APRENDIZAGEM POR REFORÇO EM REDES NEURAIS MULTICAMADAS APLICADAS EM CONTROLE AUTÔNOMO

FRANK BORSATO E MAURÍCIO FIGUEIREDO

Departamento de Computação, Universidade Federal de São Carlos
13565-905 - São Carlos - SP

E-mails: frankhelbert@yahoo.com.br, firedo@dc.ufscar.br

Abstract—A neural system is described. It is potentially capable for autonomous control applications. Psychology and Neurophysiology areas furnish bases for the system design. The architecture consists of three neural modules: basic behavior generation, learning management and input-output mapping. Learning is based on the conditioning theory. Synaptic weight adjustment is possible for internal layers (input-output mapping network). An autonomous control application is adopted to appraise the capabilities of the system. Simulation results confirm the good expectations: knowledge acquisition from environment interactions.

Keywords— autonomous intelligent systems, unsupervised neural networks, reinforcement learning, autonomous control.

Resumo— Descreve-se um sistema neural potencialmente hábil para aplicações associadas a controle autônomo. Recursos da Psicologia do Comportamento e da Neurofisiologia estabelecem as bases da concepção do sistema. A arquitetura consiste de três repertórios neurais: geração de comportamentos básicos, gerenciamento de aprendizagem e mapeamento entrada-saída. A aprendizagem está baseada na teoria do condicionamento e permite o ajuste dos pesos sinápticos em qualquer camada (rede de mapeamento entrada-saída). As características do sistema são apreciadas quando aplicado a um problema de controle autônomo. Resultados de simulação confirmam a capacidade de aquisição de conhecimento a partir da interação com o ambiente.

Palavras-chave— sistemas autônomos inteligentes, redes neurais não-supervisionadas, aprendizagem por reforço, controle autônomo.

1 Introdução

Redes neurais compõem uma das mais cativantes áreas da Inteligência Computacional [1]. Diversos aspectos podem explicar seu forte apelo, muito certamente pela associação com as intrigantes características da contra-parte biológica, e.g., processamento paralelo.

Esforços têm sido conduzidos no sentido de alcançar suporte teórico para projetos de alta complexidade. Neste sentido algumas propostas vão muito além dos modelos tradicionais, entre outras: modelagem de código temporal em redes pulsadas [2], processamento paralelo em modelos dinâmicos que não obedecem às condições de unicidade [3] e; dinâmicas caóticas [4] e auto-realimentação em modelos de neurônios [5] visando alcançar características avançadas de memória. Particularmente, no contexto dos sistemas autônomos inteligentes, investigações com base na teoria do Condicionamento e na Neurofisiologia são de forte interesse [6] [7] [8].

Não sem motivo a estratégia de aprendizagem por reforço tem despertado atenção. Entendendo que a autonomia se refere à capacidade de aquisição de habilidades cognitivas sem intervenção externa, observe-se que aprendizagem supervisionada e aprendizagem auto-organizada (as alternativas possíveis) não são convenientes para uma classe ampla de aplicações, a saber, controle autônomo (e.g., navegação autônoma de robôs [9]). No caso da aprendizagem supervisionada a dissonância provém da ausência de modelos cognitivos (eventualmente, até pelo fato do ambiente ser desconhecido), fator crítico e essencial para a estratégia. A segunda alternativa, muito embora prescindindo de modelos cognitivos, é inviável por não estabelecer um mapeamento entrada-saída (fundamental para aplicações em controle).

Apesar da relevância evidente associada à aprendizagem por reforço, a estratégia não se encontra plenamente ajustada às redes neurais [10, 11, 12, 13]. Felizmente a Psicologia do Comportamento e a Neurofisiologia possibilitam indícios importantes para o seu desenvolvimento.

Este artigo descreve uma nova classe de redes neurais, resultante da exploração de tais áreas, tendo como objetivo a concepção de sistemas autônomos inteligentes, ou seja, redes neurais capazes de aprender a partir de sua interação com o ambiente em que atua. A fase de concepção também teve como orientação o objetivo específico de tornar a rede versátil, ou seja, capaz de aplicações distintas sem fortes exigências de alterações na arquitetura e aprendizagem, necessitando de um mínimo de conhecimento *a priori*. Os neurônios são topologicamente arranjados em camadas. A estratégia de aprendizagem segue os princípios da aprendizagem por reforço clássica (baseada na Psicologia do Comportamento). Modelos da Neurofisiologia são adotados para compor uma estrutura que suporte os mecanismos de ajuste sináptico, incluindo as neuromoléculas, seus atributos e processos de dispersão. Um modelo dinâmico de neurônio é definido segundo diferentes modos de operação, dependentes de classes de estímulos e valores de parâmetros. Resultados de simulação confirmam, de um lado, a reprodução de comportamentos bem descritos pela Psicologia, no contexto do condicionamento operante; de outro, as expectativas associadas à autonomia cognitiva, ou seja, o sistema demonstra potencialidades para aplicações em controle autônomo.

