Proceedings of the IV Brazilian Conference on Neural Networks - IV Congresso Brasileiro de Redes Neurais
pp. 356-360, July 20-22, 1999 - ITA, São José dos Campos - SP - Brazil

Treinamento de Redes Neurais para Tarefas de Inspeção Mediante Reforços de Múltiplos Críticos

Paulo Baz Agra, Takashi Yoneyama, Cairo Lúcio Nascimento Júnior
Divisão de Engenharia Eletrônica
Instituto Tecnológico de Aeronáutica
12.228-900 - São José dos Campos - SP
E-mails: takashi@ita.cta.br, cairo@ita.cta.br

Abstract

The main objective of this work is to investigate the concepts of game theory in the context of reinforcement learning with multiple critics. The neural net is assumed to perform an inspection task, where desirable input patterns should produce high 'accept' output values while undesirable input patters are expected to yield high 'reject' output values. The multiple critics may or may not cooperate, so that game theoretic situations arise. In this context, the stochastic learning automata is shown to converge to Nash equilibrium points or Pareto solutions, depending on the nature of the information state.

1. Introdução

O problema de aprendizado com reforço imediato [19] empregando índices de desempenho multicritérios é estudado utilizando-se os conceitos da Teoria de Jogos. A estrutura empregada neste trabalho encontra-se ilustrada na figura 1.

O problema é classificado em quatro situações particulares, todas no contexto do problema de inspeção e de complexidade crescente no tocante a estrutura de sinais de reforço [1]:

a) treinamento de uma rede com um único objetivo: por exemplo, aceitar apenas as cartas vermelhas dos baralhos.
b) treinamento de uma rede com múltiplos objetivos não conflitantes: por exemplo, um dos críticos aceita cartas vermelhas, enquanto o outro se preocupa em verificar se o número é par ou ímpar.
c) treinamento de uma rede com múltiplos objetivos conflitantes mas com críticos cooperativos: por exemplo, o padrão de entrada pode consistir de 2 cartas, sendo que o primeiro crítico deseja uma primeira carta com valor elevado e segunda com valor baixo, enquanto o segundo crítico deseja uma primeira carta com valor baixo e segunda com valor elevado. Dependendo da estrutura de informação que os críticos utilizam para ajustar a rede, este pode aprender a considerar como o padrão de referência um par de cartas de valores medianos.
d) treinamento de uma rede com múltiplos objetivos conflitantes e críticos não cooperativos: considerando novamente o problema de analisar um par de cartas e dependendo da estrutura de informação, a rede pode aprender a considerar como o padrão de referência duas cartas altas ou duas cartas baixas.

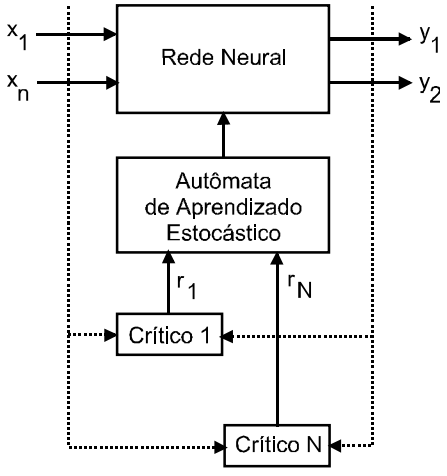

Figura 1 - Uma estrutura de aprendizado por reforço com múltiplos críticos

Tendo-se em vista a facilidade de visualização dos conceitos de jogos utilizando-se as curvas de nível dos índices de desempenho utilizados por cada crítico, os padrões adotados neste trabalho são pares ordenados (x_1,x_2) com x_1 e $x_2 \in R$, ao invés de números inteiros (como o caso do baralho).

