos jogadores do time (VI) influencia no resultado final da partida (VD).

[2]http://www.minitab.com

TABLE III. MODELOS DE REGRESSÃO LOGÍSTICA.

Modelo	β_0	β_1	p-valor
1	-0,60578	-0,0043498	0,463
2	-2,45881	0,0021671	0,543
3	-2,35237	0,0019863	0,450
4	-3,41161	0,0049826	0,016
5	-3,06561	0,0042678	0,012
6	-2,85671	0,0038310	0,004
7	-3,19468	0,0045961	0,000
8	-3,28458	0,0049335	0,000

- H_a: a soma de toques na bola dos jogadores do time (VI) influencia no resultado final da partida (VD).

Dessa forma, adotando efeitos significativos a um nível de confiança de 95%, se o p-valor for menor ou igual a 0,05, indica que deve-se rejeitar H_0 e, consequentemente, aceitar H_a. Porém, se o p-valor for maior que 0,05, deve-se aceitar H_0.

Os modelos de 1 à 3 apresentam p-valor>0,05, ou seja, à nível de significância de 5% não é válido afirmar estatisticamente que o número toques da bola influencia no resultado final da partida. Já os modelos de 4 à 8 apresentam p-valor<0,05, ou seja, nestes casos a VI influencia em VD.

Dessa forma, os modelos com p>0,05, representam apenas as simulações em período inicial, onde ainda o sistema de AR não esboçava convergência. Como o algoritmo de aprendizado por reforço necessita de um certo número de episódios de treinamento para encontrar sua estabilização, a partir do modelo 4, pode-se observar o início da convergência do sistema. Na Figura 5 - (a), pode-se notar este fato para os valores de p-valor.

Observando também o valor do coeficiente (β_1) em cada modelo, é possível tirar conclusões semelhantes quanto a convergência do sistema. Na Figura 5 - (b), nota-se que β_1 tende a ficar estável a partir do modelo 4.

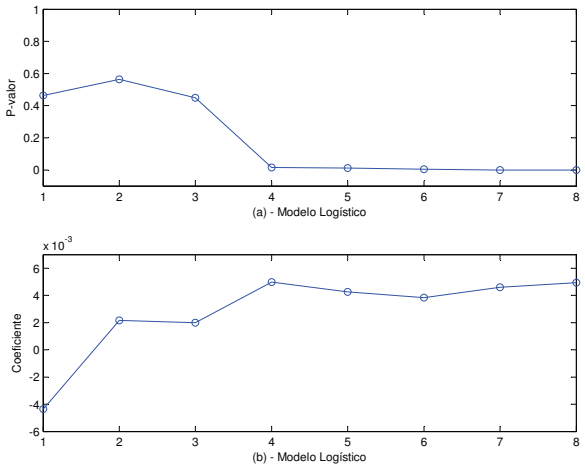

Fig. 5. (a) Número do modelo logístico versus o seu de p-valor. (b) Número do modelo logístico versus o valor ajustado para o coeficiente (β_1).

VIII. Conclusão

Este trabalho teve como objetivo analisar a convergência de um sistema de aprendizado por reforço via modelos de regressão logística. Como estudo de caso foi abordado o futebol de robôs simulado da RoboCup. Além disso, a modelagem do sistema de AR (ações, estados e recompensas) é a mesma descrita pelos autores no trabalho [17], publicado no CBIC 2013 (Congresso Brasileiro de Inteligência Computacional)[3].

Para o estudo, foram ajustados 8 modelos de regressão logística, adotando como variável independente, a soma de toques na bola e como variável dependente, o resultado final do jogo (1: vitória ou 0: não vitória).

Os resultados mostraram que a partir do modelo 4, os valores para p-valor e β_1 tendem a um equilíbrio. Ou seja, os modelos de 4 à 8, mostram uma estabilização do AR entre as simulações de número 100 e 200 de cada repetição. Já para os modelos de 1 à 3, não é possível afirmar a influencia da VI em VD, a um nível de confiança de 95%.

Nos próximos trabalhos, será analisado o desempenho do aprendizado por reforço aplicado no domínio da navegação autônoma para desvios de obstáculos. Além disso, pretende-se desafiar o Q-learning na resolução de problemas de otimização combinatória, como o Problema do Caixeiro Viajante. Dessa forma, a técnica de análise via modelos de regressão logística será abordada e aprimorada para outros estudos de casos.

Vale ressaltar que este artigo contempla uma sequência dos estudos realizados por estes autores envolvendo aprendizado por reforço e análise estatística. Em trabalhos anteriores, estes autores estudaram a aplicação e comparação de métodos de aprendizado por reforço para a tomada de decisão multiagente [23], [17]. Além disso, foram estudadas metodologias estatísticas úteis na análise do desempenho do AR [14]. Recentemente, estes autores adotaram os modelos de regressão logística para o estudo da convergência do AR no ambiente de futebol de robôs simulado [24], [25]. Esses dois últimos trabalhos publicados em formato de resumo nos Anais do Simpósio Nacional de Probabilidade e Estatística 2014.

Agradecimentos

Agradecemos à FAPEMIG, CAPES, CNPQ, UFSJ e UTFPR pelo apoio.

Referências

[1] C. J. Watkins and P. Dayan. "Technical note Q-learning". *Machine Learning*, 1992.

[2] S. J. Russell and P. Norving. *Inteligência Artificial*. Campus, second edition, 2004.

[3] R. A. C. Bianchi. "Uso de Heurística para a aceleração do aprendizado por reforço." Master's thesis, Tese (Doutorado) Escola Politécnica da Universidade de São Paulo, 2004.