O restante do artigo está organizado conforme segue. A Seção 2 oferece um texto básico direcionado para a aprendizagem por reforço e suas bases biológicas. A Seção 3 descreve a arquitetura e aprendizagem do sistema. Resultados de experimentos acompanhados de breves análises são encontrados na Seção 4. A última seção dedica-se às conclusões e futuras propostas de pesquisa.

2 Aprendizagem por Reforço

2.1 Introdução

Duas classes de estratégias de aprendizagem têm sido bem descritas no contexto da teoria de redes neurais: não-supervisionadas e supervisionadas.

Aprendizagem por reforço pode ser considerada um caso especial de aprendizagem não-supervisionada, pois torna o sistema capaz de adquirir conhecimento sem apoio externo (que disponibilize um conjunto de pares entrada-saída). A aprendizagem se faz a partir de seleção de alternativas. Cada resposta da rede é imediata ou remotamente associada a um valor de desempenho (associação não exigente de auxílio externo). O acúmulo de experiências acaba por gerar um conjunto de alternativas mais adequadas às expectativas. Assim, sistemas inteligentes podem adquirir conhecimento exclusivamente a partir de sua interação com o ambiente. Esta capacidade é essencial quando não há fonte de conhecimento disponível (inclusive, modelos cognitivos), e.g. exploração espacial ou submarina. Já, as estratégias supervisionada e auto-organizada não são, isoladas, adequadas nestes casos.

2.2 Condicionamento Operante

A aprendizagem por reforço apresenta atualmente duas abordagens distintas: moderna e clássica. A primeira está associada a problemas de controle ótimo, com soluções obtidas via mecanismos similares à programação dinâmica [11]. A aprendizagem clássica restringe-se, de uma forma geral, a modelos baseados na Psicologia do Comportamento, explicando a aprendizagem de sistemas biológicos via processos de condicionamento.

A teoria do condicionamento é baseada no comportamento animal. Seus princípios são bem conhecidos e verificados a partir de experimentos controlados. Das duas classes de condicionamento, a saber, operante e respondente, somente a primeira é de interesse no trabalho.

O condicionamento operante pode ser primeiramente explicado pela Lei do Efeito: a associação entre estímulo e resposta é afetada pela conseqüência gerada pelo comportamento [14]. De um lado o mecanismo requer um estímulo (reforçador) associado a algum valor (e.g., hedonístico, no caso de sistemas biológicos). De outro, requer uma resposta (reforçada) que é a ação que produz o reforçador. O condicionamento é totalmente voluntário, sendo possível somente se a resposta reforçada é emitida.

Antes de o sistema iniciar o condicionamento para um reforçador específico, encontra-se em nível operante. O nível operante é importante tanto para a medida da aprendizagem (permitindo comparações com a freqüência das respostas após o condicionamento) quanto para a modelagem do sistema em si (veja seções seguintes).

2.3 Reforçadores de Alta Ordem

Reforçadores podem ser de dois tipos: adquiridos ou inatos. Antes de qualquer processo de aprendizagem, somente reforçadores inatos são identificados pelo sistema nervoso (podem eliciar respostas bem definidas). Reforçadores adquiridos são formados ao longo do processo de aprendizagem. Um estímulo preliminarmente neutro adquire a característica de reforçador; especificamente, reforçador adquirido; se se tornar associado a um reforçador inato. Tal associação ocorre se o estímulo neutro elicia uma resposta (resposta condicionada) que por sua vez gera o reforçador inato. Reforçadores adquiridos também podem ser formados a partir da associação entre um estímulo neutro e um outro reforçador adquirido. Neste caso, o reforçador adquirido é de segunda ordem. Assim, de forma idêntica, reforçadores de ordem superior podem ser definidos, sempre por conta da associação com um reforçador adquirido. No processo de formação de reforçadores adquiridos de segunda ordem ou de ordem superior, sempre o estímulo neutro é seguido da emissão consecutiva de reforçadores adquiridos, culminando com a emissão do reforçador inato.

2.4 Mecanismos biológicos

Duas áreas do sistema nervoso humano são importantes na modelagem do sistema neural: cortex frontal de associação (FAC) e área tegmental ventral (VTA). Tais áreas são parte do suporte biológico ao processo de condicionamento. A FAC associa estímulos a respostas. Antes da aprendizagem, associações casuais são devidas a fracas conexões inatas entre neurônios. A aprendizagem tem como efeito biológico o fortalecimento de conexões correspondentes aos comportamentos reforçados [5]. Entretanto a área FAC não é capaz de gerenciar o fortalecimento de suas sinapses neste processo. Esta capacidade é atribuída à área VTA, que de forma difusa projeta conexões sobre a FAC. Tais conexões desprendem neuromoduladores dopamina responsáveis pela consolidação de conexões entre neurônios [15].