O modelo de neurônio utiliza funções de base radial e possui duas saídas assumindo valores no intervalo [0,1]:

$$y_1 = \begin{cases} \uparrow 1 & \text{aceita} \\ \downarrow 0 & \text{rejeita} \end{cases} \qquad y_2 = \begin{cases} \downarrow 0 & \text{aceita} \\ \uparrow 1 & \text{rejeita} \end{cases} \qquad (1)$$

de modo que, quando a saída y_1 se aproxima de 1 e a saída y_2 se aproxima de 0, então a entrada x atual é considerada estar próxima de um padrão desejável (portanto, aprovado na inspeção). Por outro lado,

quando y_1 assume valores pequenos enquanto y_2 assume valor próximo de 1, a entrada x é considerada afastada de um padrão desejável, ou seja, é um padrão indesejável (portanto, reprovado na inspeção).

Após um processo de treinamento com sucesso a saída y_1 tende a ser a negação de y_2. Entretanto, em casos onde os pesos da rede neural não estão adequadamente sintonizados, y_1 pode ser diferente de $\neg y_2$ (not y_2). Assim, de modo geral $y = f^{RN}(x,w)$ onde x é o padrão apresentado à rede, w é o valor atual dos pesos sinápticos e y é a saída.

O padrão de entrada é da forma $x^j = [\, x_1^j \;\; x_2^j \;\; \ldots \;\; x_n^j \,]$ e o conjunto de treinamento é $X \equiv \{ x^j \mid j = 1, \ldots, N\}$. Cada crítico possui um índice de desempenho (custo a ser minimizado) da forma $J_i(x, y)$ associado a uma saída y quando o padrão de entrada é x.

O crítico i produz um sinal de reforço r_i no instante (k+1) segundo:

$$r_i[k+1] = r_i(\phi_i[k], \phi_i[k+1]) \qquad (2)$$

onde:

$$\phi_i = \sum_{j=1}^{N} J_i\left(x^j, f^{RN}\left(x^j, w[k]\right)\right) \qquad (3)$$

Em muitos aplicações r_i assume valores {+1, -1}, embora, no caso geral, tal fato não seja necessário.

2. Aprendizado com reforço imediato

Sejam dadas as funções $f:X \rightarrow Y$ e $f^{RN}:X \times W \rightarrow Y$ onde $X \equiv R^n$, $Y \equiv R^m$ e $W^{p,q,r}$ são os conjuntos das entradas, saídas e pesos, respectivamente. O aprendizado, no contexto desta seção, consiste em determinar os pesos $w \in W$ de uma rede neural caracterizada pela sua relação entrada-saída $f^{RN}(.,.)$, de modo que $f(x) = f^{RN}(x,w)$, ou mais realisticamente, obter uma aproximação $f(x) \cong f^{RN}(x,w)$ ([8], [20]).

No caso de *backward error propagation* os pesos w são ajustados iterativamente de modo que $f^{RN}(.,w)$ aproxime f(.) nos pontos $\{(x_1,f(x_1)), \ldots ,(x_N,f(x_N))\}$ fornecidos pelo supervisor. No caso de *aprendizado com reforço*, os pesos w devem ser ajustados, tambem iterativamente, de modo que $f^{RN}(.,w)$ se aproxime de f(.), baseado nos valores assumidos pelas funções de reforço (ou de recompensa/*reward*, ou de punição/*penalty*) no instante k, ou seja, $r_i[k]$, i = 1, ..., N ([2], [11], [19]). O caso N = 1 corresponde ao aprendizado com reforço clássico.