[4] C. J. Watkins. "Models of Delayed Reinforcement Learning". Master's thesis, PhD thesis, Psychology Department, Cambridge University, Cambridge, United Kingdom., 1989.

[5] A. H. P. Selvatici and A. H. R. Costa. "Aprendizado da coordenação de comportamentos primitivos para robôs móveis". *Revista Controle & Automação*, vol. Vol.18 no.2, 2007.

[6] G. A. Oliveira. "Uma aplicação da aprendizagem por reforço na otimização da produção em um campo de petróleo". Master's thesis, Universidade Federal do Rio Grande do Norte, 2010.

[7] D. P. Alves. "Modelagem de Aprendizagem por Reforço e Controle em Nível Meta para melhorar a Performance da Comunicação em Gerência de Tráfego Aéreo". Master's thesis, Universidade de Brasília, 2006.

[8] L. A. Scárdua, J. J. Cruz and A. H. R. Costa. "Controle Ótimo de Descarregadores de Navios Utilizando Aprendizado por Reforço". *Revista Controle & Automação*, vol. Vol.14 no.4, 2003.

[9] J. S. Waskow. "Aprendizado por Reforço utilizando Tile Coding em Cenários Multiagente". Master's thesis, Universidade Federal do Rio Grande do Sul, 2010.

[10] L. A. Celiberto Jr and R. A. C. Bianchi. "Aprendizado por Reforço Acelerado por Heurística para um Sistema Multi-Agentes". *3rd Workshop on MSc dissertations and PhD thesis in Artificial Intelligence*, 2006.

[11] L. A. Celiberto Jr. "Aprendizado por Reforço Acelerado por Heurísticas no Domínio do Futebol de Robôs Simulado". Master's thesis, Centro Universitário da FEI, 2007.

[12] P. Stone, R. S. Sutton and G. Kuhlmann. "Reinforcement Learning for RoboCup-Soccer Keepaway". *Adaptive Behavior*, vol. 13, no. 3, pp. 165–188, 2005.

[13] A. L. C. Ottoni, R. D. Lamperti, E. G. Nepomuceno, M. S. Oliveira and F. F. Oliveira. "Modelagem e Simulação de um Sistema de Aprendizado de Reforço para Robôs". *VIII Encontro Mineiro de Engenharia de Produção, ISSN 1983 - 0629*, 2012.

[14] A. L. C. Ottoni, R. D. Lamperti, E. G. Nepomuceno and M. S. Oliveira. "Desenvolvimento de um sistema de aprendizado por reforço para times de robôs - Uma análise de desempenho por meio de testes estatísticos". *XIX Congresso Brasileiro de Automática, ISBN 978-85-8001-069-5*, pp. 3557–3564, 2012.

[15] J. R. F. Neri, C. H. F. Santos and J. A. Fabro. "Descrição Do Time GPR-2D 2011". *Competição Brasileira de Robótica*, vol. 2011, 2011.

[16] A. L. C. Ottoni, E. G. Nepomuceno, F. F. Oliveira and M. S. Oliveira. "Análise do comportamento de sistemas multiagentes cooperativos por meio de testes estatísticos". *X Encontro Mineiro de Estatística*, 2011.

[17] A. L. C. Ottoni, E. G. Nepomuceno, M. S. Oliveira and R. D. Lamperti. "Análise do Aprendizado por Reforço Aplicado a Otimização em Tomadas de Decisões Multiagente". *1st BRICS Countries Congress (BRICS-CCI) and 11th Brazilian Congress on Computational Intelligence (CBIC)*, 2013.

[18] D. W. Hosmer, S. Lemeshow and R. X. Sturdivant. *Applied Logistic Regression*. Third Edition, New York: John Wiley & Sons, 2013.

[19] S. T. Monteiro and C. H. C. Ribeiro. "Desempenho de Algoritmos de Aprendizagem por Reforço sob Condições de Ambiguidade Sensorial em Robótica Móvel". *Revista Controle & Automação*, vol. Vol.15 no.3, 2004.

[20] E. S. Fraccaroli. *Análise de Desempenho de Algoritmos Evolutivos no Domínio do Futebol de Robôs*. Dissertação apresentada à Escola de Engenharia de São Carlos da Universidade de São Paulo, 2010.

[21] E. S. Fraccaroli and P. M. Carlson. "Time GEARSIM 2010 da categoria Robocup Simulation 2D". *In Competição Latino Americana de Robótica*, vol. 2010, 2010.

[22] H. Akiyama, H. Shimora, T. Nakashima, Y. Narimoto and T. Okayama. "HELIOS2011 Team Description". *Robocup 2011*, 2011.

[23] R. D. Lamperti, E. G. Nepomuceno and A. L. C. Ottoni. "Aprendizado por Reforço no Dominio do Futebol de Robôs 2D". *XI SBAI - Simpósio Brasileiro de Automação Inteligente*, 2013.

[24] A. L. C. Ottoni and M. Oliveira. "Regressão Logística Aplicada na Análise do Aprendizado por Reforço". *21º SINAPE - Simpósio Nacional de Probabilidade e Estatística*, 2014.

[25] A. L. C. Ottoni, M. S. Oliveira and L. T. Cordeiro. "Regressão Logística Aplicada na Análise Comparativa entre Times de Futebol de Robôs". *21º SINAPE - Simpósio Nacional de Probabilidade e Estatística*, 2014.

[3] BRICs e CBIC 2013: http://www.brics-cci.org.