3 A rede neural

A rede neural proposta reproduz qualitativamente algumas das estruturas biológicas associadas ao condicionamento e identificadas pela Neurofisiologia [16]. O modelo consiste de três repertórios neurais: rede de condicionamento ou rede de mapeamento entrada-saída (CN), rede de comportamentos básicos (IBN) e rede de regulação (RN); sendo o primeiro correspondente à FAC e os demais à VTA (Figura1).

3.1 Rede de Condicionamento (CN)

Camadas de neurônios topologicamente arranjados em toróide compõem a estrutura básica da rede de condicionamento. A primeira camada (camada de entrada) recebe estímulos do ambiente, enquanto a última camada (camada de saída) define respostas correspondentes a ações aplicadas sobre o ambiente. As demais camadas, internas, estabelecem associações entre estímulos e respostas, compondo um mapeamento entrada-saída.

Os neurônios estabelecem três tipos de sinapses: excitatórias intercamadas, excitatórias intracamadas e inibitórias intracamadas. As excitatórias intercamadas

conectam neurônios de camadas sucessivas, de forma que cada neurônio pré-sináptico estabelece conexões segundo uma distribuição Gaussiana com média na mesma posição relativa do neurônio pré-sináptico. As sinapses intracamadas também seguem a mesma estratégia de distribuição Gaussiana; mas, para as sinapses inibitórias, as conexões são efetivas apenas para neurônios distantes (neste caso as conexões definem uma área em forma de coroa circular).

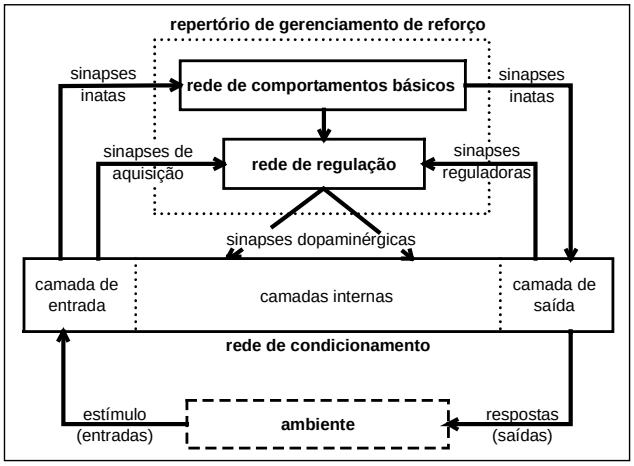

Figura 1: Diagrama de blocos da rede neural.

3.2 Repertório de gerenciamento de reforço (RMR)

As redes de comportamentos básicos (IBN) e de regulação (RN) compõem o repertório de gerenciamento de reforço.

A rede IBN gera respostas inatas (não-condicionadas, imutáveis), apenas eliciadas por estímulos particulares bem definidos (reforçadores inatos).

A rede RN controla a difusão do neuromodulador dopamina sobre a rede CN. Os repertórios CN e RMR interagem de acordo com quatro tipos de sinapses, classificadas segundo suas funções: inatas, aquisição, reguladoras e dopaminérgicas. Estímulos chegam à rede IBN por meio das sinapses inatas estabelecidas com a camada de entrada da rede CN. Diferentes sinapses inatas possibilitam que as respostas da rede IBN estimulem neurônios da camada de saída da rede CN, os quais efetivamente geram a resposta que atua sobre o ambiente.

Os estímulos captados pela camada de entrada de CN chegam à rede RN via sinapses de aquisição (excitatórias), responsáveis por definir reforçadores adquiridos. Os estímulos gerados na camada de saída de CN também alcançam RN via sinapses reguladoras (inibitórias), garantindo que um único reforçador adquirido seja emitido a partir de uma resposta particular de CN. Em um fluxo contrário, os sinais nas sinapses dopaminérgicas partem de RN e estimulam CN, mas sem influenciar na ativação dos neurônios. Diferentemente, modelam a liberação de dopamina na rede CN, correspondendo ao ajuste dos pesos sinápticos intra e inter-camadas, de acordo com a Lei de Hebb.

3.3 Raciocínio e aprendizagem na rede CN

Considere que $a(j,t) \in [0,1]$ seja a ativação do neurônio j de CN na iteração t. A atividade do neurônio na camada de entrada é definida tal como segue:

$$a(j,t) = \begin{cases} 1.0, & \text{if } E(j,t) \neq 0; \\ a(j,t-1)\varphi, & \text{if } E(j,t) = 0 \text{ e } a(j,t\text{-}1) > 0.1; \\ 0.0, & \text{caso contrário;} \end{cases} \quad (1)$$

em que: $E(j,t) \in [0,1]$ é o estímulo que capturado pelo neurônio j na iteração t, e $\varphi \in [0,1]$ é uma constante.