O reforço imediato é aquele fornecido tão logo uma ação seja realizada, sem haver necessidade de se memorizar as ações passadas. A ação, no caso de redes neurais, seria a atualização dos pesos de acordo com alguma regra pré-estabelecida. Portanto, a cada atualização dos pesos seria recebida uma realimentação do meio ambiente (através dos críticos), no sentido de informar se tal ação foi boa ou não. Considerando-se que r[k] é obtido a partir de $J_i(x,f^{RN}(x,w[k]))$, e utilizando-se o abuso de notação para indicar a alteração nos valores dos pesos w[k] como uma ação "a", a atualização dos pesos para o valor w^* deveria ser tal que

$$J_i\left(x, f^{RN}\left(x, w^*\right)\right) = \min_{a \in A} J_i(x,a) \;\forall x \in X \qquad (4)$$

ou seja, o ponto de ótimo seria aquele valor w^* que minimiza a função custo de cada crítico i, para os padrões $x \in X$. Obviamente, existindo conflito de interesses, há necessidade de um mecanismo para seleção de a, uma vez que não há como minimizar todos os J_i de forma simultânea.

Uma forma de determinar w^* seria testar todos os valores $a \in A$, para cada x, e selecionar aquele que, segundo algum critério seja o melhor, mas os métodos de enumeração são custosos numericamente, mesmo no caso de A ser um conjunto finito $A = \{a_1, \ldots , a_N\}$.

Um método alternativo seria o que emprega índices de mérito (ou densidade de probabilidade, no caso de se utilizar a versão estocástica). Considere-se a função de mérito $g:X \rightarrow R^N$ que fornece, para cada x, um vetor $z \in R^N$, cuja componente z_i é proporcional ao mérito de se utilizar a ação a_i. A função de mérito g(x) pode, por sua vez, ser realizada por uma rede neural $g^{RN}(x,v)$, com pesos ajustáveis v.

Seja k tal que $z_k \geq z_i$, $\forall\, i \neq k$, ou seja, a ação a_k possui o maior mérito no momento, dado por z_k.

O problema de construir a função de mérito $g^{RN}(x,v)$ pode ser resolvido pelo mecanismo de aprendizado:

$$g_k(x,v^{novo}) := g_k(x,v^{velho}) + \\ + \alpha\psi(r_1(x,a_k), \ldots , r_N(x,a_k)) \qquad (5)$$

$$g_{i\neq k}(x,v^{novo}) := g_{i\neq k}(x,v^{velho}) - \\ - \alpha\psi(r_1(x,a_k), \ldots , r_N(x,a_k)) \qquad (6)$$

onde ψ é a função que permite representar a estrutura de informações dos múltiplos críticos.

Este mecanismo de atualização dos pesos pode ser realizado por um *stochastic learning automata*. É, basicamente, uma máquina abstrata que seleciona ações "a_j" em função de g_j (que pode ser a densidade de probabilidade) e recebe realimentações r_i do ambiente avaliando essas ações. As densidades g_i são armazenadas internamente pelo automata e são atualizada de acordo com a realimentação avaliativa provida pelo ambiente.

A idéia é que as ações que recebem avaliações mais favoráveis se tornem as escolhas mais prováveis pelo automata. Em geral, o ambiente, através dos críticos e função ψ pode prover sua avaliação de acordo com um

conjunto de distribuições, uma para cada ação possível de automata.

3. Equilíbrio na Teoria de Jogos

O objetivo deste tópico é apresentar alguns tipos de equilíbrio que ocorrem em problemas de múltiplos critérios ([7],[14], [15]), particularmente em situações nas quais ocorrem conflitos de interesse, tipicamente em jogos. Mais especificamente, serão abordados aqui dois tipos de equilíbrio: os de Pareto e de Nash [9].

Como J_i depende de y, que por sua vez é obtida a partir de w^*, que é ajustada un função de r_1, ... , r_N, seja $J_i(x,r_1,...,r_N)$, novamente valendo-se de abuso de notação, o custo associado com o crítico i em um instante onde o padrão apresentado é x. Nestas condições, o equilíbrio Nash corresponde a sinais de reforço r_i^* tal que, $\forall r_i$:

$$E\left[J_i\left(x,r_1^*,...,r_i^*,...,r_N^*\right)\right] \le E\left[J_i\left(x,r_1^*,...,r_i,...,r_N^*\right)\right] \quad (7)$$