Nas camadas internas a atividade é definida por:

$$a(j,t) = \begin{cases} a'(j,t), & \text{if } j \in G(c,t); \\ 0.0, & \text{caso contrário;} \end{cases} \quad (2)$$

se $a'(j,t)$ e $G(c,t)$ são tais como definidos em seguida.

Para $a'(j,t)$, considere, inicialmente, que o total de estímulos excitatórios e inibitórios no neurônio j na iteração t, $exc(j,t)$ e $inh(j,t)$, respectivamente, sejam assim definidos:

$$exc(j,t) = \sum_i a(i,t).w(i,j,t) ; \quad (3)$$

$$inh(j,t) = \sum_i a(i,t).w(i,j,t) ; \quad (4)$$

em que: $w(i,j,t) \in [0,1]$ é o peso sináptico entre os neurônios pré e pós-sinápticos i e j.

Desta forma, $a'(j,t)$ é determinado tal como segue a:

$$a'(j,t) = \begin{cases} S(exc(j,t)) + \\ \quad \tau.S(exc(j,t-1))[1 - S(1 - exc(j,t))] - S(inh(j,t)), \\ \qquad \text{se } (exc(j,t) \geq \Theta(j,t) \text{ e } exc(j,t) > inh(j,t)); \\ a(j,t-1) - \kappa.a(j,t-1)[1 - a(j,t-1)], \\ \qquad \text{se } (exc(j,t) < \Theta(j,t) \text{ e } exc(j,t) > inh(j,t)); \\ 0.0, \quad \text{se } exc(j,t) \leq inh(j,t); \end{cases} \quad (5)$$

em que: $\Theta(j,t)$ é um número aleatório Gaussiano; $S(x) = 1/(1 + \exp[(-x + \gamma)/\delta])$ é a função logística; τ e $\kappa \in [0, 1]$; e γ e $\delta \in \Re$.

A definição de $G(c,t)$, conjunto de neurônios ξ que estão em algum grupo de neurônios da camada c na iteração t (um grupo de neurônios consiste de neurônios espacialmente próximos que estabelecem uma atividade colaborativa), é dada por:

$$G(c,t) = \{\xi / d(\xi, \vartheta(\mathbf{k} + \eta \mathbf{V}(c,t))) < \bar{r}_c, \mathbf{k}(k), k \in \Omega(c,t)\}; \quad (6)$$

em que:

$$\mathbf{V}(c,t) = \begin{cases} \mathbf{V}(c,t-1) + \mathbf{D}(c,t), \\ \quad \text{if } |\mathbf{V}(c,t-1) + \mathbf{D}(c,t)| \leq |\mathbf{V}(c,t-1)|; \\ \mathbf{V}(c,t-1) + [\mathbf{D}(c,t).(\rho - |\mathbf{V}(c,t-1)|)/\rho], \\ \quad \text{caso contrário;} \end{cases} \quad (7)$$

$$\mathbf{D}(c,t) = \begin{cases} \mathbf{F}_{tend}(t), & \text{se } c = 1; \\ \mathbf{V}(c-1,t).\zeta, & \text{se } c > 1 \text{ e } |\mathbf{V}(c-1,t)| \geq 1.0; \\ (0.0,0.0), & \text{se } c > 1 \text{ e } |\mathbf{V}(c-1,t)| < 1.0; \end{cases} \quad (8)$$

$$\boldsymbol{F}^{\intercal}_{tend}(t)=\begin{pmatrix} r_2[\cos(\theta(t))-\cos(\theta(t-1))]+r_1[\cos(\Phi(t))-\cos(\Phi(t-1))] \\ r_2[\mathrm{sen}(\theta(t))-\mathrm{sen}(\theta(t-1))]+r_1[\mathrm{sen}(\Phi(t))-\mathrm{sen}(\Phi(t-1))] \end{pmatrix} \quad (9)$$

$$\Omega(c,t)=\{j/\hat{a}(j,t)>\chi,\ j\in C(c)\}; \quad (10)$$

$$\hat{a}(j,t)=\sum_{m\in R(j)}[a'(m,t)/(1+d(j,m))]; \quad (11)$$

em que: $C(c)$ é o conjunto dos neurônios da camada c; $\xi\in C(c)$; $R(j)$ é o conjunto pré-definido de neurônios próximos ao neurônio j; $d(i,j)$ é a distância Euclidiana entre os neurônios i e j; $\vartheta(.)$ retorna o neurônio mais próximo do seu argumento; $\eta=1$, se $\hat{a}(k,t)<\bar{\mu}$, caso contrário $\eta=0$; $\boldsymbol{k}(k)$ é o vetor associado à posição do neurônio k; $\Phi(t)=t.\varepsilon_1$; $\theta(t)=t.\varepsilon_2$; ζ, r_1, r_2, ε_1 e $\varepsilon_2\in$ [0,1]; e $\bar{r}_c$, χ, ρ e $\bar{\mu}\in\Re$.

Em geral a atividade da última camada também segue (2); a não ser quando a rede IBN recebe um estímulo reforçador inato. Neste caso a resposta de IBN ativa neurônios específicos da última camada de CN, produzindo a resposta instintiva que atua no ambiente.

O ajuste dos pesos sinápticos depende das atividades dos neurônios pré e pós-sinápticos (i e j) e da concentração $H(t)$ de dopamina liberada sobre CN (Equação 20), tal como segue:

$$w(i,j,t)=\begin{cases} w(i,j,t-1)+\alpha\, a(j,t)\, H(t)\, p(i,t)\, r(j,t) \text{ if } H(t)>0; \\ w(i,j,t-1)-\beta\, w(i,j,t-1)\, a(i,t)\, a(j,t) \text{ if } H(t)\le 0; \end{cases} \quad (12)$$

$$p(i,t)=\frac{a(i,t).w(i,j,t-1)}{N}; \quad (13)$$

$$r(j,t)=1-\sum_{l} w(l,j,t); \quad (14)$$

em que: N assume o valor de $exc(j,t)$ ou de $inh(j,t)$ dependendo do tipo de sinapse (excitatória ou inibitória, respectivamente); l representa qualquer neurônio conectado ao neurônio j; e α e $\beta\in$ [0, 1].

3.4 Raciocínio e aprendizagem no repertório RMR

O repertório de gerenciamento de reforço consiste da rede IBN e da rede RN (Figura 1). A rede IBN gera comportamentos entrada-saída não-condicionados (pré-definidos / inatos e imutáveis). Portanto não há aprendizagem da rede IBN, permanecendo fixos seus pesos sinápticos.

Um único neurônio representa a rede RN. Sua atividade regula a quantidade de dopamina lançada em CN. Uma composição de estímulos excitatórios e inibitórios definem o comportamento de RN. Respostas são eliciadas (com liberação de dopamina) por estímulos excitatórios originados em: IBN, se esta rede é estimulada por reforçadores inatos; ou CN, se reforçadores adquiridos chegam à RN via sinapses de aquisição (Figura 1). Assim:

$$exc_r(t)=exc_{ibn}(t)+exc_{as}(t); \quad (15)$$

$$exc_{ibn}(t)=\begin{cases} 1.0, \text{ se o estímulo é um reforçador inato;} \\ 0.0, \text{ caso contrário;} \end{cases} \quad (16)$$

$$exc_{as}(t)=\begin{cases} 1, \text{ se } \sum_{s}[a_s(s,t).w_{as}(s,t)]\ge\varsigma; \\ 0, \text{ caso contrário;} \end{cases} \quad (17)$$

em que: $exc_r(t)$, $exc_{ibn}(t)$ e $exc_{as}(t)$ representam a composição de estímulos excitatórios, estímulos de IBN e estímulos de CN, respectivamente; $a_s(s,t)$ é a atividade do neurônio pré-sináptico s (na rede CN), $w_{as}(s,t)$ é o peso sináptico entre o neurônio s e o neurônio de RN; e $\varsigma\in\Re$.

Por outro lado, estímulos provenientes da camada de saída de CN inibem RN, tal como modelado em seguida:

$$inh_r(t)=\begin{cases} \varepsilon\sum_{v}[a_v(v,t)\, w_v(v,t)]; \\ \qquad \text{se } \varepsilon\sum_{v}[a_v(v,t)\, w_v(v,t)]\le 1; \\ 1.0, \text{ caso contrário;} \end{cases} \quad (18)$$

em que: $a_v(v,t)$ é a atividade do neurônio pré-sináptico v (em CN); $w_v(v,t)$ é o peso sináptico entre o neurônio v e o neurônio de RN; e $\varepsilon\in$ [0, 1].

Assim, a atividade do neurônio de RN é definida por:

$$a_r(t)=exc_r(t)-inh_r(t); \quad (19)$$

A quantidade $H(t)$ de dopamina lançada sobre a rede CN na iteração t é definida pela atividade do neurônio de RN, tal como segue:

$$H(t)=a_r(t). \quad (20)$$

As sinapses de aquisição se convenientemente ajustadas, para efetivamente eliciar respostas em RN, passam a definir quais estímulos assumem o papel de reforçadores adquiridos (Figura 1). Os respectivos pesos sinápticos $w_{as}(s,t)$ de tais sinapses são definidos em (21):

$$w_{as}(s,t)=\begin{cases} w_{as}(s,t-1) \\ \quad +([1-w_{as}(s,t-1)]\, a_s(s,t).\hat{\upsilon}), \\ \quad \text{se } 0.0<H(t)<\Phi; \\ w_{as}(s,t-1)-w_{as}(s,t-1)\, a_s(s,t)\, \hat{\chi}, \\ \quad \text{caso contrário;} \end{cases} \quad (21)$$

em que: $\hat{\upsilon}$ e $\hat{\chi}\in$ [0, 1]; e $\Phi\in\Re$.