e a otimalidade de Pareto é caracterizada por:

$$\begin{aligned} &E\left[J_i\left(x,r_1,...,r_N\right)\right] \le E\left[J_i\left(x,r_1^{\#},...,r_N^{\#}\right)\right] \Rightarrow \\ &E\left[J_i\left(x,r_1,...,r_N\right)\right] = E\left[J_i\left(x,r_1^{\#},...,r_N^{\#}\right)\right], i=1,...,N \end{aligned} \quad (8)$$

e o conjunto $\Lambda = \{r^{\#}\}$ é denominado de conjunto de soluções Pareto ótimas ou soluções não-inferiores.

Por simplicidade, considere-se um jogo disputado por apenas dois jogadores. Cada jogador jusca minimizar o seu próprio custo J_i associado a um padrão caracterizado por apenas 2 coordenadas (x_1,x_2), tanto mais aceitável pelo jogador i quanto menor o valor J_i (de onde se obtém a função $J_i(x,y)$). No caso particular de $J_1 = -J_2$, diz-se que o jogo é de soma zero e, muitas vezes, o ponto de ótimo corresponde a um ponto de sela [11], [13] e [14].

Considere-se agora as curvas de nível de J_i. Estas são de tal modo distribuídas que, ao se aproximar do centro da distribuição, o valor de J_i diminui. A figura 2 fornece uma visualização do plano $\{x_1,x_2\}$ e os conjuntos de curvas de nível para dois jogadores com índices de desempenho distintos, bem como as soluções Nash e as soluções Pareto-ótimas. Métodos numéricos para solução de problemas com múltiplos critérios podem ser encontrados em [6], [13]. [14].

4. Exemplo Numérico

O exemplo numérico considera um padrão com 2 componentes (x_1 e x_2) e o treinamento deve fazer com que o pico da função $y = f^{RN}(x,w)$ esteja localizado em torno de um padrão considerado desejável, segundo os múltiplos critérios adotados pelos críticos. No caso, a função é monomodal e o seu pico deve se localizar em

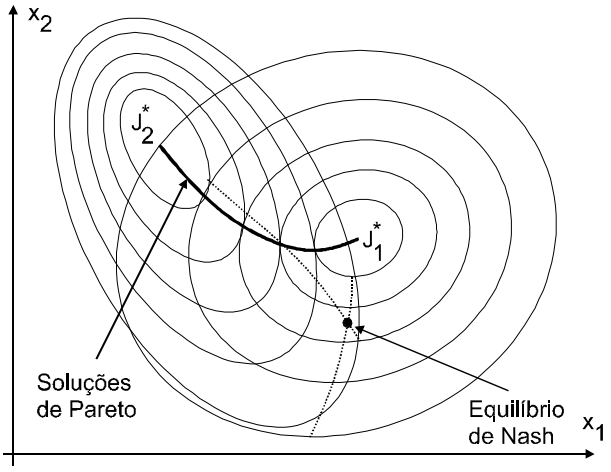

Figura 2 - Conjunto de Soluções de Pareto e um Ponto de Equilíbrio Nash para um Jogo de Soma Não-Zero.

um ponto de equilíbrio para os múltiplos críticos (cooperativos ou conflitantes).

A seqüência de treinamento nada mais é que um laço iterativo com ações pré-determinadas pelo método de recompensa-penalidade. Primeiro, apresenta-se o primeiro elemento do vetor de entradas à rede. Partindo da configuração inicial, calculam-se as saídas da rede. O próximo passo é a escolha da ação. A ação é sorteada por uma rotina baseada em números aleatórios, dentre o elenco das alternativas possíveis, e é executada. Ou seja, se a ação é, por exemplo, aumentar o peso w_1, esse valor é acrescido de d. A nova saída da rede é então calculada. A seguir, a entrada e as saídas iniciais e finais da rede são apresentadas a cada crítico que emite seu parecer – bom ou ruim – através da variável booleana r_i. Com base nesse parecer é calculada a função ψ que leva em consideração a estrutura de informação e, então, é aumentada ou diminuída a probabilidade associada a cada ação. Importante ressaltar que o crítico não tem nenhum acesso aos parâmetros da rede. Ele apenas observa a entrada e as saídas e emite sua opinião. Detalhes de técnicas de aprendizado com reforço podem ser encontradas em [2], [3], [11] e [16], entre outros.