As sinapses reguladoras impedem que diferentes estímulos tornem-se reforçadores adquiridos após condicionamento de um mesmo reforçador (inato ou não) [6]. Os respectivos pesos sinápticos $w_v(v,t)$ são ajustados tal como segue:

$$w_v(v,t)=\begin{cases} w_v(v,t-1)+([1-w_v(v,t-1)]\, \phi a_v(v,t).), \\ \qquad \text{se } a_v(v,t)\ge\partial \text{ e } H(t)>0; \\ w_v(v,t-1)-\delta\, w_v(v,t-1), \\ \qquad \text{se } a_v(v,t)\ge\partial \text{ e } H(t)\le 0; \\ w_v(v,t-1), \text{ caso contrário;} \end{cases} \quad (22)$$

em que: ϕ e $\delta\in$ [0, 1]; e $\partial\in\Re$.

3.5 Dinâmica conjunta CN - RMR

A cada iteração a camada de entrada de CN recebe um estímulo $E(t)$, que pode pertencer a uma das três classes: reforçador inato, reforçador adquirido e dissociado (estímulo não inato para o qual não há uma resposta condicionada associada). A dinâmica devida às interações de CN e RMR a partir da chegada do estímulo é descrita em seguida.

Se $E(t)$ é um reforçador inato ou adquirido então elicia uma resposta bem definida (inata ou condicionada) estabelecida por IBN ou CN, respectivamente ($E(t)$ chega à IBN via sinapses inatas). Diferentemente, um estímulo dissociado estimula CN definindo uma dinâmica no nível operante, ou seja, neurônios em geral apresentam atividade reduzida a menos de momentos escassos sem qualquer coerência ou correlação com o estímulo.

Ainda, se $E(t)$ é um reforçador inato ou adquirido, RN é estimulada no sentido de produzir e lançar dopamina sobre CN. Se assim acontece, as seguintes classes de sinapses são ajustadas: inter e intracamadas em CN, aquisição e reguladoras. Se $E(t)$ é dissociada, RN não é estimulada, portanto não há ajuste sináptico.

4 Resultados

O problema descrito em seguida não é complexo mas satisfaz as condições necessárias para avaliar as potencialidades do sistema em dois aspectos: geração de reforçadores adquiridos e condicionamento de segunda ordem. O problema modela o ajuste de posição uma câmera de forma que o alvo de interesse "deslize" para o centro da imagem.

No experimento simulado, cinco camadas, cada qual com 20 posições por dimensão (para um total de 400 neurônios), compõem a rede CN. Cada estímulo pode ser identificado de acordo com o padrão de atividade que causa nos neurônios da camada de entrada da rede CN. Somente estímulos do tipo padrão são considerados significativos, ou seja, capazes de estimular a rede CN. São 25 os estímulos-padrão E_Z, z = 1,…, 25; cada qual formado por 4 neurônios adjacentes ativados (para uma iteração t) em cada conjunto de 16 neurônios tal como definidos na Figura 2 (que ilustra E_1). Assim, se $E(t)$ é um estímulo, a seguinte notação é valida: $E(t) = E_\Phi(t) \Leftrightarrow E(t) = E_\Phi$.

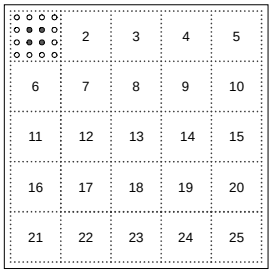

Figura 2: Conjunto de estímulos e respostas.

As respostas da rede em geral não são do tipo padrão (análogas aos estímulos-padrão). Portanto, a resposta da rede passa a ser considerada como o padrão O_Z que mais se aproxima da resposta. Desta forma, a notação adotada para os estímulos é válida para a resposta da rede $O(t)$ eliciada por $E(t)$ na iteração *t*. Além disso, embora $E(t)$ possa eliciar qualquer resposta, nem todas são viáveis (implementáveis). Para $E_\Phi(t)$, somente respostas $O_Q(t)$ mais próximas de O_Φ (mesmo índice de $E_\Phi(t)$) são viáveis. Assim, se $E_\Phi(t) = E_8$ então as respostas viáveis $O_Q(t)$ são tais que $Q \in \{2, 3, 4, 7, 8, 9, 12, 13, 14\}$ (Figura2).

O experimento apresentado em seguida consiste de várias provas, cada qual iniciada a partir de um estímulo selecionado aleatoriamente dentre os possíveis padrões E_Z, z = 1,…, 25; e finalizada caso o estímulo $E(t) = E_{13}$. Há um único reforçador (inato), emitido se $O(t) = O_{13}$.

O estímulo a cada iteração é definido tal como segue:

$$E(t+1) = \begin{cases} E(t), \text{se } O(t) \text{ não é viável;} \\ E_\Theta, \text{se } O(t) = O_\Theta(t) \text{ é viável.} \end{cases} \quad (23)$$

Portanto os estímulos acompanham as respostas viáveis.

A dinâmica esperada para o experimento pode ser descrita resumidamente tal como segue. Para cada prova iniciada na iteração inicial t_0, os seguintes passos se sucedem:

1. Seleção aleatória de $E(t_0) = E_z$; $z \in \{1, 2, ..., 25\}$;
2. Apresentação de $E(t)$ ao sistema (à CN);
3. Se $O(t)$ não é viável ou $O(t) \neq O_{13}$, passo 2, observando (23) e $t = t+1$; caso contrário, passo 4;
4. $O(t) = O_{13}$, então $E(t+1) = E_{13}$ (reforçador); ajuste dos pesos sinápticos e encerramento da prova;
5. Retorno passo 1 para início de nova prova e $t_0 = t+1$; ou encerramento do experimento.

Nos gráficos apresentados em seguida o estímulo inicial de cada prova é representado por um retângulo; reforços adquiridos, por triângulos; reforços inatos, por asteriscos (representando o fim da prova); e demais estímulos, por círculos (os estímulos são definidos na ordenada).

Em uma fase inicial do experimento não se encontram seqüências breves e bem definidas de estímulo/resposta (considerando a relação definida por (23)) tal que o estímulo inicial $E(t_0)$ é conduzido ao estímulo final E_{13} (Figura 3).

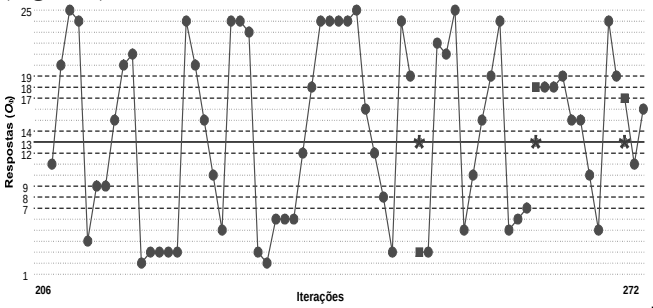

Figura 3: Desempenho do sistema: fase de exploração.

Em uma fase mais avançada do experimento é possível identificar seqüências de estímulo/resposta que rapidamente forçam o encerramento das provas, e.g., $E_7 \rightarrow E_{12} \rightarrow E_{13}$ (Figura 4). Observa-se ainda que o período de duração das provas (entre asteriscos consecutivos) é relativamente reduzido (para comparações veja Figura 3).

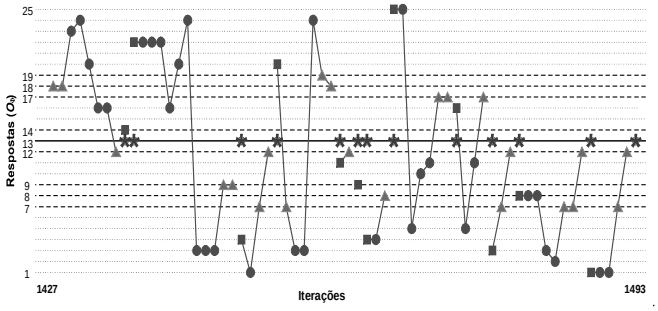

Figura 4: Desempenho do sistema: regularidade de comportamentos.

Ao longo do experimento, os estímulos vizinhos do reforçador inato tornam-se reforçadores adquiridos, confirmados após 1400 iterações (veja Figure 5; observe também triângulos na Figura 4 e na Figura 3). O número de iterações necessárias para que o sistema encerre uma prova ($E(t_0)$ conduzido a $E_{13}(t)$) é reduzido à medida que a aprendizagem se processa (Figura 6).

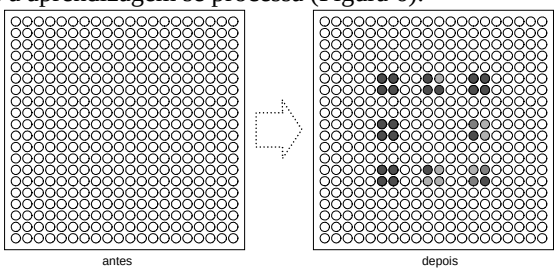

Figura 5: Sinapses de aquisição: antes e após aprendizagem (círculos correspondem às sinapses entre RN e a primeira camada de CN; quanto mais escuros, mais eficientes são as sinapses).