As figuras 3 e 4 ilustram a convergência do processo de treinamento para pontos Pareto-ótimos, correspondendo a cooperação entre os críticos. Nestes casos $J_1(x_1,x_2)$ e $J_2(x_1,x_2)$ são:

$$J_1(x_1,x_2) = (x_1-2.5)^2 + (x_2+0.5)^2 \quad (9)$$

$$J_2(x_1,x_2) = (x_1-0.5)^2 + (x_2-2.5)^2 \quad (10)$$

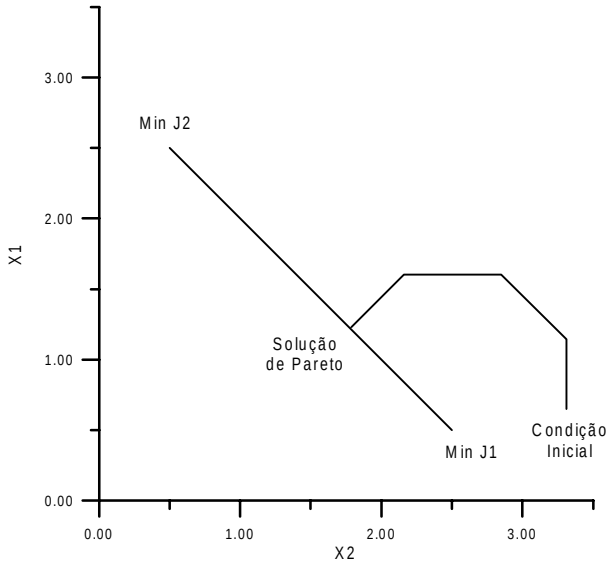

Figura 3 - Treinamento com Críticos Cooperativos para uma solução de Pareto a partir de (3.3,0.5)

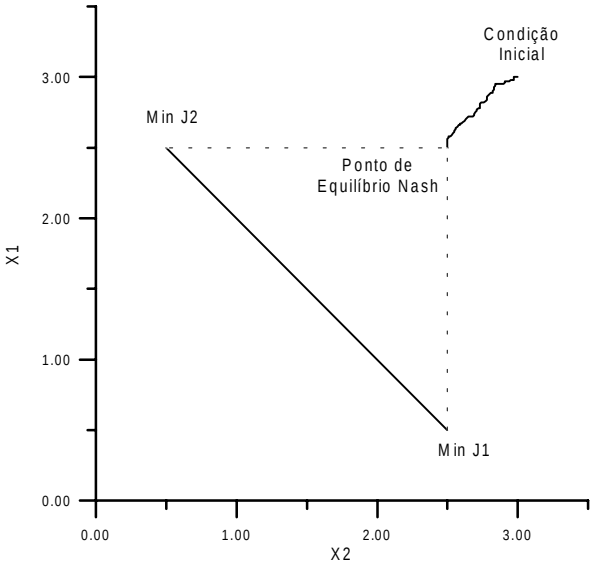

Figura 5 - Treinamento Não Supervisionado com Críticos Conflitantes Racionais a partir de (3.0,3.0)