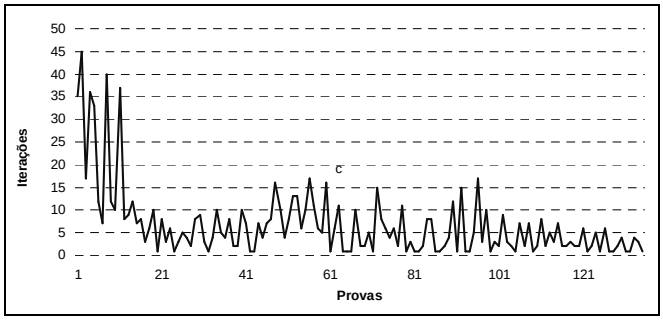

Figura 6: Número de iterações por prova.

5 Conclusões e trabalhos futuros

Autonomia cognitiva permite que sistemas inteligentes ampliem seu conhecimento independentemente de auxílios externos.

O principal objetivo deste trabalho é de apresentar um sistema com potencialidades para assumir tarefas em que a autonomia é uma característica essencial. Psicologia do Comportamento e Neurofisiologia oferecem as bases teóricas para este desafio. O sistema corresponde a uma rede neural concebida com suporte à estratégia de aprendizagem por reforço. Entre outras características relevantes, citam-se: arquitetura topológica e multicamada, modelo dinâmico para o neurônio; aprendizagem não supervisionada; e ajuste sináptico de camadas internas. Para sua avaliação preliminar adota-se uma aplicação associada ao controle autônomo. Os resultados de simulação confirmam as expectativas: o sistema é capaz de assimilar habilidades de controle sem qualquer auxílio externo, gerando seqüências de respostas que levam o ambiente de um estado inicial (aleatório) a um estado final desejado.

O sucesso preliminar alcançado indica apenas potencialidades do sistema. Aplicações pouco mais complexas deixariam o sistema ineficaz. Investigações têm sido dedicadas no sentido de ampliar suas características para aplicação em navegação autônoma de robôs.

Agradecimentos

Frank Borsato agradece à Fundação Araucária pelo apoio financeiro durante curso para titulação a Mestre em Ciências.

Referências

[1] Haykin, S.; Neural Networks: a comprehensive foundation, Prentice Hall, New York, EUA, 1994.

[2] Maass, W. e Bishop, C. (Eds); Pulsed Neural Networks; MIT Press, Cambridge, EUA, 1999.

[3] Zak, M.; "Terminal attractors in neural networks", Neural Networks (2), 259-274, (1989).

[4] Crook, N. e Scheper, T.; "A novel chaotic neural network architecture"; Proc. of the European Symposium on Artificial Neural Networks; Bélgica, pp. 295-300, 2001.

[5] Bakker, B.; Zhumatiy, V.; Gruener, G. e Schmidhuber, J.; "A robot that reinforcement-learns to identify and memorize important previous observations"; Proc. of the 2003 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, 2003.

[6] Donahoe, J. e Palmer, D.; Learning and Complex Behavior, Massachusetts, Simon & Schuster Inc., 1994.

[7] Gluck, M. e Myers, C.; Gateway to Memory: an introduction to neural network modeling of the hippocampus and learning, MIT Press, Londres, 2001.

[8] Edelman, G.; Neural Darwinism: the theory of neuronal group selection, Basic Books, EUA, 1987.

[9] Antonelo, E. e Figueiredo, M.; "Intelligent autonomous navigation for mobile robots: spatial concept acquisition and object discrimination"; Proc.6th IEEE Int. Symp. on Computational Intelligence in Robotics and Automation, Finlândia, 2005.

[10] Millán, J.; "Rapid, safe, and incremental learning of navigation strategies", IEEE Transactions on SMC – Part B, vol. 26, no.3, 1996.

[11] Sutton, R. e Barto, A.; Reinforcement Learning: an introduction, MIT Press, Cambridge (1998).

[12] Crestani, P.; Figueiredo, M. e Von Zuben, F.; "A hierarchical neuro-fuzzy approach to autonomous navigation," in Proc. of 2002 Int. Joint Conference on Neural Networks, EUA, 2002.

[13] Calvo, R. e Figueiredo, M.; "Reinforcement learning for hierarchical and modular neural network in autonomous robot navigation," in Proc. of 2003 Int. Joint Conference on Neural Networks, EUA, 2003.

[14] Thorndike, E. e Bruce, D. (Introdução), Animal Intelligence: experimental sudies, Transaction Publishers, 1999.

[15] Donahoe, J.; Burgos, J. e Palmer, D.; "A slectionist aproach to rinforcement", J. of the Exp. Analysis of Behavior, 60, 17-40, 1993.

[16] Borsato, F.; Autonomia Cognitiva em Rede Neural Topológica Multicamada de Plasticidade Sináptica Intracamada, dissertação de mestrado, Universidade Estadual de Maringá, 2006.