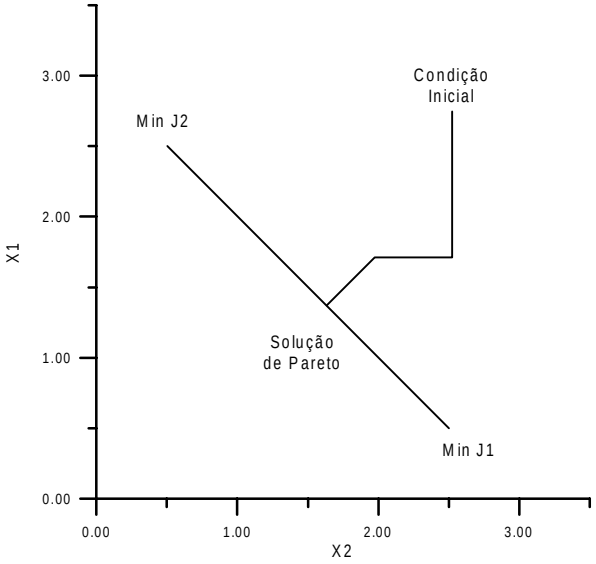

Figura 4 - Solução de Pareto atingido a partir da condição inicial (2.5, 3.0)

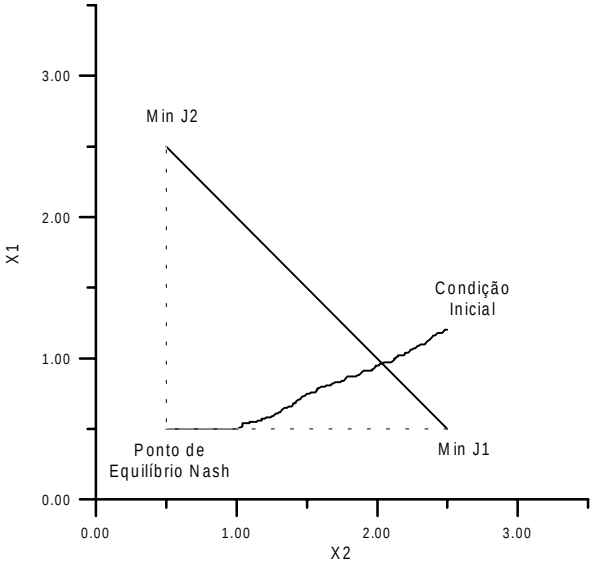

Figura 6 - Convergência para um ponto de equilíbrio Nash a partir da condição inicial (2.5,1.25)

Verifica-se que a rede, devidamente modelada para um caso de conflito de objetivos mas de comportamento cooperativo, apresenta resultados que podem ser previstos pela Teoria de Jogos. Nota-se, no caso de equilíbrio de Pareto, que mediante cooperação, o padrão de referência vem se localizar em um ponto (x_1,x_2) tal onde os custos tendem a piorar para algum dos críticos, caso se desloque de (x_1,x_2) e o custo atingido neste ponto por cada índice é inferior ao de equilíbrios racionais como o de Nash. Obviamente, para críticos não racionais, os custos podem ser melhores ou piores para o crítico racional, dependendo da realização.

As figuras 5 e 6 ilustram a convergência do processo de treinamento para pontos Nash, em vista da não cooperação dos críticos.

Novamente pode-se observar que a rede, devidamente modelada para uma situação de múltiplos objetivos conflitantes porém não cooperativos, tem seu comportamento previsto pela Teoria dos Jogos.

5. Conclusões

Entre as vantagens observadas no enfoque de aprendizado não-supervisionado com multi-objetivos pode-se citar a flexibilidade, quando as redes possuem múltiplas possibilidades de satisfazer os objetivos desejados; a adaptabilidade, visto que o método de

recompensa-penalidade pouco foi alterado em cada situação para que as simulações fossem feitas; e a independência ou isolamento, pois durante todo o processo de aprendizagem, os críticos nada precisaram saber do estado interno da rede, apenas seu comportamento.

Foi observada também a completa correspondência entre o comportamento da rede e os conceitos da Teoria de Jogos. Apesar de serem estudados apenas alguns tópicos de equilíbrio, a perspectiva de que outros conceitos como o de max-min e o de equilíbrio de Stackelberg [8] possam vir a ser aplicados a situações de treinamento multicritérios é muito grande. Isso faz com que aumente a previsibilidade a respeito do comportamento de redes neurais artificiais treinadas com múltiplos objetivos.

Agradecimentos

Os autores desejam manifestar os seus agradecimentos ao Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq pelo suporte a este trabalho na forma do Projeto ProTeM-CC/SHI^2, processo 680092/95-1.

Referências

[1] Agra, P. B., *Aprendizado com Reforço Empregando Índices de Desempenho com Múltiplos Critérios*, Relatório Interno IEES-001, ITA, São José dos Campos, SP, 1997.

[2] Anderson, C. W., *Strategy Learning with Multilayer Connectionist Representations*, GTE Labs, Waltham, MA, Report TR 87-509.3, 1988.

[3] Barto, A. G. & Anandan, P., Pattern-Recognizing Stochastic Learning Automata, *IEEE Trans, Syst. Man Cybern.*, vol. 15, 1985, pp. 360-375.

[4] Basar, T., Nash Strategies for N-person Differential Games with Mixed Information Structures, *IEEE TAC*, vol. 20, no. 3, jun 1995, pp. 320-328.

[5] Blackwell, D. & Girshick, M. A., *Theory of Games and Statistical Decisions*, John Wiley, New York, 1954.

[6] Goodwin, G. C. et al, On the Optimization of Vector-Value Performance Criterion, *IEEE TAC*, dez 1975, pp. 803-804.

[7] Haykin, S., *Neural Networks - A Comprehensive Foundation*, Macmillan College Publishing Company, New York, 1994.

[8] Ho, Y. C. & Starr, A. W., Nonzero-sum Differential Games, *J. Opt. Theory Appl.*, vol. 3, no. 3, 1969, pp. 184-206.

[9] Karlin, S., *Mathematical Methods and Theory in Games, Programming and Economics*, Addison-Wesley, Reading, MA, 1962.

[10] Keerthi, S. S. & Ravindran, B., *A Tutorial Survey of Reinforcement Learning*, Indian Institute of Science, Bangalore, 1996 (URL: ftp://archive.cis.ohio-state.edu/pub/neuroprose/keerthi.rl-survey.ps.Z).

[11] Luce, R. D. & Raiffa, H., *Games and Decisions: Introduction and Critical Survey*, John Wiley, 1957.

[12] McCulloch, W. S. & Pitts, W. H., A Logical Calculus of the Ideas Imminent in Nervous Activity, *Bull. Math. Biophy.*, 1943.

[13] Mukai, H., Algorithms for Multicriterion Optimization, *IEEE TAC*, vol. 25, no. 2, abr 1980, pp.177-186.

[14] Osyczka, A., *Multicriterion Optimization in Engineering*, Halstead Press, 1984.

[15] Salukvadze, M. E., *Vector Value Optimization Problems in Control Theory*, Academic Press, New York, 1979.

[16] Sutton, R. S.; Barto, A. G.; Willians, R. J., Reinforcement Learning is Direct Adaptive Control, *IEEE Control Systems Magazine*, apr 1992, pp. 19-22.

[17] Von Neumann, J. & Morgenstern, O. , *Theory of Games and Economic Behaviour*, Princeton University Press, 3ª ed., 1953.

[18] Widrow, B., Gupta, N. K., Maitra, S., Punish/Reward: Learning with a Critic in Adaptive Threshold Systems, *IEEE Trans. Syst. Man and Cybern.*, vol. 3, no. 5, sep 1973, pp 455-465.

[19] Williams, R. J., *Reinforcement Learning in Connectionist Networks: A Mathematical Analysis*. Tech. Report nº 8605, Institute for Congitive Science, University of California, San Diego, 1986.

[20] Zurada, J. M., *Introduction to Artificial Neural Systems*, West Publishing Company, St. Paul, MN, 1992.